Valide und vergleichbare Erfassung bildungsrelevanter Konstrukte bei Schülerinnen und Schülern mit sonderpädagogischem Förderbedarf Lernen

Inauguraldissertation

in der Fakultät Humanwissenschaften

der Otto-Friedrich-Universität Bamberg

vorgelegt von

Dipl.-Päd. Lena Mariana Nusser

aus Bad Nauheim

Bamberg, den 31.08.2017

URN: urn:nbn:de:bvb:473-opus4-511036
DOI: https://doi.org/10.20378/irbo-51103

Tag der mündlichen Prüfung: 13.11.2017

Dekan: Universitätsprofessor Dr. Stefan Hörmann

Erstgutachter: Universitätsprofessor Dr. Claus H. Carstensen

Zweitgutachterin: Universitätsprofessorin Dr. Cordula Artelt

Danksagung

Auf dem Weg zu meiner Dissertation haben mich viele Menschen gestärkt, motiviert und inspiriert, denen ich dafür sehr dankbar bin.

Zunächst und ganz besonders möchte ich meinem Betreuer Prof. Dr. Claus H. Carstensen und meiner Betreuerin Prof. Dr. Cordula Artelt für ihre Beratung und Begleitung während meiner Dissertationsphase danken. Ihre Anregungen, Kritik und ihr Mentoring haben mich stets motiviert und zum Gelingen meiner Dissertation wesentlich beigetragen. Darüber hinaus bin ich Prof. Dr. Sabine Weinert sehr dankbar, die mir die Chance gegeben hat, an diesem innovativen Projekt der Machbarkeitsstudien an Förderschulen Lernen im Rahmen des Nationalen Bildungspanels (NEPS) mitzuwirken. Ihre Unterstützung zu entscheidenden Meilensteinen meiner Dissertation war sehr wichtig für mich.

Dankbar bin ich auch für die vielen wissenschaftlichen Diskussionen mit meinen KollegInnen im NEPS (Kompetenzsäule) und LIfBi e.V. (Arbeitsbereich Kompetenzentwicklung), ihren konstruktiven Anmerkungen zu meinen Schriften. Insbesondere möchte ich die Arbeit und Unterstützung meiner KollegInnen Jana Heydrich und Markus Messingschlager im NEPS-Förderschulteam (2010-2014), die sehr stark an der Gestaltung der Studien und Datenerhebungen beteiligt waren, hervorheben.

Besonderer Dank gilt Karin Gehrer, die mit ihrer positiven Energie und liebevollen Begleitung den Weg immer wieder etwas leichter gemacht hat.

Abschließend möchte ich meiner Familie danken, die meinen Weg zur Dissertation mit großem Interesse begleitet hat.

Inhalt

1 Einleitung

Seit der Jahrtausendwende hat die Anzahl an und Bedeutung von Schulleistungsstudien stetig zugenommen. Im Fokus stehen hierbei sowohl Trendanalysen (z.B. Programme for International Student Assessment [PISA], Klieme et al., 2010; IQB-Ländervergleiche, Pant et al., 2013; Stanat, Pant, Böhme & Richter, 2012) als auch die Erfassung individueller Entwicklungsverläufe von Kompetenzen über die Lebensspanne (Nationales Bildungspanel [NEPS]; Blossfeld, Roßbach & von Maurice, 2011). Large-Scale-Studien betrachten in unterschiedlichen Altersstufen verschiedene Kompetenzen, die für den Bildungserfolg, die gesellschaftliche Teilhabe sowie die Berufskarriere relevant sind (Baumert, Artelt, Carstensen, Sibberns & Stanat, 2002; Weinert et al., 2011). Mit Hilfe dieser Daten können z. B. Einflussfaktoren für die Kompetenzentwicklung, soziale Disparitäten oder Determinanten von Bildungserfolg aufgedeckt werden. In Schulleistungserhebungen werden bundesweit repräsentative Stichproben gezogen, um Aussagen über das deutsche Bildungssystem generieren zu können. SchülerInnen mit einem sonderpädagogischen Förderbedarf (SPF)[1] haben in diesem Kontext in Deutschland zunächst sehr wenig Aufmerksamkeit erhalten; sie wurden in der Regel zu Repräsentativitätszwecken mit kleiner Fallzahl in die Stichproben inkludiert (Hörmann, 2007), so dass kaum differenzielle Analysen möglich waren. Obwohl der Anteil der SchülerInnen mit einem diagnostizierten SPF jährlich anwächst (Autorengruppe Bildungsbericht, 2016; Dietze, 2012), ist ein verstärktes nationales Interesse an dieser Gruppe erst relativ spät entstanden. Zunächst existierten überwiegend regional begrenzten Studien an Förderschulen, z. B. die Studien

[1] In Deutschland existieren insgesamt sieben Förderschwerpunkte, welche durch die Kultusministerkonferenz der Länder (2017) definiert wurden: 1) emotionale und soziale Entwicklung, 2) geistige Entwicklung, 3) Hören, 4) körperliche und motorische Entwicklung, 5) Lernen, 6) Sehen, 7) Sprache sowie die Sammelkategorie Lernen, Sprache und emotionale und soziale Entwicklung (KMK - Sekretariat der Ständigen Konferenz der Kultusminister der Länder, 2017).

Lernausgangslage an Förderschulen (LAUF, Wocken, 2005), Berliner Erhebung arbeitsrelevanter Basiskompetenzen von Schülerinnen und Schülern mit Förderbedarf „Lernen" (BELLA, Lehmann & Hoffmann, 2009), Kompetenzen und Einstellungen von Schülerinnen und Schülern an Hamburger Schulen zu Beginn der Jahrgangsstufe 7 (KESS7, Bos, Gartmeier & Gröhlich, 2009) oder die Panel Study at the Research School „Education and Capabilities" in North Rhine-Westphalia an Förderschulen (PARS-F, Müller, Stubbe & Bos, 2013). Ein Oversample von SchülerInnen mit SFP Lernen bzw. Lernen, Sprache und emotionale und soziale Entwicklung, welches differenzierte Analysen ermöglichen würde, wurde erstmals 2010 im Nationalen Bildungspanel (NEPS, Heydrich, Weinert, Nusser, Artelt & Carstensen, 2013), 2011 im IQB-Ländervergleich (Kocaj, Kuhl, Kroth, Pant & Stanat, 2014) und 2012 im Rahmen von PISA (Gebhardt, Sälzer, Mang, Müller & Prenzel, 2015) in bundesweite Large-Scale-Assessments integriert.

Für die anfänglich marginale Berücksichtigung dieser Schülergruppe in Schulleistungserhebungen in Deutschland kann vor allem ein Hauptgrund angenommen werden. Für eine Integration von SchülerInnen mit SPF in Large-Scale-Studien, die vergleichende Analysen erlauben, sind angemessene Instrumente und Administrationsbedingungen zur Erfassung bildungsrelevanter Konstrukte notwendig. In internationalen Studien ist die Implementation von *Akkommodationen* eine häufige Maßnahme, um behinderungsbezogene, aber konstrukt-irrelevante Hindernisse zu minimieren (Koretz & Barton, 2004). Die Anwendung von spezifischen Anpassungen bezieht sich in der Regel auf das Erhebungsmaterial und die Administration der Instrumente (z.B. Zeitvorgaben, Einzel- oder Kleingruppensitzung). Inwieweit solche zielgruppenspezifischen Anpassungen die Validität der Messung beeinflussen, kann auf Grund der heterogenen Befundlage noch nicht abschließend beurteilt werden (Cormier,

Altman, Shyyan & Thurlow, 2010). Somit gibt es keine eindeutige Entscheidungsgrundlage, wie mit neuen Instrumenten zu verfahren ist. Vor allem für Deutschland gibt es keine Leitlinien, die für die Anwendung im Rahmen von Schulleistungsstudien oder Large-Scale-Studien empfohlen werden (Müller et al., 2017; Südkamp, Pohl, Hardt, Jordan & Duchhardt, 2015). Die Testinstrumente großer Leistungsstudien verfolgen in ihrer Anlage die Messung von Kompetenzen einer großen Population, die ein breites Leistungsspektrum aufweist. Daraus ergibt sich, dass die eingesetzten Maße vor allem im stark vertretenen mittleren Fähigkeitsbereich differenzieren und weniger stabil im dünner besiedelten oberen und unteren Fähigkeitsbereich messen. In einer aktuellen Studie von Gebhardt et al. (2015) erreichte die Mehrheit der SchülerInnen an Förderschulen mit dem Schwerpunkt Lernen nicht die unterste PISA-Kompetenzstufe in Lesen und Mathematik. Somit fehlen Daten, welche differenzielle Analysen zu Stärken und Schwächen (Kompetenzprofile) dieser SchülerInnengruppe erlauben. Ob die Tatsache, dass SchülerInnen mit SPF Lernen die unterste Kompetenzstufe nicht erreichen, mit der Itemschwierigkeit oder ggf. auch mit einer fehlenden „Zugangsfertigkeit" (i. S. v. geringes Instruktionsverständnis, niedrige Dekodiergeschwindigkeit; Kettler, Braden & Beddow, 2011; Renner & Mickley, 2015) der SchülerInnen mit SPF zusammenhängt, wurde noch nicht empirisch entflochten.

Darüber hinaus ist die Erfassung anderer bildungsrelevanter Konstrukte, die häufig über schriftliche Befragungen erhoben werden, ein weiterer zentraler Aspekt in Large-Scale-Assessments. Auch bei schriftlichen Befragungen werden Anforderungen an die Lesekompetenz und Aufmerksamkeit der SchülerInnen gestellt, die einen Einfluss auf die Validität haben können (Bell, 2007; Borgers, De Leeuw & Hox, 2000). Somit ist es insgesamt erforderlich, angemessene Instrumente und Administrationsbedingungen für diese SchülerInnengruppe zu entwickeln (Heydrich et al., 2013; Oser & Biedermann, 2006). Diese

Kompetenztests und Befragungsinstrumente müssen jedoch nicht nur den Anforderungen der gängigen Gütekriterien, wie der Validität, entsprechen, damit sichergestellt wird, dass das theoretisch zugrundeliegende Konstrukt erfasst wurde. Zusätzlich müssen die Daten der einzelnen Aufgaben oder Fragen messäquivalent mit anderen SchülerInnengruppen sein, damit vergleichende Analysen und Aussagen über SchülerInnen mit SPF möglich werden.

Zunehmend kann in der Öffentlichkeit, der Politik und dem Bildungssystem ein gesteigertes Interesse wahrgenommen werden, mehr über die SchülerInnen mit Förderbedarf und ihre Kompetenzentwicklung und Lebenswege zu erfahren und somit diese Gruppe bei Schulleistungserhebungen und Large-Scale-Assessments nicht mehr auszusparen. Wie eine valide und vergleichbare Erfassung bildungsrelevanter Konstrukte gelingen kann, ist Hauptanliegen der vorliegenden Arbeit. Unter bildungsrelevanten Variablen werden Maße verstanden, die direkten oder indirekten Einfluss auf die Bildungsverläufe und Kompetenzentwicklung von Individuen haben. An ausgewählten Inhalten und Beispielen werden die Forschungsdesiderata angegangen, wie insbesondere der Gruppe der SchülerInnen mit einem sonderpädagogischen Förderbedarf Lernen (SPF-L) der Zugang zu Inhalten und Aufgaben erleichtert werden kann, damit eine valide und vergleichbare Messung relevanter Variablen im Rahmen von Schulleistungserhebungen oder anderen Large-Scale Studien möglich wird.

Die vorliegende Arbeit gliedert sich in drei Teile. Zunächst wird in einer theoretischen Einbettung die Zielgruppe der Schülerschaft mit SPF-L näher beschrieben. Zur Erklärung der Herausforderungen, die sich für eine valide Erfassung bildungsrelevanter Konstrukte abzeichnen, werden zwei kognitive Modelle herangezogen: Das Modell in Anlehnung an

die Prinzipien des dynamischen Testens nach Carlson und Wiedl (1992), welches die Probleme einer suboptimalen Anwendung der tatsächlichen Kompetenz und der daraus resultierenden Testperformanz aufgreift sowie das Modell nach Tourangeau, Rips und Rasinksi (2000), welches die kognitiven Prozessschritte, die ein/e Befragte/r bei der Beantwortung von Fragebogenitems durchläuft, fokussiert. Die Ausgangslage der SchülerInnen mit SPF-L wird den theoretischen Modellen gegenübergestellt und somit die Herausforderungen bei der Bearbeitung von Fragebögen oder Tests und die Implikationen für Large-Scale-Studien herausgearbeitet. Anschließend werden die übergeordneten Forschungsfragen dieser Arbeit und die zugrundeliegenden Schriften vorgestellt. Die Ergebnisse werden in Bezug auf die dargelegten Forschungsfragen diskutiert und entsprechend der theoretischen Einbettung eingeordnet, bevor abschließend die Grenzen dieser Arbeit und die Implikationen für Large-Scale-Assessments unter Einbezug von SchülerInnen mit SPF-L dargelegt werden.

2 Theoretischer Rahmen

In diesem Abschnitt werden zunächst die SchülerInnen mit einem sonderpädagogischen Förderbedarf Lernen beschrieben und die Besonderheiten dieser Gruppe aufgezeigt werden. Darauf aufbauend werden die speziellen Herausforderungen bei der Befragung und Testung dieser Personengruppe konkretisiert und bezüglich möglicher Implikationen diskutiert.

2.1 Schülerinnen und Schüler mit sonderpädagogischem Förderbedarf Lernen

2.1.1 Definition

Die Anzahl der SchülerInnen mit einem sonderpädagogischen Förderbedarf (SPF) erreichte in den letzten Jahren immer wieder neue Höchststände. Zuletzt wurden im Schuljahr 2014/15 rund 7 % aller SchülerInnen in Deutschland mit einem sonderpädagogischen Förderbedarf diagnostiziert (Autorengruppe Bildungsbericht, 2016), der einem der sieben von der deutschen Kultusministerkonferenz (KMK) definierten Förderschwerpunkte zugeordnet ist (Bundschuh, 2006). Die Diagnostik eines solchen sonderpädagogischen Förderbedarfes und somit die Praxis der Zuweisung entsprechender Fördermaßnahmen sind auf Grund des föderalistischen Systems von Bundesland zu Bundesland unterschiedlich (Sälzer, Gebhardt, Müller & Pauly, 2015). Dies wird auch an den stark differierenden Förderquoten in den einzelnen Bundesländern evident (Dietze, 2012). Der Großteil (44 %) der SchülerInnen mit SPF weist einen sonderpädagogischen Förderbedarf Lernen (SPF-L) auf (Autorengruppe Bildungsbericht, 2016). Die KMK geht von SPF-L für SchülerInnen aus, wenn diese „in ihrer Lern- und Leistungsentwicklung so erheblichen Beeinträchtigungen unterliegen, dass sie auch mit zusätzlichen Lernhilfen der allgemeinen

Schulen nicht ihren Möglichkeiten entsprechend gefördert werden können“ (KMK, 1999, S. 4).

Der Terminus *sonderpädagogischer Förderbedarf Lernen* hat den Begriff *Lernbehinderung* im Zuge der Etablierung und Empfehlungsformulierung für verschiedene Förderschwerpunkte durch die KMK in den 1990er-Jahren abgelöst (Kretschmann, 2006). Die internationale Begriffsbestimmung von *learning disability* ist nach Lloyd, Keller & Hung (2007) der deutschen Definition von Lernbehinderung bzw. SPF-L ähnlich. Hierunter werden schulische Schwierigkeiten gefasst, für die weder eine andere Behinderung noch fehlende Beschulung verantwortlich gemacht werden können (Lloyd et al., 2007). Auch wenn die Definitionen nicht umfassend äquivalent sind, so kann doch von einer großen Überlappung der Populationen ausgegangen werden.

Zur näheren Beschreibung von Lernbehinderung und Abgrenzung gegenüber Lernstörungen im deutschen Schulsystem haben Klauer und Lauth (1997) ein Konzept der dimensionalen Klassifikation erarbeitet, das sowohl den Zeitraum (vorübergehend vs. persistent) als auch das Ausmaß (bereichsspezifisch vs. generalisiert) der Beeinträchtigung berücksichtigt. Im Sinne einer individuellen Bezugsnorm werden partielle Lernschwierigkeiten definiert, die sich in domänenspezifischen Lernstörungen äußern, wobei diese sowohl vorübergehend als auch persistent (z. B. Dyskalkulie oder Legasthenie) sein können (Klauer & Lauth, 1997). Generalisierte Lernschwierigkeiten mit einer SchülerInnenperformanz unterhalb der Klassen- oder Gruppennorm, die zeitlich überdauernd sind, werden im Schema nach Klauer und Lauth (1997) als Lernbehinderung bezeichnet. Als Oberbegriff nutzen die beiden Autoren die Bezeichnung

Lernschwierigkeiten. Diese werden von Zielinski (1996) als „Probleme der Informationsaneignung durch ein Individuum" beschrieben.

2.1.2. Genese

Die Entstehung von Lernschwierigkeiten, also von Schwierigkeiten beim Erwerb neuer Wissensinhalte, wird multifaktoriell verstanden. Zielinski (1996) nennt insgesamt drei interne Faktoren ((1) mangelndes Instruktionsverständnis, (2) mangelndes aufgabenspezifisches Vorwissen, (3) mangelnde Lernmotivation) und drei externe Bedingungen ((4) nicht ausreichende Lernzeit, (5) mangelnde Unterrichtsqualität, (6) moderierende Zusatzbedingungen), die zur Entstehung von Lernschwierigkeiten beitragen[2].

(1) *Instruktionsverständnis:* Das Verständnis verbaler Instruktionen, welche im Unterricht zur Vermittlung von Lerninhalten genutzt werden, erfordert eine zügige Enkodierung der sprachlichen Reize (Schlee, 1974). Bereits sehr frühe rezeptive Sprachfähigkeiten können die spätere Performanz und Schulleistung voraussagen (Hohm, Jennen-Steinmetz, Schmidt & Laucht, 2007; Kotzerke, Röhricht, Weinert & Ebert, 2013). Die sprachlichen Kompetenzen begünstigen demnach den Erwerb von Wissen und schulischen Erfolg (Bos, Buddeberg, Bremerich-Vos & Schwippert, 2012). Ein „mangelndes Instruktionsverständnis" (Zielinski, 1996, S. 371) könnte somit die Genese von Lernschwierigkeiten begründen. Andere Autoren nennen diese elementaren Fähigkeiten (auditive Informationen aufnehmen, relevante Informationen herauslösen und analysieren) Basisfertigkeiten, die hinreichend

[2] Die von Zielinski (1996) postulierten Faktoren sind angelehnt an ein Modell von Haertel, Walberg und Weinstein (1983) zu Bedingungen und Einflussfaktoren des generellen Schulerfolges.

bewältigt werden müssen, um weitere Kompetenzen zu erwerben (Lauth, Brunstein & Grünke, 2014).

(2) *Aufgabenspezifisches Vorwissen:* Individuelles Vorwissen hilft, neue Lerninhalte zu strukturieren, zu kategorisieren, Verknüpfungen herzustellen, die Aufmerksamkeit auf relevante (neue) Inhalte zu lenken (Shuell, 1986). Weinert, Helmke und Schneider (1989) konnten zeigen, dass das Vorwissen ein stärkerer Prädiktor für Gedächtnisleistung bzw. die Ergebnisse eines Mathematiktests darstellt als die nonverbale Intelligenz der SchülerInnen (siehe auch Stern, 1998, 2009). Der Wissensstand beeinflusst den Erwerb neuer Kenntnisse (Grube & Hasselhorn, 2006; Siegler, 1983). Daher geht Zielinski davon aus, dass „mangelndes aufgabenspezifisches Vorwissen" (1996, S. 373) bei SchülerInnen mit Lernschwierigkeiten zu finden ist. Da SchülerInnen mit Lernschwierigkeiten generell über eine kleinere, weniger stark vernetzte Wissensbasis verfügen, kann vermutet werden, dass der Erwerb weiterer Kenntnisse besonders erschwert ist (Lauth et al., 2014). Zudem scheinen sie weniger Erfahrungen mit konkreten Aufgaben, wie sie im schulischen Kontext zu erwarten sind, zu haben (Lauth et al., 2014).

(3) *Lernmotivation:* Wenn die Aussichten auf schulischen Erfolg gering sind, ist auch die Bereitschaft für Lernanstrengung weniger ausgeprägt (Heckhausen & Heckhausen, 2009). Werden erlebte Misserfolge stärker internal attribuiert – also auf die eigenen (mangelnden) Fähigkeiten zurückgeführt – führt dies in der Konsequenz zu einer geringeren Lernmotivation (Wilbert, 2010a). Zielinski geht davon aus, dass „mangelnde Lernmotivation" (1996, S. 375) ein dritter internaler Bedingungsfaktor für Lernschwierigkeiten darstellt. Da SchülerInnen mit Lernschwierigkeiten mehr Misserfolge beim schulischen Lernen als ihre Peers ohne Beeinträchtigungen erleben, kann

angenommen werden, dass ihre Anstrengungsbereitschaft und ihre Motivation weniger ausgeprägt sind (Lauth, 2000).

(4) *Lernzeit:* Auf Grund der beschriebenen Differenzen hinsichtlich der individuellen Vorkenntnisse von SchülerInnen ist davon auszugehen, dass die Kinder und Jugendlichen nicht alle im gleichen Tempo neue Lerninhalte aufnehmen und verstehen können. Somit müssen in einem adäquaten Unterricht binnendifferenziert die heterogenen Ausgangsvoraussetzungen berücksichtigt und SchülerInnen unterschiedlich viel Lernzeit und Unterstützung gewährt werden (Weinert, 1997). Zielinski nennt „nicht ausreichende Lernzeit" (1996, S. 377) als weiteren Bedingungsfaktor für Lernschwierigkeiten. Dies korrespondiert mit den beschriebenen Merkmalen von Kindern mit Lernbehinderung, für die festgestellt wurde, dass sie langsamer und insgesamt weniger lernen sowie mehr Wiederholungen benötigen (Grünke, 2004). Es gibt jedoch auch Hinweise in der Richtung, dass SchülerInnen mit Lernschwierigkeiten insgesamt weniger Zeit für ihre Lernbemühungen investieren (Lauth et al., 2014) und häufiger oberflächlich arbeiten, eher raten und weniger Aufmerksamkeit auf spezifische Anforderungen der ihnen gestellten Aufgaben richten (Klauer & Lauth, 1997; Scruggs, Bennion & Lifson, 1985).

(5) *Unterrichtsqualität:* Als relevante Einflussgröße für erfolgreiches Lernen wird qualitativ hochwertiger Unterricht angesehen (Helmke, Schneider & Weinert, 1986). Als wichtige Faktoren für einen erfolgreichen Unterricht werden effiziente Klassenführung, explizite Instruktionen, angeleitete Übungen und eindeutige, zeitnahe Rückmeldungen an die SchülerInnen (Schrader & Helmke, 2008; Vaughn, Gersten & Chard, 2000) identifiziert. Laut Zielinski wirken „mangelnde Unterrichtsqualität" (1996, S. 378) sowie „ineffektives

Klassenmanagement“ (1996, S. 379) begünstigend auf die Entstehung einer Lernschwierigkeit.

(6) *Moderierende Zusatzbedingungen*, die den sozialen Kontext der SchülerInnen sowohl im Klassengeschehen als auch in der Familie in den Vordergrund stellen, werden von Zielinski als weitere Faktoren benannt (1996, S. 379). SchülerInnen mit einem angenommenen SPF sind bereits in der ersten Klasse einem höheren Ausgrenzungsrisiko ausgesetzt (Krull, Wilbert & Hennemann, 2014), was negativen Einfluss auf die Leistungsentwicklung haben kann (Entwisle, Alexander, Cadigan & Pallas, 1986). Zur Entstehung von Lernschwierigkeiten tragen jedoch auch vorschulische Erfahrungen bei. Ein bildungsfernes Elternhaus mit niedrigem Einkommen und wenig kognitiver Aktivierung kann die Ausgangs- und Startbedingungen für Kinder im Schuleintrittsalter negativ beeinflussen (Euen, Vaskova, Walzenburg & Bos, 2015; Koch, 2007; Weinert & Ebert, 2013). Ein geringes elterliches Unterstützungsverhalten wird auch im Laufe der Schullaufbahn möglicherweise die Entstehung einer Lernschwierigkeit begünstigen (Tiedemann & Faber, 1990).

Die sechs Faktoren, die Zielinksi (1996) beschreibt, bedingen, dass von ihnen betroffene SchülerInnen im schulischen Kontext im Falle eines nicht hinreichend differenzierten Unterrichts nicht gleichermaßen von Lernangeboten profitieren, ihr Wissen und ihre Kompetenzen nicht entsprechend ausbauen können wie ihre KlassenkameradInnen. Aus dieser Konstellation können Lernschwierigkeiten entstehen; diese sind dann gegeben, wenn SchülerInnen in ihrer Leistung zurückbleiben (Kretschmann, 2007). Die Bestimmung des „Zurückbleibens“ ist von der jeweiligen Bezugsnorm abhängig (Zielinski, 1996; siehe auch Abschnitt 2.1.1).

Lernbehinderung gilt als „eine besonders ausgeprägte Form einer Minderleistung bei der absichtsvollen und aktiven Verarbeitung sowie der Abspeicherung von Wissen. Die Einschränkungen zeigen sich in erster Linie beim Erwerb kognitiv-verbaler und abstrakter Inhalte“ (Grünke & Grosche, 2014, S. 76). Meist steht diese Minderleistung, die einen Leistungsabstand von mindestens zwei bis drei Jahren zu gleichaltrigen SchülerInnen ohne Beeinträchtigung bedeutet, in Verbindung mit einer geringeren Intelligenz (bei Lernbehinderung in der Regel ein IQ von 55 bis 85; Grünke & Grosche, 2014). In Deutschland werden diese Kinder und Jugendlichen überwiegend in einem historisch gewachsenen und nach Förderschwerpunkten ausdifferenzierten Förderschulsystem beschult (Autorengruppe Bildungsbericht, 2016; Dietze, 2012). Hier werden sie in kleineren Lerngruppen nach einem speziellen Lehrplan unterrichtet, der ihnen eine adäquate Förderung in der „kognitiven, sprachlichen, emotionalen, und sozialen Entwicklung“ (KMK, 1999, S. 15) zukommen lassen soll. Lediglich ein Viertel dieser Gruppe besucht im Sinne der Inklusion – wie sie mit der Ratifizierung der UN-Behindertenrechtskonvention seit 2009 rechtlich verbindlich ist (Werning, 2010) – eine allgemeinbildende Schule (Dietze, 2012).

Während Bleidick systemkritisch konstatierte, dass „lernbehindert ist, wer eine Schule für Lernbehinderte besucht“ (1998, S. 106), kann nicht davon ausgegangen werden, dass alle SchülerInnen, die an einer Förderschule Lernen unterrichtet werden, auch im Sinne des oben beschriebenen definitorischen Ansatzes eine Lernbehinderung haben und aus diesem Grund mit einem entsprechenden Förderbedarf diagnostiziert wurden. Die Schülerschaft an Förderschulen stellt sich als sehr heterogen dar; die kognitive Leistungsfähigkeit der Kinder kann von durchschnittlich bis hin zu einer leichten geistigen Behinderung beschrieben werden (Bos, Müller & Stubbe, 2010, S. 384). Dementsprechend

heterogen fallen auch die inter- und intraindividuellen Kompetenzprofile von SchülerInnen mit SPF-L aus.

Ergänzend zu Zielinskis sechs Faktoren zur Entstehung von Lernschwierigkeiten müssen an dieser Stelle noch (meta)kognitive Strategien bzw. Lernstrategien der SchülerInnen erwähnt werden. So verfügen SchülerInnen mit SPF-L meist nicht über die entsprechenden Strategien, um ihre Lernprozesse adäquat zu organisieren und ihre Lernzeit effektiv zu nutzen (Lauth et al., 2014). Viele SchülerInnen mit Lernbehinderung zeigen Defizite beim spontanen Einsatz kognitiver Strategien (Torgesen, 1977). Im Besonderen fehlen dieser SchülerInnengruppe metakognitive Kompetenzen, welche für die Regulierung des Lernens und den Erwerb von Kompetenzen notwendig sind (Wong, 1991). So zeigen verschiedene Studien, dass die Kompetenzen von SchülerInnen mit SPF-L geringer ausgeprägt sind als jene der SchülerInnen ohne Beeinträchtigungen. In der Studie KESS7 wurden SchülerInnen im 7. Jahrgang an Hamburger Förderschulen mit Kindern am Ende der Grundschulzeit verglichen. Sowohl im Bereich Leseverständnis als auch im Bereich Mathematik fallen die Leistungen der SchülerInnen an Förderschulen etwa eine halbe bzw. eine ganze Standardabweichung schlechter aus (Bos et al., 2009). Auch zum Ende der Pflichtschulzeit sind deutliche Unterschiede zwischen SchülerInnen mit und ohne SPF-L zu finden (Lehmann & Hoffmann, 2009). Diese Ergebnisse werden durch Analysen der PISA-Daten 2012 bestätigt. Der Großteil der SchülerInnen mit SPF-L an Förderschulen erreichten in keinem der drei Kompetenzbereiche (Lesen, Mathematik, Naturwissenschaften) die Kompetenzstufe I (Gebhardt et al., 2015). Die Vergleichbarkeit der Kompetenzmessung ist jedoch auf Grund von Differential Item Functioning (DIF; Roussos & Stout, 2004) nicht umfassend gegeben, so dass die Ergebnisse mit Vorsicht interpretiert werden müssen (Müller et al., 2017). Insgesamt stützen die empirischen Befunde dennoch die theoretische

Definition, dass SchülerInnen mit SPF-L in ihrer Kompetenzentwicklung etwa zwei bis drei Schuljahre zurückliegen (Grünke & Grosche, 2014).

In der aktuellen gesellschaftlichen Diskussion bezogen auf die Umsetzung eines inklusiven Schulsystems gibt es zunehmend Studien, die die Entwicklung von SchülerInnen mit SPF in segregierten und inklusiven Schulsettings vergleichen. Dabei gibt es Hinweise, dass SchülerInnen mit SPF-L von einer inklusiven Beschulung in einer Klassengemeinschaft mit SchülerInnen ohne Beeinträchtigungen mehr profitieren und höhere Kompetenzwerte erzielen (Kocaj et al., 2014; Wild et al., 2015). Schulsettings können aber auch Einfluss auf sozio-emotionale Variablen haben. So ist das akademische Selbstkonzept von Kindern und Jugendlichen mit Beeinträchtigungen, die eine inklusive Klasse besuchen, geringer ausgeprägt als jener, die eine Förderschule besuchen (Sauer, Ide & Borchert, 2007). Dennoch haben Kinder mit Lernschwierigkeiten und Teilleistungsstörungen eine differenzierte Wahrnehmung ihrer Stärken und Schwächen (Schuchardt et al., 2015). Ob diese verschiedenen Befunde tatsächlich auf die Art der Beschulung bzw. die Klassenkomposition oder möglicherweise auf Selektionsprozesse im Rahmen von Schulentscheidungen zurückzuführen ist, konnte methodisch bisher noch nicht ausreichend differenziert werden.

Zusammenfassend soll betont werden, dass es sich bei SchülerInnen mit SPF-L um eine sehr heterogene Gruppe handelt. Es sind stark ausgeprägte Leistungsspannen (Gebhardt, Oelkrug & Tretter, 2013) sowie intra- und interindividuelle Kompetenzprofile vorzufinden (Bos et al., 2010). Eine hinreichend differenzierte und umfassende Beschreibung dieser Personengruppe bedarf daher der Berücksichtigung einer Vielzahl von Perspektiven und Variablen.

2.2 Kognitive Anforderungen bei Befragungen und Testungen

Neben den inhärenten Anforderungen eines Kompetenztests bzw. Fragebogens werden bei der Administration dieser Instrumente im Rahmen von Large-Scale-Assessments über die reine Bearbeitung der Aufgaben hinaus weitere kognitive Anforderungen an die teilnehmenden Personen gestellt. Diese werden nachfolgend für die jeweiligen Erhebungsformen (Befragung und Testung) beleuchtet, um anschließend Implikationen für die Administration entsprechender Verfahren abzuleiten.

2.2.1 Bearbeitung von Kompetenztests

Im Rahmen großer Schulleistungsstudien oder Large-Scale-Assessments werden Kompetenzen häufig in Anlehnung an die Definition von Weinert (2001) verstanden. Er versteht Kompetenzen als „die bei Individuen verfügbaren oder durch sie erlernbaren kognitiven Fähigkeiten und Fertigkeiten, um bestimmte Probleme zu lösen, sowie die damit verbundenen motivationalen, volitionalen und sozialen Bereitschaften und Fähigkeiten, um die Problemlösungen in variablen Situationen erfolgreich und verantwortungsvoll nutzen zu können“ (Weinert, 2001, S. 27f).

Die Erfassung der Kompetenzen erfolgt in der Regel durch standardisierte Testverfahren mit einer vorgegebenen Testzeit. Die Kompetenzmessung kann entweder fächerübergreifend oder domänenspezifisch sein (Pepper, 2011; Weinert et al., 2011). Auf Grund der Performanz einer Person in einer spezifischen, zeitlich begrenzten Testsituation wird auf ihre zugrundeliegende Kompetenz geschlossen. Diese Annahme birgt jedoch auch Probleme in sich, wie Carlson und Wiedl (2000) beschreiben:

> “If a student fails to solve an arithmetic problem such as ‘‘Mary has 8 apples and Peter brings her 5 more, how many apples will Mary have?’’ the conclusion usually would be that

> he or she cannot add 8 plus 5: But how sure can we be about our conclusion? Perhaps the student knows what 8 + 5 is but does not recognize that the addition operation is required for this word problem. Or perhaps the student cannot read the question and, unknowingly, we have given the student a reading test rather than a math test. Or perhaps the student selected an answer randomly, never reading the question. Again, unknowingly, we may have tested the student's attitude, motivation, ability to follow directions to comply with our expectations or whatever. The point is that for a variety of reasons the student's performance may or may not reflect what we think it does and the conclusions we draw. Accordingly, the inferences made may be inaccurate and misleading, and we have a validity problem." (S. 681)

Die Gefahr, dass Messungen nicht zu validen Ergebnissen führen, ist eine der größten Risiken bei der Erfassung von Kompetenzen bei spezifischen Gruppen, wie auch bei SchülerInnen mit SPF-L. Die Annahmen über und Schilderungen der Kompetenzen und Fähigkeiten von SchülerInnen mit SPF-L, wie in Abschnitt 2.1 beschrieben, beruhen in weiten Teilen darauf, dass Testergebnisse dieser Kinder und Jugendlichen mit denjenigen von SchülerInnen ohne Beeinträchtigung verglichen werden. Basierend auf differenziellen Resultaten und Testscores entsteht die Annahme, dass SchülerInnen mit SPF-L in diesem oder jenem Kompetenzbereich Defizite aufweisen (Torgesen, 1977). Dieses Vorgehen wird aber auch kritisiert, da standardisierte Testverfahren Minderheitengruppen und auch SchülerInnen mit SPF benachteiligen können. Ein schlechteres Testabschneiden kann im Rahmen dieser Argumentation auf vielfältige Gründe zurückgeführt werden (Feuerstein, Rand & Hoffmann, 1979). Es ist jedoch nicht zwingend anzunehmen, dass eine schlechtere Testperformanz die wenig ausgeprägte Kompetenz reflektiert (Tzuriel, 2012). Vielmehr können fehlende Lernstrategien, mangelnde kognitive Grundfähigkeiten, geringe aufgabenbezogene Motivation etc. als interagierende Wirkmechanismen für niedrige Testresultate angenommen werden (Feuerstein et al., 1979). Eine theoretische und konzeptionelle Unterscheidung der Fähigkeit bzw. Kompetenz und der Performanz ist

daher dringend notwendig (Torgesen, 1977). Laut Bortner und Birch (1969) reflektieren die eigenen Fähigkeiten, wenn manifestiert in Performanz, eine Interaktion zwischen den vorhandenen Potenzialen einer Person und den spezifischen Aufgabenanforderungen. In einem Modell zum dynamischen Testen nach Carlson und Wiedl (1992) wird diese Interaktion aufgegriffen. Die Relation zwischen der tatsächlichen Kompetenz (in ihrem Fall Intelligenz) und der Performanz in einem Kompetenztest (bzw. Intelligenztest) kann durch verschiedene Faktoren beeinflusst werden. Die suboptimale Anwendung der vorhandenen Kompetenz, die sich in einer nicht-adäquaten Testperformanz ausdrückt, kann auf Grund verschiedener Faktoren – unter anderem (1) nicht vorhandener Testfairness, (2) geringer Testwiseness, (3) weniger Vertrautheit mit den Aufgaben oder der Testsituation oder (4) weniger ausgeprägter Baseline Reserve Capacity – entstehen (Carlson & Wiedl, 1992). Wenn eine suboptimale Performanz vermutet wird, ist im Rahmen des dynamischen Testens das übergeordnete Ziel, die Testbedingungen so anzupassen, dass eine verbesserte Performanz gelingt, die die tatsächliche Kompetenz der Person reflektiert. Auf die verschiedenen Faktoren, welche die Performanz beeinflussen, wird im Folgenden kurz eingegangen.

(1) *Testfairness* soll die Gleichbehandlung aller Testteilnehmenden im Testprozess garantieren, um eine Verzerrung der Ergebnisse auf Grund unterschiedlicher Administrationsbedingungen der Ergebnisse auszuschalten (American Educational Research Association, American Psychological Association & National Council on Measurement in Education, 1999). Kane (2010) unterscheidet zwei relevante Ebenen für die Evaluation von Testfairness. Die prozedurale Ebene bezieht sich primär auf administrative Aspekte, wie gleichwertige Bedingungen und Inhalte für alle Teilnehmenden (Kane, 2010). Zur Sicherstellung der prozeduralen Testfairness werden in empirischen Studien

standardisierte Vorgehensweisen eingesetzt. Die substantielle Ebene der Testfairness bezieht sich auf die Interpretation der Testergebnisse und daraus abgeleitete Inferenzen und mögliche Bildungsentscheidungen (Kane, 2010). Dies bedeutet, dass Personen mit ähnlicher Ausprägung auf einer Kompetenzdimension den gleichen Testwert zugewiesen erhalten sollten.

Ein zu stark standardisiertes Vorgehen kann jedoch gruppenspezifische Hindernisse bei der Bewältigung von Testanforderungen und -aufgaben in sich bergen, vor allem für SchülerInnen mit Beeinträchtigungen oder SPF (Sireci, Scarpati & Li, 2005). Bestimmte Aspekte oder Vorgehensweisen einer standardisierten Testadministration können zu behinderungsbezogener, aber konstrukt-irrelevanter Varianz für bestimmte Gruppen führen (Sireci et al., 2005). Konstrukt-irrelevante Varianz entsteht durch systematische Messfehler, die mit Eigenschaften einer Person, einer Gruppe oder mit der Anlage eines Tests zusammenhängen (Haladyna & Downing, 2004). Hierbei kann es sich sowohl um Personen mit einer SPF handeln, als auch um Gruppen aus unterschiedlichen Schultypen (Schwabe, McElvany & Trendtel, 2015), Gruppen mit Migrationshintergrund (Haag, Heppt, Stanat, Kuhl & Pant, 2013) oder anderen Sprachhintergründen (vgl. Artelt & Baumert, 2004), die im Rahmen einer standarisierten Testung benachteiligt oder bevorzugt werden. Die Quelle nicht zufälliger Messfehler ist vielfältig und nicht immer identifizierbar, führt in der Konsequenz jedoch zur systematischen Unterschätzung der Fähigkeiten bestimmter Personen und somit zu einer Gefährdung der Validität und Vergleichbarkeit der eingesetzten Maße (Messick, 1995). Um Testfairness zu garantieren, muss die besondere Ausgangslage der SchülerInnen mit SPF-L berücksichtigt werden.

(2) Des Weiteren kann die *Testwiseness* einen Einfluss auf die Performanz haben. Testwiseness ist als jene Fähigkeit einer Person definiert, die Eigenschaften und Formate eines Tests und/oder einer Testsituation effektiv nutzen zu können, um einen hohen Testscore zu erreichen (Millman, Bishop & Ebel, 1965). Diese Fähigkeit ist im Prinzip unabhängig von dem gemessenen Konstrukt. Banks und Eaton (2014) beschreiben daher die besten Testperformer als Personen, die über Wissen verfügen, wie z. B. das Verstehen der Ziel, Restriktionen und Anforderungen eines Tests. Personen mit geringer bzw. ohne Testwiseness sind daher in Testungen benachteiligt gegenüber Teilnehmenden, die über entsprechende Fähigkeiten verfügen (Banks & Eaton, 2014). Die Elemente von Testwiseness beinhalten vor allem Strategien zur effektiven Zeitnutzung, sowie Fehlervermeidung, Raten, bzw. deduktives Ableiten der korrekten Antwort (Millman et al., 1965). Inzwischen gibt es viele Studien, die die Effekte von Schulungen in Testwiseness, sogenannten Testtaking-Programmen, und deren Effekte auf die Testperformanz untersuchen (Brunner, Artelt, Krauss & Baumert, 2007; Hong, Sas & Sas, 2006; Ritter & Idol-Maestas, 1986; Rothman & Cohen, 1988). Interventionen zur Vorbereitung auf Testungen kann auch die Performanz von Personen verschiedener Minderheitspopulationen verbessern (Johns & Vanleirsburg, 1992). Programme für SchülerInnen mit SPF-L, die auf spezifische Strategien im Zusammenhang mit dem Fragestimulus, Itemformat und Zeitmanagement fokussieren, sind erfolgreich (Scruggs, 1984; Scruggs, White & Bennion, 1986). So können auch SchülerInnen mit geringem Leseverständnis von abgestimmten Strategieinstruktionen profitieren (Ritter & Idol-Maestas, 1986).

(3) Auch die *Vertrautheit* der SchülerInnen mit der Testsituation und den administrierten Aufgaben scheint ein relevanter Einflussfaktor bei der Bearbeitung von Tests zu sein. Sind Testaufgaben oder -situationen nicht gänzlich unbekannt, so können Teilnehmende sich

unproblematischer auf die Anforderungen des Tests einlassen. Sowohl spezifische Coaching-Programme als auch Vortests (im Sinne eines Vertrautwerdens mit den Aufgaben) können positive Effekte auf Testergebnisse haben. Das Ausmaß der Steigerung der Testergebnisse ist abhängig davon, wie ähnlich die Aufgaben im Rahmen des Coachings oder Vortests mit den tatsächlichen Testaufgaben sind (Carter et al., 2005; Kulik, Kulik & Bangert, 1984). Wiederum andere Studien mit einem Pre-/Post-Test-Design konnten jedoch nur teilweise signifikante Verbesserungen der Testergebnisse nach einem Vortest mit anschließendem Coaching von SchülerInnen an Gymnasien feststellen (Brunner et al., 2007). Für SchülerInnen mit SPF-L zeigt sich, dass für sie vor allem die ersten Items eines Tests systematisch schwieriger zu lösen sind im Vergleich zu SchülerInnen ohne Beeinträchtigung an Regelschulen (Pohl, Südkamp, Hardt, Carstensen & Weinert, 2016). Dies könnte als Hinweis interpretiert werden, dass die Testsituation oder die zugehörigen Testaufgaben und -formate unbekannt sind und sich die SchülerInnen mit SPF-L erst an die Situation gewöhnen müssen. Weitere Interventionsstudien haben darüber hinaus gezeigt, dass SchülerInnen mit Lernschwierigkeiten nach der Phase einer Vertrautmachung mit den Testmaterialien eine verbesserte Testleistung bei einem Matrizentest zeigten (Hessels, 2009).

(4) Zudem nennen Carlson und Wield (1992) in ihrem Modell die *Baseline Reserve Capacity* als weiteres Element bei der Translation von Kompetenz zu Performanz. Die Baseline Reserve Capacity wird als latentes Leistungspotenzial einer Person verstanden (Neher & Sowarka, 1999). Wenn alle existierenden Reservekapazitäten einer Person genutzt würden, wäre sie in der Lage, in einem Test ihre maximale Performanz zu demonstrieren (Baltes, 1987). Die Gründe, warum Personen ihr maximales Potenzial in Testsituationen nicht abrufen (können), sind vielfältig (Banks & Eaton, 2014). Hierzu gehören z. B. mangelnde

Motivation (Baumert & Demmrich, 2001; Duckworth, Quinn, Lynam, Loeber & Stouthamer-Loeber, 2011; Kukla, 1974), Testängstlichkeit (Bryan, Sonnefeld & Grabowski, 1983; Cassady & Johnson, 2002) oder Bedrohung durch Vorurteile (Good, Aronson & Inzlicht, 2003; Spencer, Steele & Quinn, 1999; Wilbert, 2010b). Für SchülerInnen mit SPF-L kann angenommen werden, dass weitere gruppenspezifische Faktoren existieren, die eine passende Translation der vorhandenen Kompetenz in eine entsprechende Performanz beeinträchtigen.

Zusammenfassend beschreibt das Modell von Carlson und Wiedl (1992) die Relation zwischen der tatsächlichen Kompetenz und der demonstrierten Performanz in einem Kompetenztest, die durch verschiedene Faktoren moderiert wird. Ein jüngeres Theoriemodell greift die spezifischen Anforderungen eines Kompetenztests sowie deren Interaktion mit den Charakteristika der Testteilnehmenden auf (Beddow, Elliott & Kettler, 2013). Die Testung wird als Situation gerahmt, in der sowohl die Anforderungen eines Tests als auch die Fähigkeiten der teilnehmenden Person aufeinandertreffen (siehe Abbildung 1). Die Passung zwischen den Testeigenschaften und den Charakteristiken der Teilnehmenden beschreibt im Sinne der Accessibility Theory (Zugangstheorie) nach Beddow und Kollegen (2013) den Zugang der teilnehmenden Personen zum Instrument. Zugang wird hiernach als das Ausmaß verstanden, in dem ein Test und die zugehörigen Items es den Teilnehmenden erlauben, ihre Kompetenzen bezogen auf das zu messende Konstrukt demonstrieren zu können. In diesem Sinne betont die Accessibility Theory die Relevanz der Zugangsfähigkeiten, welche notwendig sind, um die Anforderungen eines Tests zu verstehen und zu bewältigen. Die Teilnahme an einer Testung erfordert in diesem Sinne nicht nur die Anwendung der definierten kognitiven Fähigkeiten des zu messenden Konstrukts, sondern weitere Fähigkeiten der Person, die Performanz und das Testergebnis

beeinflussen können. Sollten auf Grund fehlender Zugangsfertigkeiten bzw. besonderer Anforderungen eines Tests gruppenspezifische Hindernisse bei der Bearbeitung eines Tests entstehen, ist die Testfairness gefährdet und die Validität des Instruments eingeschränkt (Beddow, Kurz & Frey, 2011).

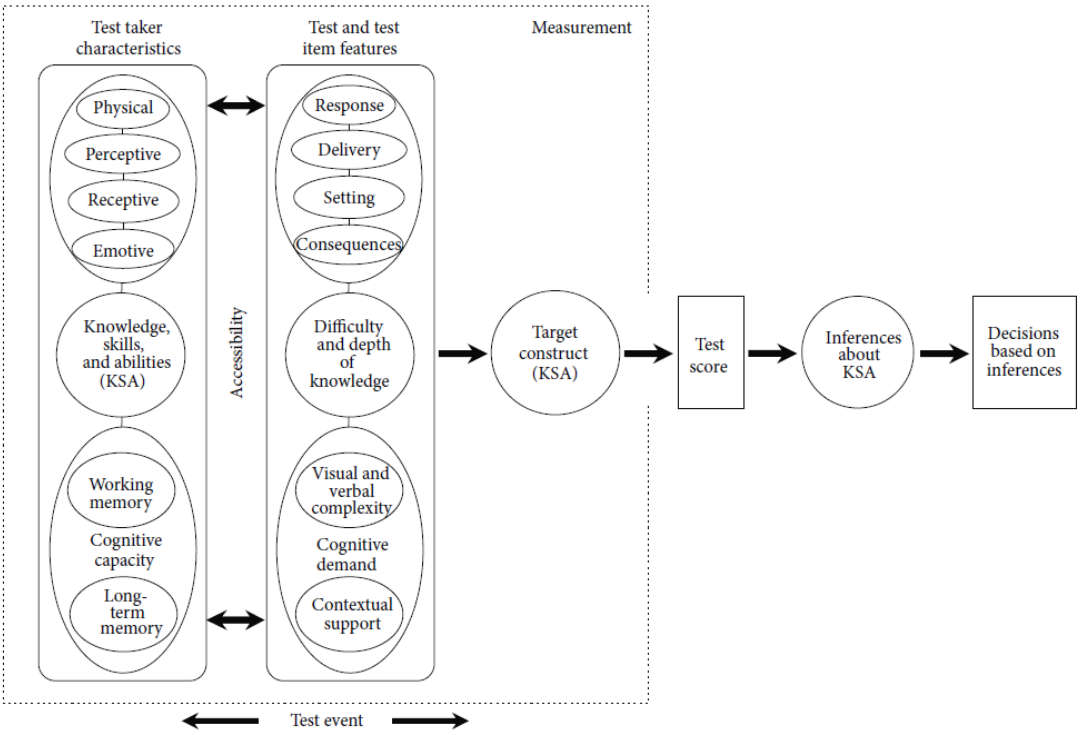

Abbildung 1. Accessibility theory (Beddow et al., 2013, S. 2)

Eine wichtige Zugangsfähigkeit zur Bearbeitung von Testaufgaben ist das Verständnis von Testinstruktionen (Abedi, 2011). Für SchülerInnen mit SPF-L kann vermutet werden, dass dieser Aspekt eine wichtige Bedeutung hat, da auf Grund ihrer Ausgangslage das Verständnis sprachlicher Instruktionen eingeschränkt sein kann (Zielinski, 1996). Somit stellen versprachlichte Testinstruktionen die Schülerschaft mit SPF-L vor eine größere Herausforderung (Hessels-Schlatter, 2002). Im Sinne der Testfairness wird für Testinstruktionen gefordert, dass diese unabhängig von Vorkenntnissen oder Sprachkenntnissen der Teilnehmenden sein sollen. Testinstruktionen müssen eindeutig,

konsistent und in verständlicher Sprache formuliert sein, so dass die teilnehmenden Personen die Aufgaben entsprechend der Intentionen der TestentwicklerInnen bearbeiten können (American Educational Research Association et al., 1999). So erweitern Beddow und Kollegen (2013) mit ihrem Modell den theoretischen Diskurs um einen wichtige Faktoren, die insbesondere in Bezug auf SchülerInnen mit SPF-L Beachtung finden sollten.

Nachdem die Anforderungen bei der Bearbeitung von Kompetenztests beschrieben wurden, geht es nachfolgend um die besonderen Herausforderungen bei Befragungen und die damit verbundenen kognitiven Anforderungen.

2.2.2 Beantwortung von Fragebögen

Ein Großteil gewonnener Daten in der Bildungsforschung basiert auf Befragungen (Bortz & Döring, 2005). Die Voraussetzung für schriftliche Befragungen ist die Lese- und teilweise auch Schreibfähigkeit der teilnehmenden Personen (Diersch & Walther, 2010). Darüber hinaus ist eine Befragung ein kognitiv aufwendiger Prozess. Während dieses Prozesses können unterschiedlichste Fehlerquellen die Selbstauskünfte beeinflussen oder verzerren (Strack, 1994). Bisher gibt es eine Vielzahl an Modellen, welche die kognitiven Anforderungen, die im Rahmen einer Befragung von den teilnehmenden Personen durchlaufen und bewältigt werden müssen, darstellen (Jobe & Hermann, 1996). Die meisten dieser Modelle haben eine gemeinsame Grundlage von insgesamt vier Phasen, welche sich im von Tourangeau, Rips und Rasinski (2000) vorgeschlagenen Modell wiederfinden lassen (Jobe, 2003). Tourangeau und Kollegen (2000) nehmen an, dass befragte Personen insgesamt vier kognitive Prozessschritte durchlaufen, die zu akkuraten Antworten führen (siehe Abbildung 2). Bei jedem dieser Schritte kann es im kognitiven

Antwortverhalten zu stufenspezifischen Fehlern kommen, die einer reliablen und validen Beantwortung einer Frage zuwiderlaufen.

Component	Specific Processes
Comprehension	Attend to questions and instructions Represent logical form of question Identify question focus (information sought) Link key terms to relevant concepts
Retrieval	Generate retrieval strategy and cues Retrieve specific, generic memories Fill in missing details
Judgment	Assess completeness and relevance of memories Draw inferences based on accessibility Integrate material retrieved Make estimate based on partial retrieval
Response	Map judgment onto response category Edit response

Abbildung 2. Die vier Komponenten des Antwortprozesses (Tourangeau et al., 2000, S. 8)

Comprehension: An erster Stelle steht zunächst das Verständnis der Fragen. Es ist erforderlich, dass der Frage und der damit verbundenen Instruktion Aufmerksamkeit geschenkt wird. Der Kerngedanke der Frage muss inhaltlich verstanden, d.h. kognitiv erfasst werden. Hierzu gehört es auch, dass ggf. unbekannte Worte, Fachausdrücke, grammatikalisch anspruchsvolle oder vage Formulierungen erkannt und entsprechend dekodiert bzw. aus dem Kontext erschlossen werden (Tourangeau et al., 2000). Für die Befragung von Kindern werden deshalb vor allem einfache Worte und eine wenig komplexe Grammatik für die Fragenformulierung empfohlen (Bell, 2007; de Leeuw, 2011). Auch für Kinder ohne Beeinträchtigung ist an dieser Stelle eine hinreichende Lesefähigkeit eine wichtige Voraussetzung, um die Zielsetzung der Frage korrekt zu verstehen. So können SchülerInnen, die über eine höhere Lesekompetenz verfügen, eher valide Antworten im Vergleich zu SchülerInnen mit niedrigeren Lesefähigkeiten abgeben, wenn sie beispielsweise zu ihren persönlichen Einstellungen befragt werden (Borgers et al., 2000).

Besonders mehrdeutige oder vage Formulierungen können dazu führen, dass Kinder die Inhalte der Fragen nicht im intendierten Sinne verstehen (Holaday & Turner-Henson, 1989; Robinson & Whittaker, 1987). Auch Items mit negierenden Aussagen stellen Kinder vor eine größere Herausforderung, so dass die Antworten weniger reliabel und valide ausfallen (Borgers et al., 2000; Marsh, 1986). Beruhend auf diesen empirischen Befunden kann angenommen werden, dass es bei SchülerInnen mit SPF-L auf dieser ersten Prozessstufe bereits zu weiterführenden Schwierigkeiten kommen kann. Auf Grund einer häufig weniger ausgeprägten Lesekompetenz dieser SchülerInnen (Gebhardt et al., 2015) gelingt ihnen das eigenständige Lesen von Fragen und zugehörigen Anleitungshinweisen möglicherweise nicht gleichermaßen gut und flüssig wie Kindern und Jugendlichen ohne Beeinträchtigungen. Somit kann es auf Grund der eingeschränkten Lesefähigkeit sowie eines reduzierten Wortschatzes bereits beim Dekodieren der Fragen zu Einschränkungen hinsichtlich des Verständnisse der Inhalte und Ziele der Fragestellungen kommen.

Retrieval: In einem nächsten Prozessschritt müssen die relevanten Informationen zur gestellten Frage aus dem Gedächtnis abgerufen werden (Tourangeau et al., 2000). Die erkannten Schlüsselworte werden zu vorhandenen Konzepten verbunden, individuelle Erlebnisse oder das eigene Wissen werden zum befragten Thema abgerufen. Die Passung zwischen den genutzten Begriffen der Frage und den eigenen konzeptionellen Vorstellungen eines Sachverhalts kann hier entscheidend sein, damit die angemessenen bzw. erwünschten Informationen abgerufen werden können. Bereits an dieser Stelle werden laut Modell mögliche Lücken mit weiteren Details gefüllt, damit sich in einem nächsten Schritt ein Urteil herauskristallisieren lässt. Je komplexer eine Frage, desto schwieriger ist es, die spezifischen Schlüsselwörter zu identifizieren und mit vorhandenen Konzepten zu verknüpfen. In diesem Zusammenhang spielt auch die Kapazität des

Arbeitsgedächtnisses eine Rolle, da relevante Informationen über den Prozess der Fragebeantwortung hinweg durchgehend präsent sein müssen. Ob korrespondierende Konzepte für bestimmte Inhalte bereits bei Kindern und Jugendlichen vorhanden sind, hängt an unterschiedlichen Faktoren. So unterscheiden sich verschiedene Altersgruppen stark hinsichtlich des Ausmaßes an Aufmerksamkeit und Interesse, das sie für bestimmte Inhalte oder Verhaltensweisen aufbringen (Schwarz, 2003). Vor allem jüngere Personen setzen sich eher mit einer Frage auseinander, wenn sie diese verstehen und ein gewisses Interesse für deren Inhalte aufbringen (Holaday & Turner-Henson, 1989). Aber auch die Erfahrungswelt der Kinder und Jugendlichen spielt eine wichtige Rolle. So können z. B. Fragen zum Beruf der Eltern nicht immer korrekt beantwortet werden, da dieser Bereich häufig nicht salient ist (Lipski, 2000; Looker, 1989). Für SchülerInnen mit SPF-L stellt diese kognitive Prozessstufe des Abrufens von Informationen große potenzielle Schwierigkeiten dar. So kann mit Sicherheit angenommen werden, dass für bestimmte Fragen keine entsprechenden Konzepte vorhanden sind, da SchülerInnen mit SPF-L insgesamt über eine geringere Wissensbasis verfügen als SchülerInnen ohne Beeinträchtigungen (siehe Abschnitt 2.1). Zudem muss davon ausgegangen werden, dass diese Kinder und Jugendlichen über eine geringe Arbeitsspeicherkapazität verfügen, so dass die verschiedenen Inhalte im Rahmen des Retrievals eventuell nicht alle gleichermaßen abgerufen werden können.

Judgment: Das Abrufen von Informationen bringt meist noch keine eindeutige Antwort auf eine Frage hervor, so dass nun in einem nächsten Prozessschritt die Informationen miteinander verbunden und ergänzt werden müssen (Tourangeau et al., 2000). Hierfür können verschiedene Strategien gewählt werden, je nachdem, um welche Art von Fragen es sich handelt. Zunächst muss jeweils die Korrektheit und Vollständigkeit der abgerufenen

Informationen beurteilt werden, um anschließend Inferenzen zwischen den erhaltenen Informationen zu ziehen, welche etwaige Lücken füllen können. Basierend auf einer Gesamtintegration aller dann vorliegenden Daten wird ein Urteil gefällt und somit in eine passende Antwort übertragen. Besonders herausfordernd können Fragen zu Häufigkeiten, Ereignissen oder Zeiträumen sein. Kinder haben noch eine sehr unterschiedliche Wahrnehmung von Zeiträumen und Entfernungen, so dass dies Einfluss auf eine valide Fragenbeantwortung haben kann (Kränzl-Nagl & Wilk, 2000). Für SchülerInnen mit SPF-L kann darüber hinaus die Integration verschiedener Informationen ein weiteres Hindernis innerhalb dieses kognitiven Prozesses darstellen.

Response: Nach dem Finden der individuell richtigen Antwort zur Frage, muss diese noch auf die vorhandenen Antwortoptionen abgebildet und angepasst werden (Tourangeau et al., 2000). Die angebotene Skala bzw. unterschiedliche Arten von Antwortoptionen können einen starken Einfluss auf das Antwortverhalten haben (Krosnick et al., 2001; Saris, Revilla, Krosnick & Shaeffer, 2010). Die Labels der Antwortoptionen können hier einen entscheidenden Einfluss ausüben: So macht es für Kinder einen entscheidenden Unterschied, ob die Antwortoptionen komplett oder teilweise mit vagen Labels versehen sind im Vergleich zu klar formulierten Antwortoptionen. Zwischen drei diesbezüglichen Experimentalbedingungen konnte im Mehrgruppenvergleich keine Messäquivalenz festgestellt werden (Borgers, Hox & Sikkel, 2003). Auch das Problem fehlender Werte kann an dieser Stelle ergänzend angefügt werden. Vor allem Kinder mit einer geringeren Lesefähigkeit zeigen einen erhöhten Anteil fehlender Werte in einer Befragung als SchülerInnen mit höheren Lesefähigkeiten (Borgers et al., 2000). Die Anzahl der Antwortoptionen kann diesbezüglich eine Rolle spielen. Während eine höhere Menge an Antwortoptionen zwar zur Reliabilität des Instruments beiträgt (Alwin, 1997; Lozano,

García-Cueto & Muñiz, 2008), kann dies auch eher zu Nichtbeantwortung (Item-Non-Response) führen (Borgers & Hox, 2001).

Somit stellt der Prozess der Befragung und Generierung einer validen Antwort durch die befragte Person eine komplexe Angelegenheit dar, die sich durch eine Vielzahl von kognitiven Prozessschritten auszeichnet, auf die jeweils item- oder personenspezifische Eigenschaften eine Wirkung ausüben. Jedoch kann nicht zwingend angenommen werden, dass die Befragten alle vier beschriebenen Prozessschritte dezidiert durchlaufen. Ob und welche dieser kognitiven Anstrengungen unternommen werden, hängt auch von der Bereitschaft der Personen ab, sowie ihrer Einstellung bezüglich dessen, wie akkurat ihre abgegebene Antwort sein soll (Tourangeau et al., 2000). Krosnick (1991, 2000) hat für dieses Szenario das Konzept des *Satisficing* entworfen, welches annimmt, dass von den Befragten nicht die optimale Antwort, sondern die erste akzeptable Antwort ausgewählt wird. Krosnick (1991) spricht von einem schwachen Satisficing, wenn die Prozessschritte Retrieval und Judgement nur teilweise oder verzerrt durchlaufen werden. Ein starkes Satisficing beinhaltet, dass diese beiden Schritte gänzlich übersprungen werden und die erste passende Antwort auf Grund einzelner Schlüsselworte gewählt wird. Satisficing kann aus unterschiedlichen Gründen stattfinden, z. B. um das Lesen einer Antwortliste zu vermeiden. Vor allem Kinder tendieren dazu, Satisficing-Strategien zu nutzen, wenn sie die gestellten Fragen nicht richtig verstehen oder wenig Interesse für den Inhalt der Frage haben (Holaday & Turner-Henson, 1989). Es wird angenommen, dass das Auftreten von Satisficing durch drei Faktoren bestimmt wird: (1) Schwierigkeit der Fragen oder Aufgabe, (2) Fähigkeit der Befragten und (3) Motivation der Befragten. Je schwieriger die Frage bzw. Aufgabe, desto höher muss, laut Krosnick (2000), die Fähigkeit und Motivation der teilnehmenden Personen sein, damit die Wahrscheinlichkeit, dass es zum Satisficing

kommt, sinkt. Das Satisficing-Modell von Krosnick (1991, 2000) deutet darauf hin, dass es im Prozess der Befragung zu einer Interaktion zwischen dem Befragungsinstrument und der befragten Personen kommt. Dieses Wechselspiel zwischen Anforderungen des Befragungsinstruments und Eigenschaften der Zielpersonen wird im *General Model of the Survey Interaction Process* aufgegriffen (Esposito & Jobe, 1991). In diesem Modell werden insgesamt drei Komponenten als relevant für den Befragungsprozess erachtet: (1) die Rahmenbedingungen der Befragung, (2) die Eigenschaften der Teilnehmenden sowie (3) die Interaktion zwischen InterviewerIn und den Teilnehmenden. Für die Komponenten der Teilnehmenden werden in diesem Modell nicht nur motivationale und kognitive Aspekte wie bei Krosnick (2000) berücksichtigt, sondern darüber hinaus auch der sozioökonomische Status, die Intelligenz, die Gesundheit und das Erschöpfungslevel, sowie das Interesse der Zielperson für die befragten Themen.

Zusammenfassend kann auch für den Bereich der Fragebogenbeantwortung gesagt werden, dass eine Vielzahl an Faktoren zusammenspielen und einen Einfluss auf den Prozess der Befragung und somit auf die Validität der generierten Antworten haben.

Die vorgestellten Modelle, die zum einen die Herausforderungen bei der Translation der Kompetenz in Performanz (Abschnitt 2.2.1) und zum anderen die kognitiven Prozessschritte bei der Beantwortung von Fragebögen (Abschnitt 2.2.2) konkretisieren, weisen verschiedene Aspekte auf, die jeweils auch für andere Erhebungsform relevant sind. Diese Überschneidungen und Übertragbarkeiten sollen im Folgenden beschrieben werden.

2.2.3 Unterschiede und Gemeinsamkeiten von Kompetenztests und Fragebögen

Bildungsrelevante Konstrukte können mit Hilfe unterschiedlicher Ansätze und Instrumente erfasst werden. Gängige Verfahren in Large-Scale-Assessments sind die Befragung und Testung der teilnehmenden Personen, da es sich um ökonomische Verfahren handelt (z. B. im Vergleich zu Beobachtungsverfahren). Je nach Ziel und Inhalt der Konstrukte werden die jeweiligen Instrumente entwickelt und gestaltet. Während bei Befragungen eher nach den individuellen Lebenswelten, der Herkunft, den Einschätzungen und Erfahrungen der Personen gefragt wird, wird in Kompetenztests nach „korrekten Antwort" in einem objektiven Sinne gefragt, welche die zugrundeliegende Fähigkeit erfasst. Ein Kompetenztest erfasst demnach die Performanz in einer spezifischen Testsituation. Eine Testsituation zeichnet sich dadurch aus, dass in einer vorgegebenen Testzeit eine bestimmte Anzahl an Aufgaben bearbeitet werden soll; korrekt bearbeitete Items dienen als Grundlage für die Schätzung der Kompetenz. Im Falle eines Geschwindigkeitstests kann die Zeit sehr kurz bemessen sein, so dass es in der Regel nicht möglich ist, dass die Teilnehmenden alle Aufgaben bearbeiten. Zeitbegrenzungen sind bei Befragungen zwar üblicherweise auch vorgesehen, sind aber so kalkuliert, dass die Beantwortung aller Fragen durch die Teilnehmenden möglich sein sollte.

Für beide Instrumentarien sind sogenannte Zugangsfähigkeiten (Beddow et al., 2011) notwendig, die jedoch je nach Instrument und zugehörigem Ziel unterschiedliche Gestalt aufweisen können. Sowohl für Kompetenztests als auch für Fragebögen muss ein Instruktionsverständnis vorhanden sein, damit die Art der Aufgabenbearbeitung als auch Fragenbeantwortung in intendierter Weise erfolgen kann. Lesefähigkeit und

Textverständnis sind notwendig, um die Aufgaben und Fragen eigenständig lesen und verstehen zu können.

Das Modell von Carlson und Wiedl (1992) zeigt Aspekte auf, die auch bei der Entwicklung und Beantwortung von Fragebögen eine wichtige Rolle spielen können. Im Sinne einer Testfairness ist es auch für die Gestaltung von Fragebögen erforderlich, bestimmte Personen oder Personengruppen bei der Bearbeitung nicht zu benachteiligen, damit die Erfassung valider Daten aller teilnehmenden Personen gewährleistet ist. Dies kann sowohl für die Formulierung der Fragen, der Salienz der Inhalte als auch bei der visuellen Gestaltung des Fragenbogens eine zentrale Rolle spielen. Testwiseness bzw. hilfreiche Strategien bei der Bearbeitung von Kompetenztests können bei der Beantwortung von Fragebögen eine andere Form annehmen. Die Anwendung bestimmter Strategien zur Reduzierung der kognitiven Belastung bei der Beantwortung von Fragebögen hat Krosnick (2000) mit dem Konzept des Satisficing vorgestellt. Im Prinzip sind jedoch gezielte Strategien, wie sie für die Verbesserung von Testergebnissen sinnvoll sind, bei der Bearbeitung von Fragebögen nicht notwendig.[3] Ebenso wie bei Kompetenztests kann die Vertrautheit bei der Beantwortung von Fragen eine Rolle spielen. Sind die teilnehmenden Personen bereits mit den Fragen vertraut oder haben sie sich bereits früher mit den Inhalten auseinandergesetzt bzw. eine Meinung zu einem bestimmten Sachverhalt gebildet, ist die Beantwortung der Frage leichter (Holaday & Turner-Henson, 1989; Lipski, 2000; Looker, 1989). Die Validität der Angaben kann bei komplexen Fragen eingeschränkt sein, wenn es den Teilnehmenden auf Grund eingeschränkter kognitiver Kapazitäten nicht

[3] Je nach Modus (PAPI vs. Face-to-Face-Interview) und Inhalte kann es bei Befragungen zu Tendenzen der sozialen Erwünschtheit kommen, die ebenfalls die Validität der Ergebnisse beeinflussen (Holbrook, Green & Krosnick, 2003; Krumpal, 2013).

möglich ist, alle relevanten Aspekte und Konzepte, welche für eine korrekte Beantwortung einer Frage notwendig wären, gleichzeitig präsent zu haben (Holaday & Turner-Henson, 1989; Krosnick, 2000).

Die erforderlichen kognitiven Prozessschritte bei der Beantwortung eines Fragebogens (Tourangeau et al., 2000) sind in ähnlicher Form auch bei der Bearbeitung von Kompetenztests notwendig. Geht man von dem häufig genutzten Multiple-Choice-Format als Aufgabe aus (z. B. bei Matrizentests oder auch domänenspezifischen Tests), ist es notwendig, dass auch der Aufgabenstamm sowie die einzelnen Antwortoptionen ebenso verstanden werden, wie es bei einem Fragebogen der Fall ist. Anschließend müssen die relevanten Konzepte, ob zur Identifikation logischer Abfolgen, mathematische Rechenoperationen oder Informationen eines gerade gelesenen Textes, abgerufen und entsprechend ihrer hinreichenden Passung sowie der korrekten Anwendung beurteilt werden. Abschließend muss die aufbereitete Antwort in der Regel auf eine der vorgegebenen Antwortoptionen übertragen werden. Bei einer Kompetenztestung können in diesem Prozess häufiger Testwiseness-Strategien zur Anwendung kommen, wie z. B. über Ausschlussprinzip einzelner Antwortoptionen die Ratewahrscheinlichkeit erhöhen (Millman et al., 1965). Dieses Vorgehen würde im übertragenen Sinne dem Satisficing (Krosnick, 2000, 1991) entsprechen, da nicht mehr alle kognitiven Prozessschritte wie der Abruf relevanter Konzepte und die Beurteilung der Vollständigkeit der Informationen durchlaufen werden, sondern eine leicht zu identifizierende, sinnvolle Antwortoption ausgewählt wird.

2.3 Implikationen für Befragung und Testung von SchülerInnen mit SPF-L

Die vorgestellten Ausgangsbedingungen der SchülerInnen mit SPF-L in Verbindung mit den theoretischen Modellen, die darlegen, welche kognitiven Anforderungen von den

teilnehmenden Personen in einer Testung oder Befragung bewältigt werden müssen, zeigen auf, dass die valide und vergleichbare Erfassung bildungsrelevanter Konstrukte bei dieser spezifischen Zielgruppe kein triviales Unterfangen ist. Behinderungsbezogene Einschränkungen sind bei der Bewältigung der kognitiven Prozesse und Anforderungen während der Befragung und Testung ganz unabhängig von den zu messenden Konstrukten zu erwarten. Daher gilt es, die Besonderheiten der SchülerInnengruppe mit SPF-L in entsprechenden Designelementen aufzugreifen und sie aus der Erfassung der bildungsrelevanten Konstrukte auszupartialisieren. Um solche konstrukt-irrelevanten Varianzen zu verringern, werden in internationalen Studien sogenannte Akkommodationen implementiert (Koretz & Barton, 2004; Pitoniak & Royer, 2001). Eine Vielzahl an Anpassungen, welche für die jeweils spezifischen Situationen der zu testenden Kinder bereitgehalten werden, sollen eine faire und vergleichbare Testung der SchülerInnen gewährleisten (Sireci et al., 2005). Neben einer Reduktion von Testaufgaben (z. B. mit zu hoher Itemschwierigkeit), wird diesen Schülerinnen und Schülern teilweise mehr Testzeit zur Verfügung gestellt. Für SchülerInnengruppen mit kognitiven Einschränkungen werden auch Out-of-Level-Aufgaben (konzipiert für jüngere SchülerInnen) oder Read-Aloud-Akkommodationen (lautes Vorlesen der Testaufgaben) angeboten (Koretz & Barton, 2004). Bei der Implementation von Akkommodationen soll das zu messende Konstrukt jedoch nicht verändert werden und die Vergleichbarkeit nicht eingeschränkt sein (Cormier et al., 2010). Neben der Validität als zentrales Gütekriterium eines Tests oder einer Skala, erweitert die Messäquivalenz die Güte der Daten um spezifische parametrische Annahmen einzelner Items (gleiche Faktorladungen, Intercepts etc.; Steinmetz, Schmidt, Tina-Booh, Wieczorek & Schwartz, 2009) und stellt somit sicher, dass bestimmte Personen oder Personengruppe nicht bei der Testung oder Befragung

durch itemspezifische Charakteristika benachteiligt werden. Messinvarianz belegt statistisch, dass Skalen auf gleiche Art und Weise bei unterschiedlichen SchülerInnengruppen funktionieren. Zu diesem Thema gibt es heterogene Forschungsbefunde, die eine abschließende Beurteilung der Effekte verschiedener Anpassungen noch nicht ermöglichen. Aktuell werden diese vielmehr zielgruppenspezifisch diskutiert (Pitoniak & Royer, 2001). Zentral für die Implementation von Akkommodationen ist zudem die *Differential Boost Hypothesis* (Fuchs & Fuchs, 2001). Diese Annahme geht davon aus, dass von Anpassungen nur jene SchülerInnen profitieren dürfen, die eine Beeinträchtigung haben. Würden auch SchülerInnen ohne Beeinträchtigungen auf Grund von Akkommodationen ein besseres Testergebnis erzielen, wäre der Vorteilsausgleich hinfällig und die implementierten Anpassungen nicht gültig.

Inzwischen gibt es jedoch auch den Ansatz des *Universal Designs*, welches das Prinzip verfolgt, die Erhebungen und Testungen so zu gestalten, dass sie für möglichst viele, wenn nicht sogar alle, SchülerInnen zugänglich und anwendbar sind (Beddow et al., 2011). Durch die Anwendung verschiedener Konstruktionsprinzipien soll somit auf zielgruppenspezifische Anpassungen verzichtet werden können (Thompson, Johnstone & Thurlow, 2002). Einer der konkreten Ansätze ist das *Accessibility and Modification Inventory* (TAMI; Beddow, Elliott & Kettler, 2009; Beddow et al., 2013), welches ein hilfreiches Instrument darstellt, um die Zugänglichkeit verschiedener Aspekte der Testaufgaben (z. B. Stimulus, Layout, Antwortoptionen) zu beurteilen. In einem iterativen Prozess können Items wiederholt überarbeitet werden, um den Einfluss der Zugangsfähigkeiten zu reduzieren und somit die Validität des Instruments nicht zu gefährden (Beddow et al., 2013).

3 Forschungsfragen

In der vorliegenden Arbeit wird untersucht, wie eine valide und mit anderen SchülerInnengruppen vergleichbare Erfassung bildungsrelevanter Konstrukte bei SchülerInnen mit SPF-L umgesetzt werden kann. Der Ansatz dieser Arbeit knüpft an aktuelle Diskussionen an, die sich damit auseinandersetzen, ob und wie SchülerInnen mit SPF in Large-Scale-Studien aussagekräftig integriert werden können (Heydrich et al., 2013; Kuhl et al., 2015; Müller et al., 2017). Bezüglich der hier thematisierten SchülerInnengruppe von Kindern und Jugendlichen mit SPF-L an Förderschulen stellt die reliable, valide und vergleichbare Erfassung bildungsrelevanter Konstrukte eine große Herausforderung dar. Dies begründet sich zum einen in der heterogenen Gruppe, die mit stark ausgeprägten Leistungsspannen sowie intra- und interindividuellen Kompetenzprofilen beschrieben wird (Bos et al., 2010). Zum anderen müssen die Ausgangsbedingungen der SchülerInnen mit SPF-L, wie sie im Abschnitt 2.1 dargelegt wurden, in der Darbietung und Administration der Instrumente Berücksichtigung finden. Behinderungsbezogene Einschränkungen sind bei der Bewältigung der kognitiven Anforderungen während der Befragung und Testung ganz unabhängig von den zu messenden Konstrukten zu erwarten (siehe Abschnitt 2.2).

Während im Rahmen (inter)nationaler Schulleistungsstudien bisher vor allem Inhalte von Testverfahren und deren Administrationsbedingungen angepasst wurden, sollen in dieser Arbeit darauf aufbauend weitere Aspekte fokussiert werden. Zunächst wird die Kompetenztestung bei SchülerInnen mit SPF-L thematisiert. Der Schwerpunkt hierbei liegt auf der Anpassung der Testbedingungen. Im Sinne der Zugangsfertigkeiten (siehe Abschnitt 2.2.1), die für konstrukt-irrelevante Varianz verantwortlich sein können, soll für die Stichprobe von SchülerInnen mit SPF-L an dem frühen Punkt der Testinstruktion angesetzt werden. Testinstruktionen, die als – in der Regel vorgelesene – Informationen

über den Testinhalt, die Zeitvorgaben sowie die korrekte Bearbeitungsweise der nachfolgenden Testaufgaben aufklären sollen, sind relevant für den Zugang zu den Anforderungen einer Testsituation und -aufgabe. Diese sprachlichen Informationen müssen von der Zielperson verarbeitet, gespeichert und umgesetzt werden, um darauf basierend den Test adäquat bearbeiten zu können.

Die Erfassung verschiedener Variablen mittels schriftlichem Fragebogen unter Zuhilfenahme international anerkannter und standardisierter Itembatterien wurde bisher für SchülerInnen mit SPF-L wenig beleuchtet. Auch wenn inzwischen einige Studien bezüglich der Befragung von Kindern und Jugendlichen ohne spezifische Beeinträchtigungen sowie abgeleitete Designempfehlungen existieren (vgl. Bell, 2007; Borgers et al., 2000; Diersch & Walther, 2010), so sind die spezifischen Herausforderungen bei der Gruppe von SchülerInnen mit SPF-L bisher wenig systematisch untersucht worden (z. B. Schwinger et al., 2015).

Aus dem breiten Spektrum offener Fragen wurden vor dem Hintergrund der dargestellten Theorien (Abschnitt 2.2) für diese Arbeit die folgenden Forschungsfragen ausgewählt, die beispielhaft an bestimmten Inhalten analysiert und veranschaulicht werden:

1) Kann der Zugang zu einem Instrument und somit die Testperformanz oder Ausfüllgenauigkeit der SchülerInnen mit SPF-L durch angepasste Instruktionen bzw. Darbietungsmodi positiv beeinflusst werden?

2) Können bildungsrelevante Konstrukte bei SchülerInnen mit SPF-L valide erfasst werden?

3) Sind die erfassten Daten bildungsrelevanter Konstrukte von SchülerInnen mit SPF-L vergleichbar mit jenen anderer SchülerInnengruppen?

Für alle drei genannten Fragestellungen werden Daten der Machbarkeitsstudien an Förderschulen Lernen (Heydrich et al., 2013) im Rahmen des NEPS (Blossfeld, Roßbach & von Maurice, 2011) genutzt. Die Daten stammen aus drei Altersbereichen: Grundschulalter (3./4. Klasse), Beginn sowie Ende der Sekundarstufe I (5. und 9. Klasse)[4]. Zur Untersuchung der Fragestellungen wurden experimentelle Designs implementiert, welche die Evaluation verschiedener Anpassungen und Interventionen erlauben.

In der ersten Fragestellung geht es um die kognitiven Anforderungen, die an teilnehmende Personen bei einer Testung oder Befragung gestellt werden. Diese Anforderungen müssen nicht zwingend Teil des zu erfassenden Konstruktes sein, sondern können bestimmte Zugangsfertigkeiten voraussetzen. Sind diese nicht vorhanden oder werden nicht angewendet, entsteht konstrukt-irrelevante Varianz. Dies beginnt mit den Testinstruktionen bzw. Ausfüllhinweisen, die von den Studienteilnehmenden aufgenommen, verstanden und angewendet werden müssen. Das Verständnis sprachlicher Anweisungen kann ein zielgruppenspezifisches Hindernis für SchülerInnen mit SPF-L darstellen, ist jedoch in der Regel nicht Teil des zu messenden Konstrukts. Somit kann das Verständnis von sprachlichen Instruktionen eine zusätzliche Fähigkeitsdimension darstellen, deren Einfluss auf den Kompetenzscore jedoch verhindert werden sollte. Gleiches gilt bei schriftlichen Befragungen. Wenn Fakten oder auch subjektive

[4] Diese Arbeit nutzt Daten des Nationalen Bildungspanels (NEPS): Startkohorte Klasse 5 (doi:10.5157/NEPS:SC3:1.0.0; doi:10.5157/NEPS:SC3:2.0.0), Startkohorte Klasse 9 (doi:10.5157/NEPS:SC4:1.0.0) und Entwicklungsstudie A54. Die Daten des NEPS wurden von 2008 bis 2013 als Teil des Rahmenprogramms zur Förderung der empirischen Bildungsforschung erhoben, welches vom Bundesministerium für Bildung und Forschung (BMBF) finanziert wurde. Seit 2014 wird NEPS vom Leibniz-Institut für Bildungsverläufe e.V. (LIfBi) an der Otto-Friedrich-Universität Bamberg in Kooperation mit einem deutschlandweiten Netzwerk weitergeführt.

Einschätzungen erfragt werden, sollte die Beantwortung unabhängig von weiteren, möglicherweise erschwerenden Anforderungen sein, wie z. B. der Lesefähigkeit der Befragten. Im Sinne konstrukt-irrelevanter Akkommodationen wurden in dieser Arbeit Anpassungen untersucht, welche die gruppenspezifischen Hindernisse der SchülerInnen mit SPF-L ausgleichen sollen, ohne das zu messende Konstrukt zu verändern. Es wurde angenommen, dass der Zugang zu einem Instrument für SchülerInnen mit SPF-L erleichtert werden kann, wenn die notwendigen Zugangsfähigkeiten durch entsprechende Maßnahmen und Interventionen aufgegriffen und deren Einfluss reduziert wird. So ist die Aufbereitung von Testinstruktionen oder Anleitungen für Fragebögen so zu gestalten, dass eingeschränkte Sprach- und Merkfähigkeiten berücksichtigt werden. Wenn die sprachliche Komplexität des Anleitungstextes reduziert sowie die Bearbeitungshinweise zusätzlich mit visuellen Demonstrationen und entsprechenden Beispielitems unterstützt werden, wurde erwartet, dass diese Maßnahmen konstrukt-irrelevante Varianz bei der Erfassung bildungsrelevanter Konstrukte bei SchülerInnen mit SPF-L reduzieren und der Zugang zu den Instrumenten verbessert wird. Auch der Darbietungsmodus der Befragungsinstrumente wurde angepasst und alle Items und Antwortoptionen mit Hilfe eines standardisierten Skriptes vorgelesen (Read-Aloud), um eine geringere Lesefähigkeit der SchülerInnen auszugleichen und das Verständnis der Fragen zu verbessern. Die Untersuchung der „Verbesserung des Zugangs“ wurde, je nach Konstrukt, durch bestimmte Testindikatoren (Anzahl korrekter Lösungen oder Anzahl nicht-instruktionskonformer Antworten) sowie die Validität der Befragungsdaten evaluiert.

Ob die vorher beschriebenen Anpassungen eine valide Erfassung bildungsrelevanter Konstrukte ermöglichen, soll in der zweiten Fragestellung untersucht werden. Hierbei werden unterschiedliche Konstrukte betrachtet. Als relevante Kontroll- oder

Gruppierungsvariablen in Fragestellungen der empirischen Bildungsforschung gelten der Migrationshintergrund (operationalisiert über Herkunft der Vorfahren bzw. Muttersprache je nach Ziel und Fragestellung; Kristen, Olczyk & Will, 2016) und der sozioökonomische Status der Familie, der sowohl als Proxy zur Generierung weiterer Statusvariablen als auch als wichtiges Maß zur Untersuchung sozialer Disparitäten dient (Stocké, Blossfeld, Hoenig & Sixt, 2011). Die Operationalisierung bzw. Berechnung solcher Hilfsvariablen für Analysen kann Effekte auf die Gruppenklassifizierung haben und somit die Ergebnisse beeinflussen (Ehmke & Siegle, 2005; Kreuter, Maaz & Watermann, 2006; Kristen et al., 2016). Es kann zudem vermutet werden, dass eine eingeschränkte Validität der Angaben eine Verzerrung der Ergebnisse nach sich zieht. Faktische Abfragen erfordern ein möglichst genaues Abrufen der relevanten Informationen und eine entsprechend präzise Übertragung auf die vorgegebenen Antwortkategorien, um valide Antworten zu erhalten (vgl. Tourangeau et al., 2000). Neben den erwähnten Variablen zum Migrationshintergrund und sozioökonomischen Familienstatus wurden weitere relevante Variablen, die eine Selbsteinschätzung erfordern, untersucht: die international anschlussfähige Skala zur Erfassung des generellen und domänenspezifischen akademischen Selbstkonzeptes (Wohlkinger, Ditton, von Maurice, Haugwitz & Blossfeld, 2011) sowie die national etablierte Skala zur Erfassung der Lesemotivation (Möller & Bonerad, 2007). Diese Skalen sind wichtige Maße für weitere inhaltliche Analysen und stellen in Anlehnung an das kognitive Antwortmodell andere Anforderungen an die teilnehmenden SchülerInnen (siehe Abschnitt 2.2.2). Die untersuchten Befragungsinstrumente wurden hinsichtlich bestimmter Aspekte angepasst, um die Erfassung valider Daten zu begünstigen. So wurde die Länge des Fragebogens reduziert um eine reduzierte Aufmerksamkeitspanne der SchülerInnengruppe mit SPF-L zu berücksichtigen. In diesem Zuge wurden auch die Inhalte

angepasst. Ausgewählte Items sollten grammatikalisch nicht zu komplex formuliert und die Inhalte salient für die SchülerInnen an Förderschulen Lernen sein. Um Ermüdungseffekte im Verlauf des Fragebogens untersuchen zu können, wurde die Reihenfolge der Items variiert (vorwärts und rückwärts). Trotz der genannten Anpassungen können spezifische Itemcharakteristika die Validität beeinträchtigen. Weiter ist zu vermuten, dass bestimmte Konzepte bezogen auf den sozioökonomischen Status und die Herkunft der Familie noch nicht ausreichend salient für Jugendliche sind und somit teilweise keine validen Angaben zu erwarten sind. Auch negativ formulierte Items könnten zu einer inkonsistenten Fragenbeantwortung führen und somit die Validität der Daten einschränken. Dennoch wurde auf Grund der akkommodierten Darbietung und zielgruppenspezifischen Gestaltung der Fragebögen erwartet, dass valide Daten von SchülerInnen mit SPF-L erfragt werden können.

Eine valide Erfassung bildungsrelevanter Konstrukte bedeutet nicht zwangsläufig, dass die erfassten Daten auch vergleichbar sind mit anderen SchülerInnengruppen und somit vergleichende Analysen über verschiedene Gruppen hinweg möglich sind. Daher schließt die dritte Frage an die Voraussetzung für vergleichende Analysen an: die Messinvarianz der Daten. Möchte man die Gruppe von SchülerInnen mit SPF-L näher beschreiben, ob bezogen auf sozial-emotionale oder kognitive Variablen, ist eine Referenzierung zu einer anderen SchülerInnengruppe zentral. Eine Darstellung der Zielgruppe ohne Bezug zu anderen SchülerInnen (z. B. ohne Beeinträchtigungen) kann wenig zu einer differenziellen Deskription der Besonderheit der SchülerInnen mit SPF-L beitragen. Vor allem im Rahmen (sonder)pädagogischer Diskussionen und auch der aktuell gesellschaftlich relevanten Debatte zum Thema Inklusion sind Vergleiche zwischen SchülerInnen in unterschiedlichen Schulsettings nicht trivial (vgl. Schwab & Helm, 2015). Für vergleichende Analysen muss

nicht nur die Validität der Daten gesichert, sondern auch die Voraussetzung der Messäquivalenz erfüllt sein (Steenkamp & Baumgartner, 1998; Steinmetz et al., 2009). Jedoch kann es auf Grund gruppenspezifischer Charakteristika zu einer eingeschränkten Vergleichbarkeit kommen (Heppt, Haag, Böhme & Stanat, 2015; Schwab & Helm, 2015; Südkamp, Pohl & Weinert, 2015). Zudem können Einschränkungen der Messinvarianz durch angepasste Administrationsbedingungen entstehen.

Die drei beschriebenen Fragestellungen der vorliegenden Arbeit werden auf unterschiedliche Art und Weise in den vier Beiträgen aufgegriffen. Hierbei werden verschiedene Ansätze genutzt, um sich den Aspekten des Zugangs zu kognitiven Anforderungen in Befragungen und Testungen, der Validität und Vergleichbarkeit der Daten anzunähern.

Der erste Artikel (*Instructions in test-taking: An appropriate approach for students with special educational needs*) ist im Grundschulalter lokalisiert und greift primär die erste Fragestellung auf. Konkret wird untersucht, inwieweit die Testinstruktion für ein standardisiertes Testverfahren zur Erfassung der Lesekompetenz (Subtest des Leseverständnistests ELFE 1-6; Lenhard & Schneider, 2006) zielgruppenspezifisch angepasst werden kann und ob diese Anpassungen zu einer verbesserten Testperformanz der SchülerInnen mit SPF-L führen. Verbesserte Testperformanz wird hier verstanden als eine Reduktion nicht-instruktionskonformer Antworten sowie ein höherer Testscore (Anzahl korrekt gelöster Antworten). Aufgrund empirischer Befunde, dass SchülerInnen mit SPF-L im Vergleich zu SchülerInnengruppen an Hauptschulen einen erhöhten Anteil nicht-instruktionskonformer Antworten in Kompetenztests aufweisen (Südkamp, Pohl & Weinert, 2015), wurde angenommen, dass dieser Befund durch ein mangelndes

Instruktionsverständnis (Zielinski, 1996) erklärt werden kann. Verschiedene Studien konnten zeigen, dass Testinstruktionen mit expliziteren Informationen sowie ergänzenden Übungsphasen inklusive direktem Feedback zu Lösungsweg und korrekter Antwort die Testergebnisse von SchülerInnen mit kognitiven Beeinträchtigungen verbessern können (Hessels, 2009; Hessels-Schlatter, 2002; Wong, Wong & LeMare, 1982). Basierend auf theoretischen Annahmen und empirischen Befunden wurden für Beitrag 1 zwei Interventionen entwickelt, um a) das Verständnis der Testanforderungen und korrekten Aufgabenbearbeitung zu verbessern und b) die Aufmerksamkeit der SchülerInnen für die Standard-Testinstruktion zu erhöhen. Die erste Intervention bestand in einer spielerischen Intensivierung der Testinstruktion und der Wiederholung von Übungsitems begleitet von einem pädagogischen Agenten; die zweite Intervention beinhaltete ein Bewegungsspiel direkt vor der Instruktion. SchülerInnen an Förderschulen Lernen in 3. und 4. Jahrgängen wurden zufällig einer von zwei Experimentalbedingungen oder einer Kontrollgruppe zugewiesen. Zusätzlich wurden relevante Kontrollvariablen (kognitive Grundfähigkeiten und Lesegeschwindigkeit) erfasst. Es wurde angenommen, dass durch eine intensivierte Testinstruktion das Verständnis der Anforderungen und die Testbearbeitung verbessert wird und eine zielgruppenspezifische Akkommodation der Instruktion somit positiven Einfluss auf die Testperformanz hat. Die Überprüfung der Hypothesen wurde anhand negativ-binominaler Regressionsmodelle (Hilbe, 2011) vorgenommen, welche die Besonderheit nicht normaldistribuierter Daten berücksichtigt.

Inwieweit SchülerInnen mit SPF-L in Klassen 5 und 9 valide Angaben bei schriftlichen Befragungen abgeben können, untersucht der zweite Beitrag (*Validity of survey data of students with special educational needs – Results from the National Educational Panel Study*). Konkret wurden hier die Angaben zum sozioökonomischen Hintergrund sowie der

familiären Herkunft bis zur dritten Generation betrachtet. Diese Variablen sind zur Beschreibung der Stichproben sowie als Kontrollvariablen für viele verschiedene Fragestellungen innerhalb der Bildungsforschung relevant (Ehmke & Siegle, 2005; Stocké et al., 2011). In diesem Beitrag wurde untersucht, ob die Angaben der SchülerInnen mit SPF-L auch als stellvertretende Angaben der Eltern (Proxy-Report) genutzt werden können. Als Außenkriterium zur Untersuchung der Validität der SchülerInnenangaben wurden die Daten der Eltern hinzugezogen. Stimmten die beiden Angaben – der Eltern und ihres Kindes – überein, wurde dies als Hinweis gewertet, dass die SchülerInnenangabe valide ist (vgl. Ensminger et al., 2000; Maaz, Kreuter & Watermann, 2006). Die Annahme, dass auf Grund der Übereinstimmung die Angaben korrekt sind, konnte jedoch nicht überprüft werden. Die Operationalisierung der Übereinstimmung erfolgte über den zufälligkeitsbereinigten Index Cohens Kappa (Wirtz & Caspar, 2002). Weiterer Bestandteil der Studie ist die Analyse einer Vergleichsstichprobe an Regelschulen.

Die Untersuchung der faktoriellen Validität und Vergleichbarkeit von Fragebogendaten an zwei ausgewählten Beispielkonstrukten (Lesemotivation, akademisches Selbstkonzept) erfolgte im dritten Artikel (*Befragung von Schülerinnen und Schülern mit sonderpädagogischem Förderbedarf Lernen: Ergebnisse zur Messinvarianz*). Die Prüfung der Validität wurde anhand konfirmatorischer Faktorenanalysen basierend auf den theoretisch postulierten Modellen überprüft. Als wichtige Referenzstichprobe wurden in dieser Arbeit SchülerInnen an Hauptschulen gewählt, da es sich um Kinder und Jugendliche handelt, die eher schwache Schulleistungen zeigen. Zudem ist eine gewisse Überlappung der Fähigkeitsbereiche zwischen den beiden Schulformen zu erwarten (Bos et al., 2010). Zur Untersuchung der Vergleichbarkeit der Daten (konfigurale, metrische und skalare

Messinvarianz; Brown, 2006) von SchülerInnen mit SPF-L mit HauptschülerInnen wurde ein Mehrgruppenvergleich mit simultaner Schätzung über die Gruppen hinweg berechnet.

Der vierte Beitrag (*There's plenty more fish in the sea. Das akademische Selbstkonzept von Schülerinnen und Schülern mit sonderpädagogischem Förderbedarf Lernen in integrativen und segregierten Schulsettings*) konzentriert sich in einem anwendungsorientierten Ansatz darauf, die in Artikel 3 untersuchten Daten in einer inhaltlichen Fragestellung zu nutzen. Die Studie untersucht den Big-Fish-Little-Pond-Effekt (BFLPE; Marsh, 2005) bei SchülerInnen mit SPF-L an Förderschulen und Regelschulen sowie deren KlassenkameradInnen ohne Beeinträchtigung. Der BFLP-Effekt entfaltet sich über soziale Vergleichsmechanismen (Dijkstra, Kuyper, van der Werf, Buunk & van der Zee, 2008), von denen angenommen wurde, dass sich diese Bezüge auch in segregierten Schulsettings nachweisen lassen. Für die SchülerInnen mit SPF-L, die in einer integrativen Klasse unterrichtet werden, die in der Regel auch ein geringeres Selbstkonzept aufweisen (Sauer et al., 2007), wurde vermutet, dass der BFLP-Effekt stärker ausfällt, da ihre Referenzgruppe innerhalb ihrer Schulklasse bezogen auf die Schulleistungen vergleichsweise stark ausfällt. Die Analysen wurden für das Fach Mathematik betrachtet, da dieses besonders stark im schulischen Kontext verankert ist (vgl. Schurtz, Pfost, Nagengast & Artelt, 2014). Die Modellierung der Mehrebenenanalyse in den drei unterschiedlichen Gruppen erfolgte getrennt voneinander, bevor die Ergebnisse miteinander verglichen wurden.

4 Darstellung und Diskussion der zentralen Befunde

In der vorliegenden Arbeit wurde untersucht, ob und wie die Performanz und Validität der Daten von SchülerInnen mit sonderpädagogischem Förderbedarf Lernen (SPF-L) durch die spezifischen kognitiven Anforderungen im Rahmen von Befragungen oder Testungen beeinträchtigt werden und welche Maßnahmen implementiert werden können, um auch bei dieser SchülerInnengruppe innerhalb von Large-Scale-Studien valide Daten zu erfassen, die vergleichbar mit anderen SchülerInnengruppen sind. Im Folgenden sollen die zentralen Ergebnisse zu den drei Fragestellungen, welchen in vier Studien nachgegangen wurde, zusammenfassend dargestellt und im Rahmen einer Einordnung in die theoretischen Grundlagen diskutiert werden. Abschließend werden die Implikationen für weitere Forschung aufgezeigt.

4.1 Befunde zum Zugang

Unter der Annahme, dass SchülerInnen mit SPF-L auf Grund ihrer Ausgangslage nicht immer über die notwendigen Fähigkeiten verfügen, um einen Test oder ein Befragungsinstrument in der von den EntwicklerInnen intendierten Art und Weise zu bearbeiten und somit valide Daten zu produzieren, fokussiert die erste Fragstellung den Zugang zum Instrument. Zugangsfertigkeiten werden hier als ein Bündel von Fähigkeiten verstanden, welche die Interaktion zwischen Studienteilnehmenden und den Anforderungen der jeweiligen Instrumente beeinflusst und zu konstrukt-irrelevanter Varianz führen kann (Beddow et al., 2013). In diesem Sinne richtet sich die erste Fragestellung zunächst auf zwei spezifische Aspekte im Rahmen von Datenerhebung bei SchülerInnen mit SPF-L: (1) das Verständnis und die Anwendung von Testinstruktionen und (2) den Administrationsmodus schriftlicher Fragebögen.

(1) Das *Verständnis von Testinstruktionen* ist ein zentraler Bestandteil einer jeder Testung, die nach standardisierten Vorgaben durchgeführt wird. Für eine faire und valide Testung ist es essentiell, dass die Vorgaben und Hinweise zur vorgesehenen Bearbeitung der nachfolgenden Testaufgaben korrekt verstanden und angewendet werden. Oerter (1980) beschreibt die Instruktionen als einen Filter, der den Teilnehmenden den Zugang zu den Anforderungen des Tests ermöglichen soll. Für SchülerInnen mit SPF-L kann es eine gruppenspezifische Herausforderung darstellen, diese inhaltlich zu verstehen und adäquat anzuwenden (vgl. Zielinski, 1996). Somit kann vermutet werden, dass das Verständnis von Testinstruktionen und den Hinweisen zur korrekten Bearbeitung von Testaufgaben Einfluss auf die Performanz und Validität der Daten hat (vgl. Wiedl, Waldorf & Schöttke, 2007).

Eine systematische Überprüfung angepasster Testinstruktionen, die die vermutete Verständnisproblematik von SchülerInnen mit SPF-L adressierte, erfolgte in Beitrag 1. Die implementierten experimentellen Bedingungen dieser Studie beinhalteten in der ersten Intervention eine intensivierte Testinstruktion, welche das Verständnis der spezifischen Testanforderungen und einer adäquaten Testbearbeitung durch Wiederholung sicherstellen sollte. Im Sinne der Direkten Instruktion nach Engelmann (1980) wurden die SchülerInnen mit Hilfe von ergänzenden Beispielitems an die Testsituation, das Itemformat und die richtige Bearbeitungsweise der Aufgaben herangeführt. Im Rahmen einer zweiten Intervention wurde – wie in der Kontrollgruppe – die Standardinstruktion dargeboten. Zusätzlich begannen die Testsitzungen hier jedoch mit einem kurzen Bewegungsspiel, welches aus einem Konzentrationstraining (Krowatschek, Krowatschek & Reid, 2011) adaptiert wurde und die allgemeine Aufmerksamkeit der Kinder steigern sollte. Die Hypothese war, dass SchülerInnen mit SPF-L sowohl von einer intensiveren und längeren Instruktionsvariante als auch von dem spielerischen Aufmerksamkeitstraining profitieren

können. Ziel war es, den Einfluss der als konstrukt-irrelevant erachteten Zugangsfertigkeiten auf die Testbearbeitung zu reduzieren.

Im Vergleich zu den Ergebnissen der Kontrollgruppe konnte bestätigt werden, dass die SchülerInnen mit SPF-L in beiden Experimentalbedingungen signifikant mehr Aufgaben korrekt lösen konnten. Zudem waren SchülerInnen, die an der intensivierten Testinstruktion teilnahmen im Vergleich zur Kontrollgruppe und Intervention 2 eher in der Lage, sich an die korrekte Bearbeitungsweise zu halten und produzierten durchschnittlich weniger Antworten, die nicht instruktionskonform waren (z. B. Mehrfachantworten). Somit konnte das Ziel der vorgenommenen Interventionen, die Herausforderung verbale Hinweise aufzunehmen, zu speichern und auf nachfolgende Testaufgaben zu transferieren, erreicht werden. Für die SchülerInnengruppe mit SPF-L mit anzunehmenden geringeren Fähigkeiten bezüglich des verbalen Instruktionsverständnisses konnte diese zusätzliche Dimension der Zugangsfertigkeiten aus der zu erfassenden Kompetenz (Textverständnis) in Ansätzen auspartialisiert werden.

Einschränkend muss gesagt werden, dass die Daten keine Aussagen über die zugrundliegenden Mechanismen und Wirkweisen der Interventionen erlauben. So kann nicht erklärt werden, welche Faktoren der beiden Interventionen zu einer höheren Testleistung geführt haben. Ob das Bekanntwerden mit der Testsituation im Rahmen von Intervention 1 (vgl. Hessels, 2009; Pohl et al., 2016) oder die Promotion der Aufmerksamkeit in Intervention 2 (vgl. Budde, Voelcker-Rehage, Pietraßyk-Kendziorra, Ribeiro & Tidow, 2008; Krowatschek et al., 2011) als entscheidende Faktoren angenommen werden können, kann nicht abschließend geklärt werden. Jedoch waren beide

Maßnahmen, vor allem aber Intervention 1, zielführende Werkzeuge, um die *accessbility* bzw. den Zugang zum Instrument zu befördern (Beddow et al., 2011; Kettler et al., 2011).

(2) *Administration schriftlicher Fragebögen:* Vor dem Hintergrund der Theorie der Zugangsfertigkeiten (Beddow et al., 2011) kann argumentiert werden, dass beispielsweise Leseverständnis und Lesegeschwindigkeit nicht Teil der erfassten Konstrukte und Inhalte in schriftlichen Fragebögen sein sollten. Auch SchülerInnen mit einer geringen Leseleistung sollen Fragen ohne für den Beantwortungsprozess hinderliche Anforderungen bearbeiten können, wenn das Ziel valide Daten über z. B. familiäre Hintergründe und Selbsteinschätzungen der Personen sind. Daher wurde für die Darbietung des Fragebogeninstruments eine Read-Aloud-Akkommodation gewählt. Mit Hilfe eines standardisierten Skripts wurden die SchülerInnen an Förderschulen Lernen durch den Fragebogen geleitet. Alle Fragen, Antwortoptionen und Items wurden den teilnehmenden Kindern und Jugendlichen zusätzlich zum Mitlesen laut vorgelesen, so dass sie weniger kognitive Kapazitäten für das Selberlesen der einzelnen Fragen aufbringen mussten.

Zu der Anpassung des Vorlesens wurde kein experimentelles Design implementiert, so dass eine systematische Überprüfung des Darbietungsmodus mit den vorliegenden Daten der Machbarkeitsstudie an Förderschulen Lernen nicht erfolgen kann. Wichtige Hinweise zur Eignung dieser Maßnahmen können jedoch aus weiteren Studien im Rahmen des NEPS abgeleitet werden, die im Folgenden kurz dargestellt werden.

Im NEPS wurde der Modus des Vorlesens für Fragebogeninstrumente bereits an FünftklässlerInnen an Hauptschulen positiv evaluiert. Die SchülerInnen, die den Fragebogen vorgelesen bekamen, produzierten weniger fehlende Werte im Vergleich zu jener SchülerInnengruppe, die den Fragebogen eigenständig bearbeitete. Besonders

langsam lesende Jugendliche konnten vom Vorlesen profitieren und präferierten diese Form der Fragebogenadministration (Gresch, Strietholt, Kanders & Solga, 2016). Diese Ergebnisse sind jedoch nicht uneingeschränkt übertragbar auf SchülerInnen an Förderschulen Lernen. So konnte im direkten Vergleich zwischen SchülerInnen an Förderschulen Lernen und Hauptschulen gezeigt werden, dass sich die Ergebnisse der Jugendlichen in einem Test zur Erfassung der deklarativen Metakognition in einer Vorlese-Version im Vergleich zu einer selbstständig bearbeiteten Variante nur bei den SchülerInnen an Hauptschulen verbesserten (Händel, Lockl, Heydrich, Weinert & Artelt, 2014). Dieser Befund kann jedoch nicht ganz von den bereits in der Forschung immer wieder bestätigten Ergebnissen entkoppelt werden, dass SchülerInnen mit SPF-L über weniger metakognitive Strategien verfügen (Pintrich, Anderman & Klobucar, 1994). Andererseits fanden amerikanische Studien Hinweise, dass SchülerInnen mit einer learning disability und besonderen Schwierigkeiten im Lesen bei bestimmten Items, die als schwieriger lesbar eingeschätzt wurden, von einer Read-Aloud-Akkommodation profitieren konnten (Bolt & Thurlow, 2006).

Weitere Befunde, die die Wirksamkeit der Read-Aloud-Akkommodation für SchülerInnen mit SPF-L bei Befragungen unterstützen, können den Abschnitten 4.2 und 4.3 entnommen werden, die die Validität und Vergleichbarkeit der Fragebogendaten vorstellt.

4.2 Befunde zur Validität

Im Rahmen der zweiten Fragestellung wurde untersucht, ob die erfassten Daten von SchülerInnen mit SPF-L valide sind und somit die Konstrukte messen, die intendiert waren, erfasst zu werden. Den Beiträgen 1, 2 und 3 können sowohl Hinweise als auch konkrete Ergebnisse zur Validität der Befragungs- und Testdaten entnommen werden.

In Beitrag 1 wird vor allem der Zugang zu den Aufgaben und den Anforderungen einer Testsituation evaluiert, darüber hinaus können Hinweise zur Validität des Tests extrahiert werden. So konnten in beiden experimentellen Bedingungen der Testinstruktion die SchülerInnen mit SPF-L eine bessere Testperformanz zeigen als ihre KlassenkameradInnen in der Kontrollgruppe. Damit kann angenommen werden, dass die SchülerInnen mit SPF-L über eine höhere Lesekompetenz verfügen als die Ergebnisse unter unveränderten Bedingungen vermuten lassen würden. Dieser Argumentation folgend kann die Hypothese formuliert werden, dass die angepassten Testinstruktionen eher zu validen Daten und Ergebnissen hinsichtlich der Kompetenzen der SchülerInnen führten (vgl. Carlson & Wiedl, 1992; Wiedl et al., 2007). Wenn Testinstruktionen nicht verstanden werden, kann die Testung nicht fair und valide sein (vgl. Wong et al., 1982).

Die Beiträge 2 und 3 konzentrieren sich auf Befragungsdaten, und auf die Frage, inwieweit die zielgruppenspezifischen Anpassungen der Anleitungen zum Fragebogen und der Darbietung des Fragebogens (Read-Aloud) zu validen Daten führten.

Beitrag 2 untersuchte die Angaben der SchülerInnen zum sozioökonomischen Status und der Herkunft ihrer Familien. Für einige dieser Items konnten gute Validitätswerte im Sinne einer Übereinstimmung mit der Angabe der Eltern gefunden werden. Doch konnten innerhalb der SchülerInnengruppe mit SPF-L auch itemspezifische Herausforderungen identifiziert werden. Die Betrachtung faktischer Abfragen zur Herkunft der eigenen Familie zeigten, dass die Lebensnähe der Iteminhalte eine entscheidende Rolle zu spielen scheint (vgl. Looker, 1989). Die Wahrscheinlichkeit, dass ein Item überhaupt beantwortet wird, sinkt, je weiter weg die Inhalte vom eigenen Lebensalltag sind. Die SchülerInnen, die diese Lebensalltag-fernen Fragen dennoch beantworteten, taten dies wenig akkurat, so dass die

Validität der Daten eingeschränkt ist. Valide Antworten werden andererseits häufiger gegeben, wenn die SchülerInnen mit SPF-L nach ihrer eigenen Sprache und der Sprache ihrer Eltern gefragt wurden. Hierbei handelte es sich um Fakten, die im alltäglichen Austausch besonders relevant und somit für die Jugendlichen salient sind. Auch die Abfrage der Geburtsländer der Eltern sowie des eigenen Geburtslandes führten in der Regel zu validen Antworten, wobei die jüngeren SchülerInnen in den 5. Klassen scheinbar etwas mehr Mühe hatten, korrekt zu antworten als die Jugendlichen in den 9. Klassen. Die zusätzliche Erfassung der Geburtsländer der Großeltern führte zum einen zu einem drastischen Anstieg fehlender Werte sowie zu einem deutlichen Abfall valider Antworten. Somit können die Daten der SchülerInnen mit SPF-L kaum als Proxy-Report zur Bestimmung der 3. Generation von Migranten, so wie von Olczyk, Will und Kristen (2016) für das NEPS vorgeschlagen, genutzt werden. Aber auch SchülerInnen an Regelschulen hatten deutliche Schwierigkeiten bei der umfassenden und validen Beantwortung der Fragen zur Herkunft ihrer Familie, wobei die Übereinstimmungswerte bei fast allen Items höher ausfielen als für SchülerInnen an Förderschulen. Die SchülerInnen aller 9. Klassen wurden zusätzlich zum sozioökonomischen Status (Schulabschluss, Ausbildung, Beruf der Eltern) befragt. Die Anteile der fehlenden Werte reichten hier von 10 bis 70 %, so dass anzunehmen ist, dass viele der SchülerInnen nicht über die entsprechenden Informationen verfügten, um die Fragen zu beantworten. Zusätzlich muss auf Grund der geringen Übereinstimmungen zwischen den SchülerInnen- und Elternangaben darüber hinaus angenommen werden, dass die wenigen abgegebenen Antworten nicht zwingend valide sind und somit die Aussagekraft der Daten eingeschränkt ist. Auch wenn dieses Befundmuster für die SchülerInnen an Regelschulen weniger gravierend ausfiel, so ist doch

auch für diese Gruppe anzunehmen, dass die Daten zum sozioökonomischen Status der Familie wenig verlässlich sind (vgl. auch Maaz et al., 2006).

Die Annahme, dass das gewählte Außenkriterium (Elternangabe) valide sei, ist ein pragmatisches Vorgehen, jedoch lediglich eine Hypothese, die hier nicht überprüfbar ist. Zudem entstand auf Grund dieser Analysestrategie ein systematischer Stichprobenausfall, der durch die Teilnahmebereitschaft der Eltern bedingt war. Es ist zu vermuten, dass die Teilnahme an einer Bildungsstudie keinen zufälligen Ausfall produziert, sondern dass bestimmte Personenkreise (z. B. bildungsnahe Familien) eher an der Erhebung teilnehmen als andere. Somit standen für jene SchülerInnen, zu denen keine zusätzlichen Informationen der Eltern zur Validierung ihrer Angaben vorlagen, keine Ergebnisse zur Datenqualität zur Verfügung. Es ist jedoch zu vermuten, dass die beiden Gruppen (SchülerInnen, deren Eltern teilnehmen und deren Eltern nicht teilnehmen) sich signifikant voneinander unterscheiden. Die vorgestellten Ergebnisse können daher nur für einen Teil der Stichproben gelten. Insgesamt sollten diese Angaben der SchülerInnen nicht unbedacht für inhaltliche Analysen genutzt werden, da die Validität der Daten für bestimmte Bereiche stark eingeschränkt bzw. gar nicht vorhanden ist.

Neben diesen faktischen Abfragen wurden im Fragebogen zusätzlich viele Konstrukte erfasst, welche eine subjektive Einschätzung der SchülerInnen ihrer Interessen, Motivation und ihres Selbstkonzeptes erforderten. Hierbei wurden andere Anforderungen an den Beantwortungsprozess gestellt. Es wurden eigene Tätigkeiten oder Interessen, bezogen auf einen bestimmten Sachverhalt (z. B. Lesen), erfragt. Hierzu bedarf es z. B. der Einordnung der eigenen Aktivitäten innerhalb eines bestimmten, vorgegebenen Zeithorizonts (vgl. Kränzl-Nagl & Wilk, 2000; Tourangeau et al., 2000) oder der Einschätzung der eigenen

Fähigkeiten (z. B. Selbstkonzept; Dijkstra et al., 2008). Für SchülerInnen in den 5. Klassen wurden beispielhaft die Konstrukte Lesemotivation (Möller & Bonerad, 2007) und akademisches Selbstkonzept (Wohlkinger et al., 2011) untersucht (Beitrag 3 und Beitrag 4). Mit geringfügigen Anpassungen der theoretisch postulierten Modelle, die nicht zu einer Schmälerung des inhaltlichen Konstruktes führten, konnte faktorielle Validität dieser beiden Maße bestätigt werden. Die Validität der Daten war unabhängig von der Position der Itembatterien innerhalb des Befragungsinstrumentes gegeben (Beginn vs. Ende des Instruments). Somit kann angenommen werden, dass der Fragebogen eine adäquate Länge für die Gruppe von SchülerInnen mit SPF-L aufwies, ohne dass eine Beeinträchtigung der Güte der Daten zu erwarten war. Dieses Ergebnis ist vermutlich unter anderem darauf zurückzuführen, dass die Anpassung Read-Aloud für die Fragebogenadministration gewählt wurde (siehe Abschnitt 4.1). Dennoch ist festzustellen, dass negativ formulierte Items weniger gut beantwortet werden konnten. So musste das negativ formulierte Item für die Erfassung der Subskala verbales Selbstkonzept sowohl bei der Förderschulstichprobe als auch bei SchülerInnen an Hauptschulen ausgeschlossen werden (vgl. Marsh, 1986). Der Wechsel zwischen positiv und negativ formulierten Items (und vice versa) scheint ebenso herausfordernd zu sein. Ein positiv formuliertes Item der Subskala Leseselbstkonzept, welches zwischen zwei negativ formulierten Items administriert wurde, musste auf Grund einer geringen Passung aus den Analysen ausgeschlossen. Die Schwierigkeiten können sich entweder auf der Ebene des Verständnisses oder eventuell auch im Rahmen der Response-Findung, in der die eigene Antwort auf eine entsprechende, zur Verfügung stehende Antwortoption übertragen werden muss (vgl. Tourangeau et al., 2000), verorten lassen. Die jeweilige neue Polung und erneute Einordnung bei konträr formulierten Aussagen könnte für SchülerInnen mit SPF-L, aber scheinbar auch für

SchülerInnen an Hauptschulen, eine kognitive Anforderung darstellen, die die Konsistenz und Validität der Antworten beeinträchtigen (vgl. Borgers et al., 2000).

Auch der Wechsel zwischen inhaltlichen Themen in einer Itembatterie mit stark ähnlich formulierten Items kann Schwierigkeiten verursachen (vgl. Borgers et al., 2000). So wies das Item „Im Fach Mathematik bekomme ich gute Noten", welches direkt auf das Item „Im Fach Deutsch bekomme ich gute Noten" folgte, eine Kreuzladung mit der verbalen Komponente des akademischen Selbstkonzepts auf. Auch in einem Vorlese-Modus ist es unabdinglich, konzentriert und genau zuzuhören, um die minimalen Unterschiede zwischen diesen beiden Formulierungen (Deutsch bzw. Mathematik) zu erkennen und bei der Beantwortung der Frage zu berücksichtigen.

Mit den vorliegenden Daten ist es nicht möglich, spezifische Probleme bezogen auf die vier kognitiven Schritte im Antwortprozess zu identifizieren (vgl. Tourangeau et al., 2000). Die kognitiven Abläufe während der Bearbeitung und Beantwortung der Fragen können nicht entflechtet werden, auch potenzielles Satisficing (Krosnick, 2000) kann nicht identifiziert werden. Es können lediglich das Ergebnis (response vs. non-response) als auch der Inhalt der Antwort analysiert werden. Der Weg zur Antwort bleibt verschlossen.

Qualitative Auswertungen der Erhebungsprotokolle lassen jedoch vermuten, dass bereits das Verständnis der Fragen, trotz des Vorlesens der Items, zu Schwierigkeiten bei SchülerInnen mit SPF-L geführt hat. Besonders die Fragen zur beruflichen Anstellung der Eltern führte zu vermehrten Nachfragen. Unverständnis bedeutet häufig auch, dass entsprechende Konzepte fehlen, so dass im Zuge des Retrievals nicht die relevanten Informationen abgerufen werden können. Somit wird Satisficing wahrscheinlicher (Krosnick, 1991). In der Konsequenz führte möglicherweise ein unzureichendes Verständnis

bzw. zu wenige Informationen zum Inhalt einer Frage zu fehlenden Antworten bzw. non-response (vgl. Borgers et al., 2000), wie sie für bestimmte Items (höchster Bildungsabschluss und Beruf der Eltern, Geburtsländer der Großeltern) zu finden waren.

Insgesamt lässt sich auf Basis der Artikel 1, 2, und 3 anführen, dass es mit den beschriebenen Maßnahmen und Akkomodationen möglich ist, valide Daten von SchülerInnen mit SPF-L zu erfassen. Dennoch ist besonderer Fokus auf die Salienz der Fragen sowie der konkreten Formulierung und Abfolge der einzelnen Items zu legen.

4.3 Befunde zur Vergleichbarkeit

Die Beschreibung der Gruppe von SchülerInnen mit SPF-L, ihrer Kompetenzentwicklung und Determinanten ihrer Bildungsverläufe ist von besonderem Interesse, wenn sie im Vergleich mit anderen Jugendlichen betrachtet werden kann. Hierfür muss zusätzlich zur Bedingung der Reliabilität und Validität der Daten die Voraussetzung der Messinvarianz erfüllt sein. Spezifisch wurde in Beitrag 3 neben der Validität der Daten auch die Vergleichbarkeit mit der Gruppe von SchülerInnen an Hauptschulen untersucht. Die faktorielle Validität zeigte für beide Gruppen vergleichbare Faktorstrukturen für die Konstrukte Lesemotivation (Möller & Bonerad, 2007) und akademisches Selbstkonzept (Wohlkinger et al., 2011), so dass von konfiguraler Messinvarianz (Byrne & van de Vijver, 2010; Dimitrov, 2010) ausgegangen werden kann. Die metrische Messinvarianz nimmt vergleichbare Faktorladungen an, die skalare Messinvarianz geht von vergleichbaren Intercepts aus (Byrne & van de Vijver, 2010; Dimitrov, 2010). Erst wenn diese, zumindest partiell, angenommen werden können, sind vergleichende Analysen sinnvoll interpretierbar (Steenkamp & Baumgartner, 1998). Bei der simultanen Schätzung der Modelle über die beiden Gruppen hinweg (SchülerInnen an Förderschulen Lernen und an Hauptschulen) zeigte sich, dass es bestimmte Items gibt, die konzeptuell unterschiedlich

verstanden wurden. Dies wurde an nicht äquivalenten Faktorladungen evident (vgl. Chen, 2008). Da für das Konstrukt Lesemotivation jede der drei Subskalen (Leselust, Lesen aus Interesse und Leseselbstkonzept) mit jeweils noch zwei invarianten Items vertreten sind, können die Daten der SchülerInnen an Förderschulen und an Hauptschulen für vergleichende Analysen genutzt werden (vgl. Cheung & Rensvold, 2002). Jedoch sollten die Items mit invarianten Faktorladungen je nach Forschungsfrage aus den Analysen ausgeschlossen werden. Das Konstrukt akademisches Selbstkonzept mit drei Subskalen (verbales, mathematisches und globales Selbstkonzept) ging auf Grund einer geringen Faktorladung (negativ formuliertes Item) und einer Kreuzladung (siehe Abschnitt 4.2) mit insgesamt nur sieben anstatt neun Items in die Prüfung der Messinvarianz ein. Die Analysen zeigten zwei Items mit invarianter Faktorladung sowie ein Item mit invariantem Intercept. Somit sind die beiden Subskalen verbales und globales Selbstkonzept mit lediglich je einem messäquivalenten Item vertreten. Nur die Subskala mathematisches Selbstkonzept weist zwei messinvariante Items auf und kann für vergleichende Analysen genutzt werden. Die Gesamtskala zum akademischen Selbstkonzept sollte gemäß dieser Ergebnisse in der administrierten Form nicht für Vergleiche zwischen den SchülerInnengruppen an Förderschulen und Hauptschulen genutzt werden. Die fehlende Vergleichbarkeit kann zum einen durch die starke Ausrichtung der Itemformulierung auf die spezifischen Leistungen der SchülerInnen (z. B. gute Noten erhalten) begründet sein. Die Handhabung von Noten kann in den beiden Schulformen deutlich unterschiedlich ausfallen, so dass die Outputorientierung zu unterschiedlichem Antwortverhalten in den beiden Schultypen führen könnte. Zum anderen ist das Konstrukt des Selbstkonzeptes nicht unabhängig von Referenzgruppeneffekten. Der erwartete BFLP-Effekt für die hier als messäquivalent befundene Subskala mathematisches Selbstkonzept wird in Beitrag 4

genauer untersucht. Für alle untersuchten Stichproben der SchülerInnen mit SPF-L an Förderschulen, SchülerInnen mit SPF-L an Regelschulen und ihre KlassenkameradInnen ohne SPF wird ein BFLP-Effekt als Resultat sozialer Vergleichsmechanismen erwartet (vgl. Dijkstra et al., 2008; Marsh, 2005). Für integrativ beschulte SchülerInnen mit SPF-L wurde jedoch angenommen, dass der BFLP-Effekt höher ausfällt, da ihre Referenzgruppe (SchülerInnen ohne Beeinträchtigung) als vergleichsweise stärker einzuordnen ist. Diese zweite Hypothese wurde in den mehrebenenanalytischen Modellen des Artikels 4 bestätigt: Für SchülerInnen mit SPF-L an Regelschulen fiel der Einfluss des mittleren Leistungsniveaus auf ihr eigenes mathematisches Selbstkonzept stärker aus als für ihre KlassenkameradInnen ohne SPF. Entgegen der Erwartungen war der BFLP-Effekt für SchülerInnen an Förderschulen nicht nachweisbar. Vielmehr scheint die Ausprägung des mathematischen Selbstkonzepts überwiegend durch die eigenen Noten bestimmt und nicht durch die Mechanismen des sozialen Vergleiches mit den Leistungen der MitschülerInnen im Klassenverband. Wenn diese Referenzierungsprozesse an Förderschulen und an Regelschulen unterschiedlich ausfallen bzw. nicht sowohl in segregierten Förderschulen als auch in Regelschulen (auch integrativen Schulsettings) gleichermaßen evident sind, kann die Hypothese formuliert werden, dass das akademische Selbstkonzept kaum äquivalent messbar ist, da die zugrundeliegende theoretische und vielfach empirisch belegte Annahme der sozialen Vergleiche möglicherweise wenig Einfluss auf die Ausbildung des akademischen Selbstkonzeptes bei SchülerInnen an Förderschulen hat. Die vielberichteten Unterschiede in der Ausprägung des Selbstkonzeptes von SchülerInnen mit SPF-L an Förderschulen und an Regelschulen (Möller, Streblow & Pohlmann, 2002; Sauer et al., 2007; Schwab, 2014) sind auf Grund dieser Befunde möglicherweise in einem anderen Licht zu interpretieren. Wenn das

Verständnis der dargebotenen Items und Aussagesätze auf unterschiedliches Antwortverhalten zurückgeführt werden kann, da die vorhandenen Konzepte nicht äquivalent sind, sind vergleichende Aussagen – auch über mittlere Ausprägungen – nicht sinnvoll.

Die in Beitrag 2 untersuchten Variablen zum sozioökonomischen Status und Migrationshintergrund der Kinder und Jugendlichen sind auf Grund der eingeschränkten Validität nur bedingt zu vergleichenden Beschreibungen der verschiedenen Stichproben nutzbar (siehe Abschnitt 4.2).

Für die in Artikel 1 implementierten Interventionen zum Verständnis der Testinstruktionen stehen Analysen zur Vergleichbarkeit noch aus. Um weitere Aussagen über die Wirksamkeit der Interventionen treffen zu können, müssten für weitere differenzielle Analysen auch Erhebungen an Grundschulen bei integrativ beschulten Kindern mit SPF-L sowie SchülerInnen ohne SPF durchgeführt werden. Erst dann kann auch im Sinne der *differential boost hypothesis* (Fuchs & Fuchs, 2001) untersucht werden, ob vor allem SchülerInnen mit SPF-L von angebotenen Interventionen profitieren, der Testscore von Kinder ohne Beeinträchtigung jedoch nicht beeinflusst wird.

4.4 Implikationen für Forschung und Praxis

Für eine faire und valide Testung bzw. Befragung von SchülerInnen mit SPF-L müssen die gruppenspezifischen Ausgangsvoraussetzungen berücksichtigt werden. Vor allem im Rahmen großangelegter Large-Scale-Studien, welche eine starke Standardisierung der Erhebungen erforderlich machen, damit differenzielle Ergebnisse nicht auf unterschiedliche Administrationsbedingungen zurückzuführen sind, sondern auf tatsächliche Unterschiede zwischen den Personen (Pitoniak & Royer, 2001), ist ein

besonderes Augenmerk auf Minderheitengruppen zu legen. SchülerInnen mit SPF-L sind auf Grund ihrer Beeinträchtigungen häufig durch fehlende Zugangsfertigkeiten wie geringeres Instruktionsverständnis oder geringe Lesekompetenz im Rahmen von Befragungen und Testungen benachteiligt, so dass es zur Beeinträchtigung der Validität und somit der Vergleichbarkeit der Daten kommt (vgl. Beddow et al., 2011). Um fehlende Ausgangs- und Zugangsfertigkeiten auszugleichen, können z. B. gruppenspezifische Akkommodationen implementiert werden, welche die konstrukt-irrelevante Varianz minimieren sollen (Koretz & Barton, 2004; Sireci et al., 2005).

Die vorliegende Arbeit zeigt verschiedene Ansätze auf, wie der Zugang zu Testungen und Befragungen gestaltet werden kann, um bildungsrelevante Konstrukte valide und vergleichbar bei SchülerInnen mit SPF-L zu erfassen. Die in Abschnitt 2.2 dargelegten theoretischen Modelle müssen bereits bei der Instrumentenentwicklung Berücksichtigung finden, damit die relevanten Faktoren der Testfairness, des Zugangs oder der Itemformulierung die Güte und Validität der Daten nicht beeinträchtigen. Die Einbeziehung von SchülerInnen mit SPF-L in Schulleistungserhebungen ist nicht trivial, sollte aber weiter forciert werden. Auch wenn für einige der Variablen und theoretische Konstrukte Hinweise für eine valide und vergleichbare Erfassung gefunden wurden, bleibt die Herausforderung einer soliden Datenqualität in diesem Zusammenhang bestehen. Die Auswahl von Items und der Zusammenstellung von Instrumenten, die vergleichende Analysen erlauben, sind mit Bedacht sowie theoretisch fundierten Annahmen bezüglich der Anforderungen an die Zielgruppen umzusetzen. Es wird an verschiedenen Stellen evident, dass eine reine zielgruppenspezifische Anpassung einzelner Bausteine der Erhebung (z. B. Administrationsbedingung Read-Aloud) nicht ausreichend ist, um Konstrukte messäquivalent zu erfassen. Vielmehr muss bereits bei der Entwicklung von

Befragungsinstrumenten sowie bei der Auswahl und Zusammenstellung einzelner Items angesetzt werden (vgl. Beddow et al., 2011). Besonders ist auf die Salienz der Inhalte (vgl. Kränzl-Nagl & Wilk, 2000; Looker, 1989), die Itemformulierungen (vgl. Bell, 2007; Borgers et al., 2000; Holaday & Turner-Henson, 1989; Marsh, 1986), eine hinreichende Differenzierung der Formulierungen innerhalb einer Itembatterie (vgl. Nusser, Carstensen & Artelt, 2015) Wert zu legen.

Ein besonderer Schwerpunkt dieser Arbeit sind die sprachlichen Anforderungen im Rahmen von Testinstruktionen, die an die SchülerInnen mit SPF-L herangetragen werden. Nur wenn diese Anleitungshinweise verstanden und angewendet werden, können valide Daten erhoben werden. Die Frage ist, ob das Verständnis der Testinstruktionen inhärent in dem Konstrukt verankert ist. Die Autorin argumentiert, dass das Verständnis sprachlicher Anweisungen ein zielgruppenspezifisches Hindernis für SchülerInnen mit SPF-L darstellen kann und somit eine zusätzliche Dimension darstellt, die aus dem zu messenden Konstrukt auspartialisiert werden sollte. In Beitrag 1 konnten Hinweise gesammelt werden, dass Variationen in der Darbietung von Testinstruktionen auf die Ergebnisse eines Tests und somit auf die Translation von Kompetenz in Performanz einwirken können (vgl. Lauth & Wiedl, 1985; Wiedl et al., 2007; Wong et al., 1982). Die Forderung der American Educational Research Association nach eindeutig und einfach formulierten Instruktionen wird an dieser Stelle nochmals hervorgehoben (American Educational Research Association et al., 1999), besonders in Bezug auf Minderheitengruppen, ob Personen mit Migrationshintergrund oder Beeinträchtigungen. Für die Entwicklung von Testinstruktionen, die adäquat für SchülerInnen mit SPF-L sind, können neben der konkreten diesbezüglichen Forschung (vgl. Hessels, 2009; Lauth & Wiedl, 1985; Wong et al., 1982) auch die vielfältigen Erkenntnisse

bezüglich schulischer Interventionen für Kinder und Jugendliche mit Lernschwierigkeiten genutzt werden (Heward, 2003; Lauth et al., 2014; Lauth & Grünke, 2005).

Der Aspekt des Instruktionsverständnisses ist nicht nur für die Administration von Kompetenzests in Bildungsstudien relevant, sondern spielt auch eine zentrale Rolle im alltäglichen Unterrichtsgeschehen. Inhalte werden häufig sprachlich durch Lehrkräfte und in der Interaktion mit den MitschülerInnen vermittelt. Können diese Reize nicht aufgenommen, enkodiert und verarbeitet werden, kann neues Wissen und Kompetenzen nicht erworben werden. Auch werden die SchülerInnen nicht in der Lage sein, die ihnen gestellten Aufgaben – ob im Unterricht, bei den Hausaufgaben oder bei Lernkontrollen – zu verstehen und einen entsprechenden Zugang zu den Anforderungen zu erlangen. Ihre Performanz könnte somit hinter ihren tatsächlichen Kompetenzen zurückbleiben.

Die dargestellten inhaltlichen als auch methodischen Herausforderungen müssen auch im sonderpädagogischen Diskurs aufgenommen werden (siehe auch Schwab & Helm, 2015). Die teilweise emotionsbeladene Debatte um eine inklusive vs. segregierte Beschulung (vgl. Werning, 2010) erfordert fundierte Studien, die sich um eine valide und vergleichbare Erfassung bildungsrelevanter Konstrukte bemüht, um belastbare empirische Evidenzen für die Bildungsverläufe und Kompetenzentwicklungen der SchülerInnen mit SPF-L zu erhalten.

Literatur

Abedi, J. (2011). Language issues in the design of accessible items. In S.N. Elliott, R.J. Kettler, P.A. Beddow & A. Kurz (Hrsg.), *Handbook of Accessible Achievement Tests for All Students* (S. 217–230). New York, NY: Springer New York.

Alwin, D. F. (1997). Feeling thermometers versus 7-point scales: Which are better? *Sociological Methods & Research*, *25* (3), 318–340. doi:10.1177/0049124197025003003

American Educational Research Association, American Psychological Association & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. American Educational Research Association.

Artelt, C. & Baumert, J. (2004). Zur Vergleichbarkeit von Schülerleistungen bei Leseaufgaben unterschiedlichen sprachlichen Ursprungs. *Zeitschrift für Pädagogische Psychologie*, *18* (3/4), 171–185. doi:10.1024/1010-0652.18.34.171

Autorengruppe Bildungsbericht (2016). *Bildung in Deutschland 2016: ein indikatorengestützter Bericht mit einer Analyse zu Bildung und Migration*. Bielefeld: Bertelsmann Verlag.

Baltes, P. B. (1987). Theoretical propositions of life-span developmental psychology: On the dynamics between growth and decline. *Developmental Psychology*, *23* (5), 611–626.

Banks, T. & Eaton, I. (2014). Improving test-taking performance of secondary at-risk youth and students with disabilities. *Preventing School Failure: Alternative Education for Children and Youth*, *58* (4), 207–213. doi:10.1080/1045988X.2013.792765

Baumert, J., Artelt, C., Carstensen, C. H., Sibberns, H. & Stanat, P. (2002). Untersuchungsgegenstand, Fragestellungen und technische Grundlagen der

Studie. In Deutsches PISA-Konsortium (Hrsg.), *PISA 2000 - Die Länder der Bundesrepublik Deutschland im Vergleich* (S. 11–38). Wiesbaden: VS Verlag für Sozialwissenschaften.

Baumert, J. & Demmrich, A. (2001). Test motivation in the assessment of student skills: The effects of incentives on motivation and performance. *European Journal of Psychology of Education*, *16* (3), 441–462.

Beddow, P. A., Elliott, S. N. & Kettler, R. J. (2009). *TAMI Accessibility Rating Matrix*. Nashville, TN: Vanderbilt University. Zugriff am 7.5.2017. Verfügbar unter: https://peabody.vanderbilt.edu/docs/pdf/PRO/TAMI_Accessibility_Rating_Matrix.pdf

Beddow, P. A., Elliott, S. N. & Kettler, R. J. (2013). Test accessibility: Item reviews and lessons learned from four state assessments. *Education Research International*, *2013*, 1–12. doi:10.1155/2013/952704

Beddow, P. A., Kurz, A. & Frey, J. R. (2011). Accessibility Theory: Guiding the science and pracice of test item design with the test-taker in mind. In S.N. Elliott, R.J. Kettler, P.A. Beddow & A. Kurz (Hrsg.), *Handbook of Accessible Achievement Tests for All Students* (S. 163–182). New York, NY: Springer.

Bell, A. (2007). Designing and testing questionnaires for children. *Journal of Research in Nursing*, *12* (5), 461–469. doi:10.1177/1744987107079616

Bleidick, U. (1998). Lernbehinderung. *Einführung in die Behindertenpädagogik. 2: Blindenpädagogik, Gehörlosenpädagogik, Geistigbehindertenpädagogik, Körperbehindertenpädagogik, Lernbehindertenpädagogik* (S. 106–131). Stuttgart: Kohlhammer.

Blossfeld, H.-P., Roßbach, H.-G. & von Maurice, J. (Hrsg.). (2011). Education as a lifelong process. The German National Educational Panel Study (NEPS). *Zeitschrift für Erziehungswissenschaft*, *14* (2).

Bolt, S. E. & Thurlow, M. L. (2006). Item-level effects of the read-aloud accommodation for students with reading disabilities. Synthesis Report 65. *National Center on Educational Outcomes, University of Minnesota*.

Borgers, N., De Leeuw, E. D. & Hox, J. (2000). Children as respondants in survey research: Cognitive development and response quality. *Bulletin de Methodologie Sociologique*, *66*, 60–75.

Borgers, N. & Hox, J. (2001). Item nonresponse in questionnaire research with children. *Journal of Official Statistics*, *17* (2), 321–335.

Borgers, N., Hox, J. & Sikkel, D. (2003). Response qualitiy in survey research with children and adolescents: The effect of labeled response options and vague quantifiers. *International Journal of Public Opinion Research*, *15* (1), 83–94.

Bortner, M. & Birch, H. G. (1969). Cognitive capacity and cognitive competence. *American Journal of Mental Deficiency*, (74), 735–744.

Bortz, J. & Döring, N. (2005). *Forschungsmethoden und Evaluation: für Human- und Sozialwissenschaftler*. Heidelberg: Springer.

Bos, W., Buddeberg, I., Bremerich-Vos, A. & Schwippert, K. (Hrsg.). (2012). *IGLU 2011: Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich*. Münster New York München Berlin: Waxmann.

Bos, W., Gartmeier, M. & Gröhlich, C. (Hrsg.). (2009). *KESS 7. Kompetenzen und Einstellungen von Schülerinnen und Schülern an Hamburger Schulen zu Beginn der Jahrgangsstufe 7.* Münster: Waxmann.

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: Springer.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research* (Methodology in the social sciences). New York: Guilford Press.

Brunner, M., Artelt, C., Krauss, S. & Baumert, J. (2007). Coaching for the PISA test. *Learning and Instruction*, *17* (2), 111–122. doi:10.1016/j.learninstruc.2007.01.002

Bryan, J. H., Sonnefeld, L. J. & Grabowski, B. (1983). The relationship between fear of failure and learning disabilities. *Learning Disability Quarterly*, *6* (2), 217–222. doi:10.2307/1510800

Budde, H., Voelcker-Rehage, C., Pietraßyk-Kendziorra, S., Ribeiro, P. & Tidow, G. (2008). Acute coordinative exercise improves attentional performance in adolescents. *Neuroscience Letters*, *441* (2), 219–223. doi:10.1016/j.neulet.2008.06.024

Bundschuh, K. (2006). Rahmenbedingungen und diagnostische Umsetzung zur Feststellung sonderpädagogischen Förderbedarfs in Bayern. In U. Petermann & F. Petermann (Hrsg.), *Diagnostik sonderpädagogischen Förderbedarfs (Tests und Trends, Band 5)* (S. 17–35). Göttingen: Hogrefe.

Byrne, B. M. & van de Vijver, F. J. R. (2010). Testing for measurement and structural equivalence in large-scale cross-cultural studies: Addressing the issue of nonequivalence. *International Journal of Testing*, *10* (2), 107–132. doi:10.1080/15305051003637306

Carlson, J. S. & Wiedl, K. H. (1992). Principles of dynamic assessment: The application of a specific model. *Learning and Individual Differences*, *4* (2), 153–166. doi:10.1016/1041-6080(92)90011-3

Carlson, J. S. & Wiedl, K. H. (2000). The validity of dynamic assessment. In C.S. Lidz & J.G. Elliott (Hrsg.), *Dynamic assessment: Prevailing models and applications* (S. 681–712). New York: Elsevier Sciences.

Carter, E. W., Wehby, J., Hughes, C., Johnson, S. M., Plank, D. R., Barton-Arwood, S. M. et al. (2005). Preparing adolescents with high-incidence disabilities for high stakes-testing with strategy instruction. *Preventing School Failure*, *49* (2), 55–62.

Cassady, J. C. & Johnson, R. E. (2002). Cognitive test anxiety and academic performance. *Contemporary Educational Psychology*, *27* (2), 270–295. doi:10.1006/ceps.2001.1094

Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of Personality and Social Psychology*, *95* (5), 1005–1018. doi:10.1037/a0013193

Cheung, G. W. & Rensvold, R. B. (2002). Evaluating Goodness-of-Fit Indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9* (2), 233–255. doi:10.1207/S15328007SEM0902_5

Cormier, D. C., Altman, J., Shyyan, V. & Thurlow, M. L. (2010). A summary of the research on the effects of test accommodations: 2007-2008. Technical Report 56. *National Center on Educational Outcomes.* University of Minnesota.

Diersch, N. & Walther, E. (2010). Umfrageforschung mit Kindern und Jugendlichen. In E. Walther, F. Preckel & S. Mecklenbräuker (Hrsg.), *Befragung von Kindern und Jugendlichen: Grundlagen, Methoden und Anwendungsfelder* (S. 297–318). Göttingen: Hogrefe.

Dietze, T. (2012). Zum Stand der sonderpädagogischen Förderung in Deutschland. Die Schulstatistik 2010/11. *Zeitschrift für Heilpädagogik*, *63* (1), 26–31.

Dijkstra, P., Kuyper, H., van der Werf, G., Buunk, A. P. & van der Zee, Y. G. (2008). Social comparison in the classroom: A review. *Review of Educational Research*, *78* (4), 828–879. doi:10.3102/0034654308321210

Dimitrov, D. M. (2010). Testing for factorial invariance in the context of construct validation. *Measurement and Evaluation in Counseling and Development*, *43* (2), 121–149. doi:10.1177/0748175610373459

Duckworth, A. L., Quinn, P. D., Lynam, D. R., Loeber, R. & Stouthamer-Loeber, M. (2011). Role of test motivation in intelligence testing. *Proceedings of the National Academy of Sciences*, *108* (19), 7716–7720. doi:10.1073/pnas.1018601108

Ehmke, T. & Siegle, T. (2005). ISEI, ISCED, HOMEPOS, ESCS. *Zeitschrift für Erziehungswissenschaft*, *8* (4), 521–539.

Engelmann, S. (1980). *The instructional design library: Direct instruction*. Englewood Cliffs, N.J: Educational Technology.

Ensminger, M. E., Forrest, C. B., Riley, A. W., Kang, M., Green, B. F., Starfield, B. et al. (2000). The validity of measures of socioeconomic status of adolescents. *Journal of Adolescent Research*, *15* (3), 392–419.

Entwisle, D. R., Alexander, K. L., Cadigan, D. & Pallas, A. (1986). The schooling process in first grade: Two samples a decade apart. *American Educational Research Journal*, *23* (4), 587–613.

Esposito, J. L. & Jobe, J. B. (1991). A general model of the survey interaction process. *Bureau of the Census of Seventh Annual Research Conference Proceedings*, 537–560.

Euen, B., Vaskova, A., Walzenburg, A. & Bos, W. (2015). Armutsgefährdete Schülerinnen und Schüler mit einem Förderbedarf im Förderschwerpunkt Lernen am Beispiel von PARS-F und KESS-7-F. 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. 101–128). Wiesbaden: Springer.

Feuerstein, R., Rand, Y. & Hoffmann, M. B. (1979). *The dynamic assessment of retarded performers: The learning potential assessment device: Theory, instruments, and techniques*. Baltimore: University Park Press.

Fuchs, L. S. & Fuchs, D. (2001). Helping teachers formulate sound test accommodation decisions for students with learning disabilities. *Learning Disabilities Research & Practice*, *16* (3), 174–181.

Gebhardt, M., Oelkrug, K. & Tretter, T. (2013). Das mathematische Leistungsspektrum bei Schülerinnen und Schülern mit sonderpädagogischem Förderbedarf in der Sekundarstufe. Ein explorativer Querschnitt der fünften bis neunten Klassenstufe in Münchner Förderschulen. *Empirische Sonderpädagogik*, *5* (2), 130–143.

Gebhardt, M., Sälzer, C., Mang, J., Müller, K. & Prenzel, M. (2015). Performance of students with special educational needs in Germany: Findings from Programme for International Student Assessment 2012. *Journal of Cognitive Education and Psychology*, *14* (3), 343–356. doi:10.1891/1945-8959.14.3.343

Good, C., Aronson, J. & Inzlicht, M. (2003). Improving adolescents' standardized test performance: An intervention to reduce the effects of stereotype threat. *Journal of Applied Developmental Psychology*, *24* (6), 645–662. doi:10.1016/j.appdev.2003.09.002

Gresch, C., Strietholt, R., Kanders, M. & Solga, H. (2016). Reading-aloud versus self-administered student questionnaires: An experiment on data quality. In H.-P. Blossfeld, J. von Maurice, M. Bayer & J. Skopek (Hrsg.), *Methodological Issues of Longitudinal Surveys* (S. 561–578). Wiesbaden: Springer.

Grube, D. & Hasselhorn, M. (2006). Längsschnittliche Analysen zur Lese-, Rechtschreib- und Mathematikleistung im Grundschulalter. In I. Hosenfeld & F.W. Schrader (Hrsg.), *Schulische Leistung. Grundlagen, Bedingungen, Perspektiven* (S. 87–105). Münster: Waxmann.

Grünke, M. (2004). Lernbehinderung. In G.W. Lauth, M. Grünke & J.C. Brunstein (Hrsg.), *Interventionen bei Lernstörungen* (S. 65–77). Göttingen: Hogrefe.

Grünke, M. & Grosche, M. (2014). Lernbehinderung. In G.W. Lauth, M. Grünke & J.C. Brunstein (Hrsg.), *Interventionen bei Lernstörungen* (S. 76–89). Göttingen: Hogrefe.

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. doi:10.1016/j.learninstruc.2013.04.001

Haertel, G. D., Walberg, H. J. & Weinstein, T. (1983). Psychological models of educational performance: A theoretical synthesis of constructs. *Review of Educational Research*, *53* (1), 75–91.

Haladyna, T. M. & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice*, *23* (1), 17–27.

Händel, M., Lockl, K., Heydrich, J., Weinert, S. & Artelt, C. (2014). Assessment of metacognitive knowledge in students with special educational needs. *Metacognition and Learning*, *9* (3), 333–352. doi:10.1007/s11409-014-9119-x

Heckhausen, J. & Heckhausen, H. (Hrsg.). (2009). *Motivation und Handeln*. Heidelberg: Springer.

Helmke, A., Schneider, W. & Weinert, F. E. (1986). Quality of instruction and classroom learning outcomes: The German contribution to the IEA classroom environment study. *Teaching and Teacher Education*, *2* (1), 1–18.

Heppt, B., Haag, N., Böhme, K. & Stanat, P. (2015). The role of academic-language features for reading comprehension of language-minority students and students from low-SES families. *Reading Research Quarterly*, *50* (1), 61–82. doi:10.1002/rrq.83

Hessels, M. G. P. (2009). Estimation of the predictive validity of the HART by means of a dynamic test of geography. *Journal of Cognitive Education and Psychology*, *8* (1), 5–21. doi:10.1891/1945-8959.8.1.5

Hessels-Schlatter, C. (2002). A dynamic test to assess learning capacity in people with severe impairments. *American Journal on Mental Retardation*, *107* (5), 340–351.

Heward, W. L. (2003). Ten faulty notions about teaching and learning that hinder the effectiveness of special education. *The Journal of Special Education*, *36* (4), 186–205.

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.

Hilbe, J. M. (2011). *Negative binomial regression*. Cambridge, UK: Cambridge University Press.

Hohm, E., Jennen-Steinmetz, C., Schmidt, M. H. & Laucht, M. (2007). Language development at ten months: Predictive of language outcome and school

achievement ten years later? *European Child & Adolescent Psychiatry*, *16* (3), 149–156. doi:10.1007/s00787-006-0567-y

Holaday, B. & Turner-Henson, A. (1989). Response effects in surveys with school-age children. *Nursing Research*, *38* (4), 248–250.

Holbrook, A. L., Green, M. C. & Krosnick, J. A. (2003). Telephone versus face-to-face interviewing of national probability samples with long questionnaires: Comparisons of respondent satisficing and social desirability response bias. *Public opinion quarterly*, *67* (1), 79–125.

Hong, E., Sas, M. & Sas, J. C. (2006). Test-taking strategies of high and low mathematics achievers. *The Journal of Educational Research*, *99* (3), 144–155.

Hörmann, B. (2007). Disappearing students PISA and students with disabilities (Schulpädagogik und pädagogische Psychologie). In S.T. Hopmann, G. Brinek & M. Retzl (Hrsg.), *PISA zufolge PISA - PISA according to PISA ; hält PISA, was es verspricht? - Does PISA keep what it promises?* (S. 157–174). Wien: LIT Verlag.

Jobe, J. B. (2003). Cognitive psychology and self-reports: Models and methods. *Quality of Life Research*, *12* (3), 219–227.

Jobe, J. B. & Hermann, D. J. (1996). Implications of models of survey cognition for memory theory. In D.J. Herrmann, M. Johnson, C. McEvoy, C. Hertzog & P. Hertel (Hrsg.), *Basic and applied memory research: Volume 2. Practical Applications* (S. 193–205).

Johns, J. & Vanleirsburg, P. (1992). Teaching test-wiseness: Can test scores of special population be improved? *Reading Psychology*, *13* (1), 99–103. doi:10.1080/027027192130108

Kane, M. (2010). Validity and fairness. *Language Testing*, *27* (2), 177–182. doi:10.1177/0265532209349467

Kettler, R. J., Braden, J. P. & Beddow, P. A. (2011). Test-taking skills and their impact on accessibility for all students. In S.N. Elliott, R.J. Kettler, P.A. Beddow & A. Kurz (Hrsg.), *Handbook of Accessible Achievement Tests for All Students* (S. 147–159). New York, NY: Springer.

Klauer, K. J. & Lauth, G. W. (1997). Lernbehinderungen und Leistungsschwierigkeiten bei Schülern (Enzyklopädie der Psychologie, Serie Pädagogische Psychologie). In F.E. Weinert (Hrsg.), *Psychologie des Unterrichts und der Schule* (S. 701–738). Göttingen: Hogrefe.

Klieme, E., Artelt, C., Hartig, J., Jude, N., Köller, O., Prenzel, M. et al. (Hrsg.). (2010). *PISA 2009: Bilanz nach einem Jahrzehnt*. Münster: Waxmann.

KMK - Sekretariat der Ständigen Konferenz der Kultusminister der Länder (Hrsg.). (1999). *Empfehlungen zum Förderschwerpunkt Lernen* (Beschluss der Kultusministerkonferenz vom 01.10.1999.). Berlin. Verfügbar unter: http://www.kmk.org/fileadmin/Dateien/pdf/PresseUndAktuelles/2000/sopale.pdf

KMK - Sekretariat der Ständigen Konferenz der Kultusminister der Länder. (2017). *Definitionenkatalog zur Schulstatistik 2017*. Berlin. Verfügbar unter: https://www.kmk.org/fileadmin/Dateien/pdf/Statistik/Defkat2017.pdf

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. doi:10.1007/s11577-014-0253-x

Koch, K. (2007). Soziokulturelle Benachteiligung (Handbuch Sonderpädagogik). In J. Walter & F.B. Wember (Hrsg.), *Sonderpädagogik des Lernens* (S. 104–116). Göttingen: Hogrefe.

Koretz, D. & Barton, K. (2004). Assessing students with disabilities: Issues and evidence. *Educational Assessment*, *9* (1–2), 29–60. doi:10.1080/10627197.2004.9652958

Kotzerke, M., Röhricht, V., Weinert, S. & Ebert, S. (2013). Sprachlich-kognitive Kompetenzunterschiede bei Schulanfängern und deren Auswirkungen bis Ende der Klassenstufe 2. In G. Faust (Hrsg.), *Einschulung. Ergebnisse aus der Studie „Bildungsprozesse, Kompetenzentwicklung und Selektionsentscheidungen im Vorschul- und Schulalter (BiKS)* (S. 111–135). Münster: Waxmann.

Kränzl-Nagl, R. & Wilk, L. (2000). Möglichkeiten und Grenzen standardisierter Befragungen unter besonderer Berücksichtigung der Faktoren soziale und personale Wünschbarkeit. In F. Heinzel (Hrsg.), *Methoden der Kindheitsforschung: Ein Überblick über Forschungszugänge zur kindlichen Perspektive* (S. 59–84). Weinheim: Juventa.

Kretschmann, R. (2006). Diagnostik bei Lernbehinderung. In U. Petermann & F. Petermann (Hrsg.), *Diagnostik sonderpädagogischen Förderbedarfs (Tests und Trends, n.F. Bd. 5)* (S. 139–162). Göttingen: Hogrefe.

Kretschmann, R. (2007). Lernschwierigkeiten, Lernstörungen und Lernbehinderung (Handbuch Sonderpädagogik). In J. Walter & F.B. Wember (Hrsg.), *Sonderpädagogik des Lernens* (Band 2, S. 4–32). Göttingen: Hogrefe.

Kreuter, F., Maaz, K. & Watermann, R. (2006). Der Zusammenhang zwischen der Qualität von Schülerangaben zur sozialen Herkunft und den Schulleistungen. In K.-S. Rehbger (Hrsg.), *Soziale Ungleichheit, Kulturelle Unterschiede. Verhandlungen des 32. Kongresses der Deutschen Gesellschaft für Soziologie in München 2004* (S. 3465–3478). Frankfurt am Main: Campus Verlag.

Kristen, C., Olczyk, M. & Will, G. (2016). Identifying immigrants and their descendants in the National Educational Panel Study. In H.-P. Blossfeld, J. von Maurice, M. Bayer & J. Skopek (Hrsg.), *Methodological Issues of Longitudinal Surveys* (S. 195–211). Wiebaden: Springer.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied cognitive psychology*, *5* (3), 213–236.

Krosnick, J. A. (2000). The threats of satisficing in surveys: The shortcuts respondents take in answering questions. *Survey Methods Newsletter*, *20* (1), 4–8.

Krosnick, J. A., Holbrook, A. L., Berent, M. K., Carson, R. T., Hanemann, W. M., Kopp, R. J. et al. (2001). The impact of "no opinion" response options on data quality. *Public Opinion Quarterly*, *66* (3), 371–403. doi:10.1086/341394

Krowatschek, D., Krowatschek, G. & Reid, C. (2011). *Marburger Konzentrationstraining (MKT) für Schulkinder*. Dortmund: Borgmann.

Krull, J., Wilbert, J. & Hennemann, T. (2014). Soziale Ausgrenzung von Erstklässlerinnen und Erstklässlern mit sonderpädagogischem Förderbedarf im Gemeinsamen Unterricht. *Empirische Sonderpädagogik*, *6* (1), 59–75.

Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: a literature review. *Quality & Quantity*, *47* (4), 2025–2047. doi:10.1007/s11135-011-9640-9

Kuhl, P., Stanat, P., Lütje-Klose, B., Gresch, C., Pant, H. A. & Prenzel, M. (Hrsg.). (2015). *Inklusion von Schülerinnen und Schülern mit sonderpädagogischem Förderbedarf in Schulleistungserhebungen*. Wiesbaden: Springer.

Kukla, A. (1974). Performance as a function of resultant achievement motivation (perceived ability) and perceived difficulty. *Journal of Research in Personality*, *7*, 374–383.

Kulik, J. A., Kulik, C.-L. C. & Bangert, R. L. (1984). Effects of practice on aptitude and achievement test scores. *American Educational Research Journal*, *21* (2), 435–447.

Lauth, G. W. (2000). Lernbehinderungen. In J. Borchert (Hrsg.), *Handbuch der Sonderpädagogischen Psychologie* (S. 21–31). Göttingen: Hogrefe.

Lauth, G. W., Brunstein, J. C. & Grünke, M. (2014). Lernstörungen im Überblick: Arten, Klassifikation, Verbreitung und Erklärungsperspektiven. In G.W. Lauth, M. Grünke & J.C. Brunstein (Hrsg.), *Interventionen bei Lernstörungen: Förderung, Training und Therapie in der Praxis* (S. 17–31). Göttingen: Hogrefe.

Lauth, G. W. & Grünke, M. (2005). Interventionen bei Lernstörungen. *Monatsschrift Kinderheilkunde*, *153* (7), 640–648. doi:10.1007/s00112-005-1167-5

Lauth, G. W. & Wiedl, K. H. (1985). Zur Veränderbarkeit der Testleistung im CFT-20 durch Instruktionsintensivierung. *Diagnostica*, *31* (2), 200–209.

de Leeuw, E. D. (2011). *Improving Data Quality when Surveying Children and Adolescents: Cognitive and Social Development and its Role in Questionnaire Construction and Pretesting*. Naantali Finland.

Lehmann, R. & Hoffmann, E. (2009). *BELLA. Berliner Erhebung arbeitsrelevanter Basiskompetenzen von Schülerinnen und Schülern mit Förderbedarf „Lernen"*. Münster: Waxmann.

Lenhard, W. & Schneider, W. (2006). *ELFE 1-6: Ein Leseverständnistest für Erst- bis Sechstklässler*. Göttingen: Hogrefe.

Lipski, J. (2000). Zur Verlässlichkeit der Angaben von Kindern in standardisierten Befragungen. In F. Heinzel (Hrsg.), *Methoden der Kindheitsforschung: Ein Überblick über Forschungszugänge zur kindlichen Perspektive* (S. 77–86). Weinheim: Juventa.

Lloyd, J. W., Keller, C. & Hung, L. (2007). International understanding of learning disabilities. *Learning Disabilities Research and Practice*, *22* (3), 159.

Looker, E. D. (1989). Accuracy of proxy reports of parental status characteristics. *Sociology of Education*, *62* (4), 257. doi:10.2307/2112830

Lozano, L. M., García-Cueto, E. & Muñiz, J. (2008). Effect of the number of response categories on the reliability and validity of rating scales. *Methodology*, *4* (2), 73–79. doi:10.1027/1614-2241.4.2.73

Maaz, K., Kreuter, F. & Watermann, R. (2006). Schüler als Informanten? Die Qualität von Schülerangaben zum sozialen Hintergrund. *Herkunftsbedingte Disparitäten im Bildungswesen: Differenzielle Bildungsprozesse und Probleme der Verteilungsgerechtigkeit* (S. 31–59). Wiesbaden: Springer.

Marsh, H. W. (1986). Negative item bias in ratings scales for preadolescent children: A cognitive-developmental phenomenon. *Developmental Psychology*, *22* (1), 37–49. doi:10.1037/0012-1649.22.1.37

Marsh, H. W. (2005). Big-Fish-Little-Pond Effect on academic self-concept. *Zeitschrift für Pädagogische Psychologie*, *19* (3), 119–129. doi:10.1024/1010-0652.19.3.119

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50* (9), 741.

Millman, J., Bishop, C. H. & Ebel, R. (1965). An analysis of test-wiseness. *Educational and Psychological Measurement*, *25* (3), 707–726.

Möller, J. & Bonerad, E.-M. (2007). Fragebogen zur habituellen Lesemotivation. *Psychologie in Erziehung und Unterricht*, *54*, 259–267.

Möller, J., Streblow, L. & Pohlmann, B. (2002). Leistung und Selbstkonzept bei lernbehinderten Schülern. *Heilpädagogische Forschung*, *28* (3), 132–139.

Müller, K., Prenzel, M., Sälzer, C., Mang, J., Heine, J.-H. & Gebhardt, M. (2017). Wie schneiden Schülerinnen und Schüler an Förderschulen bei PISA ab? Analysen aus der PISA 2012-Zusatzerhebung zu Jugendlichen mit sonderpädagogischem Förderbedarf. *Unterrichtswissenschaft, 45* (2), 175–192.

Müller, S., Stubbe, T. C. & Bos, W. (2013). Leistungsheterogenität angemessen berücksichtigen: Konzeption der Kompetenzmessung an Förderschulen mit dem Förderschwerpunkt Lernen im Rahmen von PARS-F. *Jahrbuch der Schulentwicklung. Daten, Beispiele und Perspektiven*, *17*, 264–296.

Neher, K. M. & Sowarka, D. (1999). Vergleich von zwei Kennwerten der kognitiven Plastizität bei der Früherkennung dementieller Erkrankungen. *Zeitschrift für Neuropsychologie*, *10* (3), 153–167. doi:10.1024//1016-264X.10.3.153

Nusser, L., Carstensen, C. H. & Artelt, C. (2015). Befragung von Schülerinnen und Schülern mit sonderpädagogischem Förderbedarf Lernen: Ergebnisse zur Messinvarianz. *Empirische Sonderpädagogik*, *7* (2), 99–116.

Oerter, R. (1980). *Psychologie des Denkens*. Donauwörth: Auer.

Olczyk, M., Will, G. & Kristen, C. (2016). *Immigrants in the NEPS: Identifying generation status and group of origin* (NEPS Survey Paper No. 4.). Bamberg: Leibniz-Institut für Bildungsverläufe, Nationales Bildungspanel. Verfügbar unter: https://www.neps-data.de/Portals/0/Survey%20Papers/SP_IV.pdf

Oser, F. & Biedermann, H. (2006). PISA für den Rest: Lehr- und Lernbehinderung und ihre schulische Anstrengungslogik. *Vierteljahresschrift für Heilpädagogik und ihre Nachbargebiete*, *75*, 4–12.

Pant, H. A., Stanat, P., Schroeders, U., Roppelt, A., Siegle, T. & Pöhlmann, C. (2013). *IQB-Ländervergleich 2012. Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I.* Münster: Waxmann.

Pepper, D. (2011). Assessing key competences across the curriculum—and Europe. *European Journal of Education*, *46* (3), 335–353.

Pintrich, P. R., Anderman, E. M. & Klobucar, C. (1994). Intraindividual differences in motivation and cognition in students with and without learning disabilites. *Journal of Learning Disabilities*, *27* (6), 360–370.

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. doi:10.3102/00346543071001053

Pohl, S., Südkamp, A., Hardt, K., Carstensen, C. H. & Weinert, S. (2016). Testing students with special educational needs in large-scale assessments – Psychometric properties of test scores and associations with test taking behavior. *Frontiers in Psychology*, *7*. doi:10.3389/fpsyg.2016.00154

Renner, G. & Mickley, M. (2015). Berücksichtigen deutschsprachige Intelligenztests die besonderen Anforderungen von Kindern mit Behinderung? *Praxis der Kinderpsychologie und Kinderpsychiatrie*, *64*, 88–103.

Ritter, S. & Idol-Maestas, L. (1986). Teaching middle school students to use a test-taking strategy. *The Journal of Educational Research*, *79* (6), 350–357.

Robinson, E. J. & Whittaker, S. J. (1987). Children' conceptions of relations between messages, meanings and reality. *British Journal of Developmental Psychology*, *5* (1), 81–90. doi:10.1111/j.2044-835X.1987.tb01044.x

Rothman, R. W. & Cohen, J. (1988). Teaching test taking skills. *Intervention in School and Clinic*, *23* (4), 341–348. doi:10.1177/105345128802300401

Roussos, L. A. & Stout, W. (2004). Differential item functioning analysis: Detecting DIF item and testing DIF hypotheses. In D. Kaplan (Hrsg.), *The Sage handbook of quantitative methodology for the social sciences* (S. 107–115). Thousand Oaks, CA: Sage.

Sälzer, C., Gebhardt, M., Müller, K. & Pauly, E. (2015). Der Prozess der Feststellung sonderpädagogischen Förderbedarfs in Deutschland. 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. 129–152). Wiesbaden: Springer.

Saris, W., Revilla, M., Krosnick, J. A. & Shaeffer, E. M. (2010). Comparing questions with agree/disagree response options to questions with item-specific response options. *Survey Research Methods*, *4* (1), 61–79. doi:10.18148/srm/2010.v4i1.2682

Sauer, S., Ide, S. & Borchert, J. (2007). Zum Selbstkonzept von Schülerinnen und Schülern an Förderschulen und in integrativer Beschulung: Eine Vergleichsuntersuchung. *Heilpädagogische Forschung*, *33* (3), 135–142.

Schlee, J. (1974). Rezeptive Sprachbarriere im Unterricht. *Bildung und Erziehung*, *27*, 244–256.

Schrader, F.-W. & Helmke, A. (2008). Determinanten der Schulleistung. In M. Schweer (Hrsg.), *Lehrer-Schüler-Interaktion: Inhaltsfelder, Forschungsperspektiven und methodische Zugänge* (S. 285–302). Wiesbaden.

Schuchardt, K., Brandenburg, J., Fischbach, A., Büttner, G., Grube, D., Mähler, C. et al. (2015). Die Entwicklung des akademischen Selbstkonzeptes bei Grundschulkindern mit Lernschwierigkeiten. *Zeitschrift für Erziehungswissenschaft*, *18* (3), 513–526. doi:10.1007/s11618-015-0649-z

Schurtz, I. M., Pfost, M., Nagengast, B. & Artelt, C. (2014). Impact of social and dimensional comparisons on student's mathematical and English subject-interest at the beginning of secondary school. *Learning and Instruction*, *34*, 32–41. doi:10.1016/j.learninstruc.2014.08.001

Schwab, S. (2014). Haben sie wirklich ein anderes Selbstkonzept? Ein empirischer Vergleich von Schülern mit und ohne sonderpädagogischen Förderbedarf im Bereich Lernen. *Zeitschrift für Heilpädagogik*, *65* (3), 116–121.

Schwab, S. & Helm, C. (2015). Überprüfung von Messinvarianz mittels CFA und DIF-Analysen. *Empirische Sonderpädagogik*, *7* (3), 175–193.

Schwabe, F., McElvany, N. & Trendtel, M. (2015). Reading skills of students in different school tracks: Systematic (dis)advantages based on item formats in large scale assessments. *Zeitschrift für Erziehungswissenschaft*, *18* (4), 781–801. doi:10.1007/s11618-015-0645-3

Schwarz, N. (2003). Self-reports in consumer research: The challenge of comparing cohorts and cultures. *Journal of Consumer Research*, *29*, 588–594.

Schwinger, M., Wild, E., Lütje-Klose, B., Grunschel, C., Stranghöner, D., Yotyodying, S. et al. (2015). Wie können motivationale und affektive Merkmale bei Kindern mit sonderpädagogischem Förderbedarf valide erfasst werden? 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. 273–300). Wiesbaden: Springer.

Scruggs, T. E. (1984). *The administration and interpretation of standardized achievement tests with learning disabled and behaviorally disordered elementary school children. Final report*. Salt Lake City: Developmental Center for the Handicapped, University of Utah. Verfügbar unter: http://files.eric.ed.gov/fulltext/ED311652.pdf

Scruggs, T. E., Bennion, K. & Lifson, S. (1985). An analysis of children's strategy use on reading achievement tests. *The Elementary School Journal*, *85* (4), 479–484.

Scruggs, T. E., White, K. R. & Bennion, K. (1986). Teaching test-taking skills to elementary-grade students: A meta-analysis. *The Elementary School Journal*, *87* (1), 68–82.

Shuell, T. J. (1986). Cognitive conceptions of learning. *Review of Educational Research*, *56* (4), 411–436.

Siegler, R. S. (1983). How Knowledge Influences Learning: What children already know about scientific and mathematical concepts influences how they acquire additional information. *American Scientist*, *71* (6), 631–638.

Sireci, S. G., Scarpati, S. E. & Li, S. (2005). Test accommodations for students with disabilities: An analysis of the interaction hypothesis. *Review of Educational Research*, *75* (4), 457–490.

Spencer, S. J., Steele, C. M. & Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology*, *35* (1), 4–28. doi:10.1006/jesp.1998.1373

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*. Münster: Waxmann.

Steenkamp, J.-B. & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, *25* (1), 78–107.

Steinmetz, H., Schmidt, P., Tina-Booh, A., Wieczorek, S. & Schwartz, S. H. (2009). Testing measurement invariance using multigroup CFA: differences between educational groups in human values measurement. *Quality & Quantity*, *43* (4), 599–616. doi:10.1007/s11135-007-9143-x

Stern, E. (1998). Die Entwicklung schulbezogener Kompetenzen: Mathematik. In F.E. Weinert (Hrsg.), *Entwicklung im Kindesalter* (S. 95–114). Weinheim: Beltz.

Stern, E. (2009). The development of mathematical competencies: Sources of individual differences and their developmental trajectories. In W. Schneider & M. Bullock (Hrsg.), *Human development from early childhood to early adulthood: Evidence from the Munich Longitudinal Study on the Genesis of Individual Competencies (LOGIC)* (S. 221–236). Mahwah, NJ: Erlbaum.

Stocké, V., Blossfeld, H.-P., Hoenig, K. & Sixt, M. (2011). Social inequality and educational decisions in the life course. *Zeitschrift für Erziehungswissenschaft*, *14* (S2), 103–119. doi:10.1007/s11618-011-0193-4

Strack, F. (1994). *Zur Psychologie der standardisierten Befragung* (Lehr- und Forschungstexte Psychologie) (Band 48). Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-78890-1

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.

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), 1–26.

Thompson, S. J., Johnstone, C. J. & Thurlow, M. L. (2002). *Universal design applied to large scale assessments* (Synthesis Report 44.). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Verfügbar unter: https://nceo.info/Resources/publications/OnlinePubs/Synthesis44.html

Tiedemann, J. & Faber, G. (1990). Der langfristige Stellenwert mütterlicher Erziehungsmerkmale und kognitiver Kindkompetenzen für die Leistungsentwicklung in der Grundschule: Ergebnisse einer vierjährigen Längsschnittuntersuchung. *Unterrichtswissenschaft*, *18*, 71–89.

Torgesen, J. K. (1977). The role of nonspecific factors in the task performance of learning disabled children: A theoretical assessment. *Journal of Learning Disabilities*, *10* (1), 33–40.

Tourangeau, R., Rips, L. J. & Rasinski, K. (2000). *The Psychology of Survey Response*. Cambridge: Cambridge University Press.

Tzuriel, D. (2012). Dynamic assessment of learning potential. In M.M.C. Mok (Hrsg.), *Self-Directed Learning Oriented Assessments in the Asia-Pacific* (S. 235–255). Dordrecht: Springer Netherlands. doi:10.1007/978-94-007-4507-0_13

Vaughn, S., Gersten, R. & Chard, D. J. (2000). The underlying message in LD intervention research: Findings from research syntheses. *Exceptional Children*, *67* (1), 99–114.

Weinert, F. E. (1997). Unterschiedliche Lernfähigkeiten erfordern variable Unterrichtsmethoden. *Lernmethoden, Lehrmethoden: Wege zur Selbstständigkeit*, *15*, 50–52.

Weinert, F. E. (2001). Vergleichende Leistungsmessung in Schulen – eine umstrittene Selbstverständlichkeit. In F.E. Weinert (Hrsg.), *Leistungsmessungen in Schulen* (S. 17–31). Weinheim: Beltz.

Weinert, F. E., Helmke, A. & Schneider, W. (1989). Individual differences in learning performance and in school achievement: Plausible parallels and unexplained discrepancies. *Learning and Instruction*, 461–479.

Weinert, S., Artelt, C., Prenzel, M., Senkbeil, M., Ehmke, T. & Carstensen, C. H. (2011). Development of competencies across the life span. *Zeitschrift für Erziehungswissenschaft*, *14* (S2), 67–86. doi:10.1007/s11618-011-0182-7

Weinert, S. & Ebert, S. (2013). Spracherwerb im Vorschulalter: Soziale Disparitäten und Einflussvariablen auf den Grammatikerwerb. *Zeitschrift für Erziehungswissenschaft*, *16* (2), 303–332. doi:10.1007/s11618-013-0354-8

Werning, R. (2010). Inklusion zwischen Innovation und Überforderung. *Zeitschrift für Heilpädagogik*, *8*, 284–291.

Wiedl, K. H., Waldorf, M. & Schöttke, H. (2007). Strategien zur Analyse komplexer kognitiver Beeinträchtigungen. In F. Linderkamp & M. Grünke (Hrsg.), *Lern-und Verhaltensstörungen-Genese, Diagnostik und Intervention* (S. 66–77). Weinheim: Beltz.

Wilbert, J. (2010a). *Förderung der Motivation bei Lernstörungen*. Stuttgart: Kohlhammer.

Wilbert, J. (2010b). Stereotype-Threat Effekte bei Schülern des Förderschwerpunkts Lernen. *Heilpädagogische Forschung*, *XXXVI* (4), 154–161.

Wild, E., Schwinger, M., Lütje-Klose, B., Yotyodying, S., Gorges, J., Stranghöner, D. et al. (2015). Schülerinnen und Schüler mit dem Förderschwerpunkt Lernen in inklusiven und exklusiven Förderarrangements: Erste Befunde des BiLieF-Projektes zu Leistung, sozialer Integration, Motivation und Wohlbefinden. *Unterrichtswissenschaften*, *43* (1), 7–21.

Wirtz, M. & Caspar, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität: Methoden zur Bestimmung und Verbesserung der Zuverlässigkeit von Einschätzungen mittels Kategoriensystemen und Ratingskalen*. Göttingen: Hogrefe.

Wocken, H. (2005). Andere Länder, andere Schüler? Vergleichende Untersuchung von Förderschülern in den Bundesländern Brandenburg, Hamburg und Niedersachsen. Verfügbar unter: http://bidok.uibk. ac.at/download/wocken-forschungsbericht.pdf

Wohlkinger, F., Ditton, H., von Maurice, J., Haugwitz, M. & Blossfeld, H.-P. (2011). Motivational concepts and personality aspects across the life course. *Zeitschrift für Erziehungswissenschaft*, *14* (S2), 155–168. doi:10.1007/s11618-011-0184-5

Wong, B. Y. L. (1991). The relevance of metacognition to learning disabilities. In B.Y.L. Wong (Hrsg.), *Learning about Learning Disabilities* (S. 231–258). San Diego: Academic Press.

Wong, B. Y. L., Wong, R. & LeMare, L. (1982). The effects of knowledge of criterion task on comprehension and recall in normally achieving and learning disabled children. *Journal of Educational Research*, *76*, 119–126.

Zielinski, W. (1996). Lernschwierigkeiten (Enzyklopädie der Psychologie, Serie Pädagogische Psychologie). In F.E. Weinert (Hrsg.), *Psychologie des Lernens und der Instruktion* (Band 2, S. 369–402). Göttingen: Hogrefe.

Anhang

Beitrag 1

Nusser, L. & Weinert, S. (in Druck). Instructions in test-taking: An appropriate approach for students with special educational needs. *Journal of Cognitive Education and Psychology.*

Beitrag 2

Nusser, L., Heydrich, J., Carstensen, C. H., Artelt, C., & Weinert, S. (2016). Validity of survey data of students with special educational needs – Results from the National Educational Panel Study. In H.-P- Blossfeld, J. von Maurice, M. Bayer & J. Skopek (Hrsg.), *Methodological Issues of Longitudinal Surveys: The Example of the National Educational Panel Study* (S. 251-266). Wiesbaden: Springer.

Beitrag 3

Nusser, L., Carstensen, C. H. & Artelt, C. (2015). Befragung von Schülerinnen und Schülern mit sonderpädagogischem Förderbedarf Lernen: Ergebnisse zur Messinvarianz. *Empirische Sonderpädagogik, 7* (2), 99-116.

Beitrag 4

Nusser, L. & Wolter, I. (2016). There's plenty more fish in the sea. Das akademische Selbstkonzept von Schülerinnen und Schülern mit sonderpädagogischem Förderbedarf Lernen in integrativen und segregierten Schulsettings. *Empirische Pädagogik, 30* (1), 130-143.

Beitrag 1

Nusser, L. & Weinert, S. (in Druck). Instructions in test-taking: An appropriate approach for students with special educational needs. *Journal of Cognitive Education and Psychology*.

Appropriate Test-Taking Instructions for Students with Special Educational Needs

Lena Nusser and Sabine Weinert

University of Bamberg, Germany

Author Note

Lena Nusser and Sabine Weinert, Department Psychology I – Developmental Psychology, University of Bamberg, Germany

This paper uses data from the National Educational Panel Study (NEPS). From 2008 to 2013, NEPS data were collected as part of the Framework Programme for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research (BMBF). As of 2014, the NEPS survey is carried out by the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg in cooperation with a nationwide network.

The authors thank Christina van Kraayenoord for her helpful comments on an earlier version of this paper and Markus Messingschlager for his support in planning and implementing the study.

Correspondence concerning this article should be addressed to Lena Nusser, University of Bamberg, 96045 Bamberg, Germany, E-Mail: lena.nusser@uni-bamberg.de

Abstract

If children fail to understand test instructions, measurements of their competence may be unfair and invalid. This is especially relevant for students with special educational needs (SEN), because they face greater challenges in comprehending instructions. Two interventions were designed (a) to facilitate the comprehension of test requirements by giving students intensified test instructions based on the principles of Direct Instruction and (b) to enhance students' attention by engaging them in physical activity immediately before receiving the test instructions. 348 students with SEN aged 8–12 years were randomly assigned to one of the two experimental conditions or to a control group. Results showed that even after statistically controlling for highly relevant variables (reading speed, basic cognitive skills), students participating in the interventions performed better in a reading comprehension test than controls. As hypothesized, the intensified test instructions significantly reduced the number of responses that were not compliant with test instructions. In conclusion, the study shows the importance of adapting test instructions for students with SEN, and it proposes interventions that can be implemented in many other studies and assessments.

Keywords: special educational needs, test instructions, intervention, reading comprehension, ELFE 1-6

Appropriate Test-Taking Instructions for Students with Special Educational Needs

In both large- and small-scale studies, standardized assessment conditions are crucial. They should ensure an objective, reliable, and comparable assessment of participants' individual competencies, and be designed to guarantee that test scores reflect interindividual differences instead of variations in assessment conditions (Pitoniak & Royer, 2001). In this context, test instructions play an essential role. In terms of objectivity, standardized test instructions ensure that all participants share the same information about a test (Sax, 1997) by informing them on how they are expected to process the test material and how to respond to the tasks (Vukovich, 1971).

As a cognitive activity, comprehending test instructions is closely related to intelligence—particularly verbal intelligence—and the speed of information processing (Zielinski, 1996). Comprehending test instructions can be understood as a bottleneck that precedes all tasks (Oerter, 1980) and allows participants access to the test requirements. Kettler, Elliott, and Beddow (2009) frame this access as an interaction between a person and a test. Such accessibility might be impeded by individual characteristics of the participants, and this could reduce the reliability and validity of the test results. The effect of more general cognitive abilities on test performance (Artelt, Schiefele, & Schneider, 2001; Banks & Eaton, 2014) may thus unfold at the very beginning of test taking. If participants fail to understand the instructions, their test results cannot be interpreted as an indicator of their competencies.

However, there is both theoretical and empirical evidence that standardized test instructions developed for measures used with students in regular schools may not be equally appropriate for all students. For students with special educational needs in the area of learning (SEN-L), this factor may represent a group-specific obstacle in a test situation. In Germany, students are diagnosed as students with SEN-L if their learning, learning behavior,

and academic achievement are impaired (KMK, 1999) and if their cognitive abilities are below average (Grünke & Grosche, 2014). Thus, many of these students receive special educational services including specific programs to support learning and achievement progress in accordance with a special curriculum (KMK, 1999). The description of students with SEN-L corresponds to the international definition of learning disabilities by Lloyd, Keller, and Hung (2007) that characterizes these students as having significant academic difficulties in school. In Germany—despite the UN Convention on the Rights of Persons with Disabilities that calls for an inclusive educational system (Werning, 2010)—most members of this student group do not attend regular schools but are enrolled in special schools (Autorengruppe Bildungsbericht, 2016).

Given the characteristics of students with SEN-L documented in the research literature, their comprehension of standardized test instructions represents a potential obstacle when it comes to interpreting their test results. For example, Zielinksi (1996) refers specifically to a poor ability to understand verbal instructions as an internal condition of students with learning difficulties. Moreover, students with SEN-L have shorter attention spans, need more study time, and require multiple repetitions to acquire new knowledge. They are characterized by limitations in their ability to cope with learning processes, and they have been shown extensively to have less adequate strategies for acquiring and retrieving relevant information and for organizing and controlling learning processes than their peers without SEN-L (Brown & Campione, 1986). Children with SEN-L apply more ineffective strategies; for example, they prefer to guess rather than systematically inspect all aspects of the material (Scruggs, Bennion, & Lifson, 1985a). In addition, their application and transfer of acquired knowledge and strategies to subsequent situations and/or similar tasks is often limited (Wong, 1994).

Empirical evidence supports these hypothesized difficulties in understanding test instructions. Feasibility studies with students with SEN-L (Heydrich, Weinert, Nusser, Artelt, & Carstensen, 2013) within the German National Educational Panel Study (NEPS; Blossfeld, Roßbach, & von Maurice, 2011) have shown that their comprehension of test instructions is restricted. When tested for their reading comprehension, 5th-grade students with SEN-L attending special schools showed significantly higher rates of responses that were not compliant with the test instructions (e.g., multiple responses when one choice had to be selected) than their peers without SEN-L in regular schools (Südkamp, Pohl, & Weinert, 2015). Likewise, Scruggs, Bennion, and Lifson (1985b) found that students with learning disabilities paid less attention to item formats and task demands than their peers without disabilities. However, comprehension of task demands is crucial to performance (Butler & Winne, 1995). In this vein, Wong (1991) has suggested that reduced task comprehension is associated with low test performance in students with learning disabilities. These assumptions are in line with studies showing that mean test scores can be improved by making specific modifications to the test instructions.

For instance, when given even short phases of familiarization with test materials, students with learning difficulties showed enhanced test performance on a matrices test assessing basic learning capacity (Hessels, 2009). Hessels' study (2009) trained aspects of encoding information, making systematic comparisons, abstracting concepts, and inferring relations, and this training was accompanied by additional task examples and mediated group discussion during a short 20–30 min introduction period. Lauth and Wiedl (1985) adapted test instructions to measure basic cognitive skills by using features of the direct perception of task requirements and practical activity instead of the linguistic elements of standard test instructions. Results showed that students with SEN-L performed better after the intensified compared to the standard test instructions. A study by Wong, Wong, and LeMare (1982)

showed that providing explicit information about task requirements led to improved comprehension and enhanced test scores on recall tasks in students with and without learning disability. The authors concluded that clear test instructions and communication about task requirements "substantially facilitate performance" (Wong et al., 1982, p. 125).

Drawing on previous research, the present experimental study compared two brief interventions in test instructions and test administration with the standard instructions for a reading comprehension test in students with SEN-L.

Intervention 1—as an intensified form of test instructions—explicitly aimed to enhance the students' comprehension. The test instructions were adapted and presented by a pedagogical agent to additionally promote motivation (cf. Bendel, Schnöring, & Back, 2002). Students were given repeated instructional hints on how to respond to the test items and went through a phase of item familiarization (Hessels, 2009). During this instructional phase, students were given additional task examples similar to the tasks they would encounter in the upcoming test. The introduction and explanation of the additional examples followed the principles of Direct Instruction (Engelmann, 1980). Direct Instruction—as a highly structured teaching method—has been shown to be particularly effective for students with learning difficulties (Swanson, 2001). Overall, these instructional amplifications gave the children repeated opportunities to obtain relevant information about the test in a child-oriented and motivating manner. Furthermore, it offered them the chance to become familiar with the test requirements as well as the item format.

Intervention 2 aimed to enhance the students' general attention to the task. Students started the session with a short story motivating playful physical activities. Activities were adapted from a concentration training designed to relax students and promote their concentration (Krowatschek, Krowatschek, & Schmidt, 2011). Kubesch and Walk (2009)

have argued that games combined with physical activities and a variety of changing rules may promote students' learning efficiency.

Finally, a group of students in a control condition received the standard test instructions without any modifications.

We investigated the effects of the supplemented and optimized test instructions on relevant aspects of test-taking behavior and performance. These focused on the following research questions:

1. Do students solve more items correctly and show an enhanced test performance compared to controls after participating in one of the two interventions?

2. Are students more attentive and able to process more items during test time after participating in one of the interventions compared to controls?

3. Is there a decline in the number of responses that do not comply with the test instructions after participating in one of the interventions compared to controls?

For the indicator *test performance,* we expected that students participating in either intervention would perform better (i.e., answer more items correctly) than their peers in the control group. We hypothesized that the intensified test instructions based on the principles of Direct Instruction in Intervention 1 would minimize barriers to accessibility and therefore prepare all students more comprehensively for the test situation. We expected that this would enhance test performance. In addition, we hypothesized that the improved attention due to engaging in physical activities would lead to better test performance in Intervention 2. We expected that the same reasoning could be applied to the second test-taking indicator, *test processing* (number of processed items).

For the third indicator, *test compliance* (number of responses not compliant with the test instructions) we anticipated differential effects. Whereas Intervention 2 may not necessarily lead to an improvement in correct response behavior, Intervention 1 focused

specifically on this issue. Therefore, we assumed that students in Intervention 1 would provide fewer responses that were not compliant with the test instructions compared to students in the control group and in Intervention 2.

Method

Sample and Design

The sample consisted of 369 Grade 3 and 4 students with SEN-L. Twenty-one students had to be excluded from further analysis due to missing test scores. Students were enrolled in 49 special schools (grade 3 and 4) located in four German federal states. They were aged 8–12 years (M = 9.72; SD = 0.79) and 34% were female. Informed consent was obtained from all parents whose children participated in the study.

Each student was randomly assigned to one of the two experimental conditions or to the control group. Because of the rather small class sizes in special schools (M = 10.3; KMK, 2016), only two out of three conditions could be realized within each institution. Overall, 91 testing sessions took place in the fall of 2013. The control group and Intervention groups 1 and 2 did not differ significantly with respect to gender, age, reading speed, and basic nonverbal cognitive skills (perceptual speed, reasoning; See Table 1 and Measures).

Table 1

Descriptive Statistics of Student Characteristics

	N [students]	N [test groups]	Age [years]	N [female]	Reading speed	Perceptual speed	Reasoning
Control group	124	32	9.65 (0.81)	35 (28.2%)	15.55 (9.99)	26.17 (8.72)	4.35 (2.84)
Intervention 1	108	30	9.73 (0.77)	39 (36.1%)	12.97 (8.10)	25.09 (8.18)	4.65 (3.14)
Intervention 2	116	29	9.81 (0.81)	45 (38.8%)	14.21 (9.44)	26.77 (8.18)	4.80 (0.00)
Test statistics (Kruskal–Wallis test, χ^2 test)			$H(2) = 2.28$, $p = .32$	$\chi^2(2) = 3.23$, $p = .20$	$H(2) = 3.16$, $p = .21$	$H(2) = 2.13$, $p = .35$	$H(2) = 1.94$, $p = .38$

Note. Standard deviations in parentheses.

Measures

In addition to giving students a reading comprehension test with three different versions of test instructions, we assessed reading speed and basic cognitive skills as relevant control variables.

Reading comprehension. We administrated a subtest of a German standardized and validated reading test for students from German Grades 1 through 6 (ELFE 1-6, Leseverständnis für Erst- bis Sechstklässler [Reading comprehension test for 1st- to 6th-graders], Lenhard & Schneider, 2006). The test instrument consists of 13 short texts and 20 items representing the cognitive requirements for recognizing given information, detecting anaphoric relations, and drawing conclusions. The time limit is 7 min (Lenhard & Schneider, 2006).

Reading speed. We assessed basic reading skills with a standardized test for reading speed (subtest of the SLS 1-4, Salzburger Lesescreening für die Klassenstufen 1-4 [Salzburg screening test for reading in elementary school students], Mayringer & Wimmer, 2003). This measure focuses on the speed of decoding sentences, which is assumed to be an important prerequisite to reading comprehension on the sentence and text level (Lenhard & Artelt, 2009). Students read simple sentences and judge the content of each sentence as true or false. Overall, 70 sentences are administered within 3 min and the number of sentences solved is taken as an indicator of reading speed (Mayringer & Wimmer, 2003).

Basic nonverbal cognitive skills. We assessed basic cognitive skills with two short indicators of general cognitive functioning: perceptual speed and reasoning. These cognitive skills are interrelated with the acquisition of domain-specific cognitive competencies (Weinert et al., 2011). We measured perceptual speed with a picture-symbol test. Students have to draw symbols corresponding to the given pictures using an answer key with matching pictures and symbols as a guide. The test contains two sets of 21 items with a time limit of 30 s per set (Lang, Kamin, Rohr, Stünkel, & Willinger, 2014). We assessed deductive

reasoning abilities with a traditional matrices test containing patterns of geometric elements. From six options, respondents have to select the one that best complements the given pattern. The measurement consists of two sets of six items. The testing time is 6 min in total (Lang et al., 2014).

Test Instructions

The control group received the regular test instructions for the reading comprehension test as a baseline measure. Instructions were read aloud to the students by a trained supervisor as described in the test manual (Lenhard & Schneider, 2006).

Intervention 1. Students participating in Intervention 1 received intensified test instructions aiming to ensure the comprehension of instructional clues. A friendly dragon served as a pedagogical agent to implement the optimized test instructions in a child-oriented manner. The dragon pictured in the test booklet explained the task requirements and repeated important aspects of the instructions. Essential components of the intensified test instructions were explicit hints on how the tasks had to be solved and how to respond to the questions. Specifically, they highlighted the following aspects: (a) that the correct answer can be given only after having read the corresponding text in which the answer is to be found, (b) that only one answer is correct, and (c) that only one answer should be given for each item. Additional task examples were provided to the students to introduce the item format step by step following the principles of Direct Instruction. The example items were presented in three chunks. First, the supervisor explained and displayed how the tasks were to be solved. Second, a comprehension task was jointly solved in the group. The supervisor provided scaffolding including immediate feedback about the correctness of the answer. The right answer and the strategy applied were explained. Third, in a last block, students had to work on three items independently. Afterwards, the solutions and strategies for answering the tasks were discussed once more in the group. As in Direct Instruction, guidance by the supervisor

was gradually faded out, and the children engaged increasingly in independent work on the tasks (Baumann, 1988).

Intervention 2. This group received the same test instructions as those given to the control group. However, in addition, the assessment session started with an introductory phase containing a story combined with physical activities. Children listened to a story about a family visit to the zoo and were asked to perform appropriate movements (representing an animal) when key words were mentioned.

Methodological Approach

To shed light on the effects of the interventions compared to performance in the control group, three test-taking indicators (*test processing, test compliance, test performance*) served as dependent variables. We investigated the effects of the interventions on these test-taking indicators with predictive modeling based on a stepwise inclusion of student characteristics (gender, age), centered control measures relevant to test-taking (reading speed, basic cognitive skills), and dummy variable coding for experimental group membership. Because the test-taking indicators were not normally distributed and a regular regression model would overestimate the effects, we calculated negative binomial models. These have been shown to be suitable for count data that do not meet the assumption of a normal distribution (Hilbe, 2011). Model tests indicated a better model fit compared to regular regression models. Analyses were conducted in R (Version 3.2.2) using the package MASS (Ripley et al., 2013).

Results

Overall, students participating in the study had a mean score of 3.26 (*SD* = 3.39) on the reading comprehension test. The internal consistency of the reading comprehension test was satisfactory ($\alpha = .85$), although the items appeared to be quite difficult in general for the students with SEN-L in Grades 3 and 4. Probabilities of correct answers based on all valid

responses ranged from .18 to .74 per item (M = .42, SD = 0.16). Table 2 gives an overview of the three test-taking indicators.

Table 2

Means of the Indicators of Test-Taking Behavior

	Test performance (*N* of items answered correctly)	Test processing (*N* of items processed)	Test compliance (*N* of responses noncompliant with instructions)
Control group	2.90 (3.61)	5.69 (5.27)	1.76 (3.87)
Intervention 1	3.44 (2.91)	7.37 (5.23)	0.92 (2.82)
Intervention 2	3.47 (3.56)	6.09 (4.97)	1.77 (3.92)
Kruskal–Wallis test	$H(2) = 6.33$, $p < .05$	$H(2) = 8.18$, $p < .05$	$H(2) = 5.78$, $p < .10$

Note. Standard deviations in parentheses.

Test Performance

Table 3 displays the results of the models predicting the outcome variable of test performance. Whereas gender did not contribute to the prediction, age resulted in a significant coefficient indicating that older students performed better in the reading comprehension test. This effect declined when the other control variables were included in the model. The second model, which additionally included reading speed and basic cognitive skills (nonverbal reasoning, perceptual speed), explained a major part of the variance (R^2 = .63), and introducing the dummy variable coding for Intervention 1 and 2 explained another 4% of variance (Model 3; $\Delta R^2 = 0.04$, $p < .05$). Thus, participating in either intervention improved students' test performance significantly. Reading speed was the strongest predictor of the outcome variable ($B = 0.07$, $\beta = 0.20$, $p < .001$). But even after controlling for other highly influential variables, participation in Intervention 1 ($B = 0.39$, $\beta = 0.05$, $p < .001$) and Intervention 2 ($B = 0.23$, $\beta = 0.03$, $p < .05$) proved to be a significant predictor of performance in the reading comprehension test. However, there was no

significant difference between the test performance of students in Interventions 1 and 2 when compared within the model (model with reference Intervention 2: $B = 0.17$, $\beta = 0.05$, $p = .11$).

Table 3

Predicting Test Performance

	Model 1			Model 2			Model 3		
	B	*B SD*	β	*B*	*B SD*	β	*B*	*B SD*	β
Intercept	1.25	0.06		0.95	0.05		0.73	0.08	
Gender	0.05	0.06	0.013	0.01	0.05	0.002	0.02	0.05	0.01
Age	0.35*	0.07	0.08	0.08	0.06	0.02	0.07	0.06	0.02
Reading speed				0.07*	0.01	0.19	0.07*	0.01	0.20
Perceptual speed				-0.01	0.02	0.05	-0.01	0.01	-0.02
Reasoning				0.05*	0.01	-0.03	0.05*	0.01	0.05
Intervention 1[+]							0.39*	0.11	0.05
Intervention 2[+]							0.23*	0.11	0.03
AIC	1599.9			1421.6			1412.1		
R^2	.10			.63			.67		

Note. [+] Reference is the control group. * $p < .05$.

Test Processing

Again, students' age resulted in a significant coefficient in Model 1, but its effect was reduced after controlling for reading speed and basic cognitive skills (Model 2; see Table 4). When the dummy variable coding for the interventions was entered in Model 3 ($\Delta R^2 = 0.05$, $p < .001$), reading speed ($B = 0.05$, $\beta = 0.08$, $p < .001$) and reasoning skills ($B = 0.03$, $\beta = 0.02$, $p < .05$) remained strong predictors of the dependent variable. But participation in Intervention 1 also had a significant effect on the indicator *test processing* compared to both the control group ($B = 0.41$, $\beta = 0.04$, $p < .001$) and Intervention 2 (model with reference Intervention 2: $B = 0.27$, $\beta = 0.05$, $p < .01$). Taking part in Intervention 2 and

listening to a story combined with physical activities did not lead to a significant increase in the number of processed items.

Table 4

Predicting Test Processing

	Model 1			Model 2			Model 3		
	B	*B SD*	β	*B*	*B SD*	β	*B*	*B SD*	β
Intercept	0.51	0.61		1.75	0.05		1.56	0.07	
Gender	-0.06	0.05	-0.01	-0.03	0.05	0.01	0.04	0.04	0.01
Age	0.15*	0.06	0.02	-0.01	0.06	-0.001	-0.02	0.06	-0.003
Reading speed				0.04*	0.01	0.08	0.05*	0.01	0.08
Perceptual speed				-0.002	0.01	-0.003	-0.001	0.01	-0.001
Reasoning				0.03*	0.01	0.02	0.03*	0.01	0.02
Intervention 1^{+}							0.41*	0.10	0.04
Intervention 2^{+}							0.13	0.10	0.20
AIC	2013.5			1903.2			1891.0		
R^2	.03			.34			.39		

Note. $^{+}$ Reference is the control group. * $p < .05$.

Test Compliance

Variables contributing to differences in the occurrence of responses that did not comply with the test instructions were: age, reasoning skill, and participation in Intervention 1 (see Table 5). Reading speed did not predict the outcome variable *test compliance*. Older students were more capable of generating responses complying with the test instructions than younger students (age: $B = 0.57$, $\beta = 0.13$, $p < .01$). Even after controlling for reading speed and basic cognitive skills, this predictor remained significant (age: $B = 0.51$, $\beta = 0.11$, $p < .05$). Nonetheless, participation in an intervention explained additional variance (Model 3: $\Delta R^2 = 0.03$, $p < .05$). Students participating in the intensified test instructions produced fewer responses that did not comply with the test instructions

compared to controls (B = -0.76, β = 0.10, $p < .05$) and compared to Intervention 2 (model with reference Intervention 2: B = -0.99, β = -0.26, $p < .01$).

Table 5

Predicting Test Compliance

	Model 1			Model 2			Model 3		
	B	*B SD*	β	*B*	*B SD*	β	*B*	*B SD*	β
Intercept	0.35	0.16		0.13	0.16		0.24	0.25	
Gender	-0.08	0.16	-0.02	0.04	0.16	0.01	-0.01	0.15	-0.002
Age	-0.57*	0.19	-0.13	-0.51*	0.20	-0.11	-0.55*	0.20	-0.12
Reading speed				-0.02	0.02	-0.04	-0.02	0.02	-0.05
Perceptual speed				0.02	0.02	0.05	0.01	0.02	0.02
Reasoning				-0.22*	0.05	-0.18	-0.22*	0.05	-0.18
Intervention 1[+]							-0.76*	0.37	-0.10
Intervention 2[+]							0.23	0.35	0.03
AIC	921.58			885.06			882.52		
R^2	.05			.14			.17		

Note. [+] Reference is the control group. * p < .05.

Discussion

Fair and valid measurement is not feasible if children fail to understand the test instructions. This is especially relevant when assessing academic competencies via standardized tests in students with special educational needs in the area of learning. It has been suggested that one suitable way to obtain reliable and valid test results is to give these students additional information on how items should be processed and how correct responses should be generated (cf. Wong et al., 1982). The present experimental study evaluated the effects of two brief interventions involving more in-depth instructions on the test-taking steps in a well-established standardized German reading comprehension test compared to a control group receiving no intervention. It investigated several aspects of test-taking behavior. Overall, participation in the interventions proved to be relevant for the test-taking behavior of

students with SEN-L. Even after controlling for highly relevant student characteristics such as reading speed and basic cognitive skills, participation in the instructional interventions significantly predicted the test-taking indicators under study (i.e., test performance); and with respect to Intervention 1, also test processing and test compliance.

As expected, students in Interventions 1 and 2 answered more items correctly than students in the control group who received the standard test instructions. However, only participation in Intervention 1 also predicted the number of items processed and reduced the number of responses that did not comply with the test instructions. Thus, participation in Intervention 1 had a significant effect on all three indicators: test performance, test processing, and test compliance. After they had received the intensified test instructions, students were able to give more correct answers than controls. In addition, on average, they processed more items than students in the control group and in Intervention 2, and they responded comparatively less often in ways that did not comply with test instructions. In conclusion, Intervention 1 seems to most benefit the test-taking performance of students with SEN-L. These adapted and supplemented test instructions combining issues of test-taking and response behavior with motivational factors can also be viewed as the familiarization phase applied in previous research that has led to positive results (cf. Hessels, 2009). The present results indicate that repeating adequate response behavior and presenting additional example items in the intensified test instructions of Intervention 1 were appropriate techniques for reducing the occurrence of responses that were not compliant with the test instructions. In light of the problems students with SEN-L have in comprehending test instructions reported in the introduction, it seems that Intervention 1 was highly effective in enhancing their test-taking behavior .

Intervention 2 also had a positive effect on the test performance of the students. However, it did not have a significant effect on test processing or test compliance. The aim of

enhancing the attention of the students through physical activities combined with a children's story was expedient in the sense of leading to better test performance. At the same time, students participating in Intervention 2 did not show more compliant response behavior than controls. Nonetheless, it is worth noting that students with SEN-L could benefit from such a rather simple technique and still improve their test performance.

The direct comparison between Interventions 1 and 2 seems to indicate that the former intervention provided the students with the relevant information in a way that better enabled them to comprehend the test instructions than students receiving the standard test instructions in the control group and those receiving Intervention 2. Nonetheless, the test performance of these latter students with SEN-L receiving Intervention 2 was also improved compared to controls, indicating that accessibility to the tasks may have been facilitated (Kettler, Braden, & Beddow, 2011; Wong, 1991).

As Lauth and Wiedl (1985) have pointed out, translating competence into performance can be altered by the way test instructions are presented. Interestingly, both interventions in this study resulted in a comparable improvement in test performance; and this, in the end, is the relevant measure for assessing students' competencies. This is the case although Intervention 2 did not affect the other indicators of test-taking behavior significantly. This could be interpreted as showing that enhanced attention may lead to more in-depth or accurate processing of the items and thus to fewer mistakes. But the mechanisms by which performance is improved in the two intervention conditions might well differ: Whereas Intervention 1 enhances task comprehension, Intervention 2 might enhance item processing.

Furthermore, the two test-taking indicators *test performance* and *test processing* were influenced strongly by reading speed and basic nonverbal reasoning skills. As expected, these variables, which have been shown to correlate significantly with reading comprehension in

students in regular schools (cf. Artelt et al., 2001), also predicted test performance and test processing in students with SEN-L. Moreover, control variables also contributed to the outcome variable of test compliance. Although reading speed did not explain interindividual differences in test compliance, age and basic reasoning skills did. Students who were older and students with higher scores in the reasoning test demonstrated more compliant response behavior in the reading comprehension test. This converges with results reported by Scruggs (1984) showing that test-wiseness develops in younger children throughout elementary school.

One of the challenges when interpreting the results in this study is the low probability of item solution in the reading comprehension test for the group of students with SEN-L. The items were very difficult for them to solve in the given test time; thus effects of the intervention may have been underestimated. With the present data, we cannot disentangle whether and in what way the additional explanations and physical exercises with the supervisor may have influenced the test situation. For example, could differences also be due to motivational aspects and the perceived atmosphere of testing?

Due to the heterogeneity of competence profiles in this group of students, certain students with specific characteristics might have profited more from one of the two interventions. Future research should examine potential differential effects of the interventions. Additionally, further variables should be taken into account. Assessing test motivation (Grolnick & Ryan, 1990) and test anxiety (Bryan, Sonnefeld, & Grabowski, 1983) could also deliver relevant predictors for a differential investigation and explanation of processes and differences in test taking that resulted from both interventions.

Nonetheless, in general, these results indicate that the implemented interventions provided a better test preparation than the regular, unadapted test instructions presented to controls. Although the effects and differences were quite small, this study presents an

approach that brings us one step closer to fair and valid competence measurements for students with SEN-L. Although Intervention 1 was designed for a specific reading comprehension test, the approach can be adapted easily to other standardized competence tests. This intervention may be useful when including students with SEN-L in large-scale assessments, thereby helping to learn more about this special group of students and their competence development in comparison to other students (cf. Heydrich et al., 2013).

References

Artelt, C., Schiefele, U., & Schneider, W. (2001). Predictors of reading literacy. *European Journal of Psychology of Education*, *16*(3), 363–383.

Autorengruppe Bildungsbericht (2016). *Bildung in Deutschland 2016: ein indikatorengestützter Bericht mit einer Analyse zu Bildung und Migration* [Education in Germany 2016: An indicator-based report including an analysis of education and migration]. Bielefeld, Germany: Bertelsmann Verlag.

Banks, T., & Eaton, I. (2014). Improving test-taking performance of secondary at-risk youth and students with disabilities. *Preventing School Failure: Alternative Education for Children and Youth*, *58*(4), 207–213.

Baumann, J. F. (1988). Direct instruction reconsidered. *Journal of Reading*, *31*(8), 712–718.

Bendel, O., Schnöring, K., & Back, A. (2002). *Potenziale pädagogischer Agenten* [Potential of pedagogical agents]. Arbeitsberichte des Learning Center der Universität St. Gallen, St. Gallen, Switzerland.

Blossfeld, H.-P., Roßbach, H.-G., & von Maurice, J. (Eds.). (2011). Education as a lifelong process: The German National Educational Panel Study (NEPS). *Zeitschrift für Erziehungswissenschaft*, *14*(2).

Brown, A. L., & Campione, J. C. (1986). Psychological theory and the study of learning disabilities. *American Psychologist*, *41*(10), 1059–1068.

Bryan, J. H., Sonnefeld, L. J., & Grabowski, B. (1983). The relationship between fear of failure and learning disabilities. *Learning Disability Quarterly*, *6*(2), 217–222.

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research*, *65*(3), 245–281.

Engelmann, S. (1980). *Direct instruction*. Englewood Cliffs, NJ: Educational Technology Publications.

Grolnick, W. S., & Ryan, R. M. (1990). Self-perceptions, motivation, and adjustment in children with learning disabilities: A multiple group comparison study. *Journal of Learning Disabilities*, *23*(3), 177–184.

Grünke, M., & Grosche, M. (2014). Lernbehinderung [Learning disability]. In G. W. Lauth, M. Grünke, & J. C. Brunstein (Eds.), *Interventionen bei Lernstörungen* (pp. 76–89). Göttingen, Germany: Hogrefe.

Hessels, M. G. P. (2009). Estimation of the predictive validity of the HART by means of a dynamic test of geography. *Journal of Cognitive Education and Psychology*, *8*(1), 5–21.

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.

Hilbe, J. M. (2011). *Negative binomial regression*. Cambridge, England: Cambridge University Press.

Kettler, R. J., Braden, J. P., & Beddow, P. A. (2011). Test-taking skills and their impact on accessibility for all students. In S. N. Elliott, R. J. Kettler, P. A. Beddow, & A. Kurz (Eds.), *Handbook of Accessible Achievement Tests for All Students* (pp. 147–159). New York, NY: Springer.

Kettler, R. J., Elliott, S. N., & Beddow, P. A. (2009). Modifying achievement test items: A theory-guided and data-based approach for better measurement of what students with disabilities know. *Peabody Journal of Education*, *84*(4), 529–551.

KMK (Ed.). (1999). *Empfehlungen zum Förderschwerpunkt Lernen* [Recommendation on special educational needs on the area of learning]. Berlin. Retrieved from http://www.kmk.org/fileadmin/Dateien/pdf/PresseUndAktuelles/2000/sopale.pdf

KMK (Ed.). (2016). *Sonderpädagogische Förderung in Schulen 2005 bis 2014* [Special education in schools 2005–2014]. Berlin. Retrieved from http://www.kmk.org/fileadmin/Dateien/pdf/Statistik/Dokumentationen/Dokumentation_SoPaeFoe_2012.pdf

Krowatschek, D., Krowatschek, G., & Schmidt, C. (2011). *Marburger Konzentrationstraining (MKT) für Schulkinder* [The Marburg concentration training for school students]. Dortmund, Germany: Borgmann.

Kubesch, S., & Walk, L. (2009). Körperliches und kognitives Training exekutiver Funktionen in Kindergarten und Schule [Physical and cognitive training of executive functions in preschool and school settings]. *Sportwissenschaft*, *39*(4), 309–317.

Lang, F. R., Kamin, S., Rohr, M., Stünkel, C., & Willinger, B. (2014). *Erfassung der fluiden kognitiven Leistungsfähigkeit über die Lebensspanne im Rahmen des Nationalen Bildungspanels* [Assessment of the fluid cognitive ability across the life span in the National Educational Panel Study] (NEPS Working Paper No. 43). Bamberg, Germany: Leibniz-Institut für Bildungsverläufe, Nationales Bildungspanel.

Lauth, G. W., & Wiedl, K. H. (1985). Zur Veränderbarkeit der Testleistung im CFT-20 durch Instruktionsintensivierung [Changes in test performance on the CFT 20 as a result of instruction intensification]. *Diagnostica*, *31*(2), 200–209.

Lenhard, W., & Artelt, C. (2009). Komponenten des Leseverständnisses [Components of reading comprehension]. In W. Lenhard & W. Schneider (Eds.), *Diagnostik und*

Förderung des Leseverständnisses. Tests und Trends (Vol. 17, pp. 1–17). Göttingen, Germany: Hogrefe.

Lenhard, W., & Schneider, W. (2006). *ELFE 1-6: Ein Leseverständnistest für Erst- bis Sechstklässler* [Reading comprehension test for 1st- to 6th-graders]. Göttingen, Germany: Hogrefe.

Lloyd, J. W., Keller, C., & Hung, L. (2007). International understanding of learning disabilities. *Learning Disabilities Research and Practice*, *22*(3), 159.

Mayringer, H., & Wimmer, H. (2003). *Salzburger Lese-Screening für die Klassenstufen 1-4 (SLS 1-4). Manual* [Salzburg Screening Test for Reading in Elementary School Students]. Bern, Switzerland: Hans Huber.

Oerter, R. (1980). *Psychologie des Denkens* [Psychology of thinking]. Donauwörth, Germany: Auer.

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.

Ripley, B., Venables, B., Bates, D. M., Hornik, K., Gebhardt, A., Firth, D., & Ripley, M. B. (2013). Package "MASS." *CRAN Repository*. Retrieved from http://cran. r-project. org/web/packages/MASS/MASS. pdf

Sax, G. (1997). *Principles of educational and psychological measurement and evaluation*. Belmont, CA: Wadsworth.

Scruggs, T. E. (1984). *The administration and interpretation of standardized achievement tests with learning disabled and behaviorally disordered elementary school children. Final report*. Salt Lake City, UT: Developmental Center for the Handicapped, University of Utah. Retrieved from http://files.eric.ed.gov/fulltext/ED311652.pdf

Scruggs, T. E., Bennion, K., & Lifson, S. (1985a). An analysis of children's strategy use on reading achievement tests. *The Elementary School Journal*, *85*(4), 479–484.

Scruggs, T. E., Bennion, K., & Lifson, S. (1985b). Learning disabled students' spontaneous use of test-taking skills on reading achievement tests. *Learning Disability Quarterly*, *8*(3), 205–210.

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), 1–26.

Swanson, H. L. (2001). Searching for the best model for instructing students with learning disabilities. *Focus on Exceptional Children*, *34*(2), 1–15.

Vukovich, A. (1971). Die Konstruktion psychologischer Tests [The construction of psychological tests]. In R. Heiss (Ed.), *Handbuch der Psychologie* (pp. 113–144). Göttingen, Germany: Hogrefe.

Weinert, S., Artelt, C., Prenzel, M., Senkbeil, M., Ehmke, T., & Carstensen, C. H. (2011). Development of competencies across the life span. *Zeitschrift für Erziehungswissenschaft*, *14*(2), 67–86.

Werning, R. (2010). Inklusion zwischen Innovation und Überforderung [Inclusion between demands and innovation]. *Zeitschrift Für Heilpädagogik*, 8, 284–291.

Wong, B. Y. L. (1991). The relevance of metacognition to learning disabilities. In B. Y. L. Wong (Ed.), *Learning about learning disabilities* (pp. 231–258). San Diego, CA: Academic Press.

Wong, B. Y. L. (1994). Instructional parameters promoting transfer of learned strategies in students with learning disabilities. *Learning Disability Quarterly*, *17*(2), 110–120.

Wong, B. Y. L., Wong, R., & LeMare, L. (1982). The effects of knowledge of criterion task on comprehension and recall in normally achieving and learning disabled children. *Journal of Educational Research*, *76*, 119–126.

Zielinski, W. (1996). Lernschwierigkeiten [Learning problems]. In F. E. Weinert (Ed.), *Psychologie des Lernens und der Instruktion* (Vol. 2, pp. 369–402). Göttingen, Germany: Hogrefe.

Beitrag 2

Nusser, L., Heydrich, J., Carstensen, C. H., Artelt, C., & Weinert, S. (2016). Validity of survey data of students with special educational needs – Results from the National Educational Panel Study. In H.-P- Blossfeld, J. von Maurice, M. Bayer & J. Skopek (Hrsg.), *Methodological Issues of Longitudinal Surveys: The Example of the National Educational Panel Study* (S. 251-266). Wiesbaden: Springer.

Validity of Survey Data of Students with Special Educational Needs—Results From the National Educational Panel Study

Lena Nusser, Jana Heydrich, Claus H. Carstensen, Cordula Artelt and Sabine Weinert

Abstract

Within the German National Educational Panel Study (NEPS), $N = 578$ students in Grade 5 and $N = 1{,}186$ students in Grade 9 with special educational needs in the area of learning (SEN-L) took part in feasibility studies examining how to include students with special needs in large-scale assessments like the NEPS (Heydrich, Weinert, Nusser, Artelt, & Carstensen, 2013). Among other things, written questionnaires were administered to the participating students. Alongside gaining insight into students' perspectives on educationally relevant questions, the information given by the students is also important in case of nonparticipating parents and thus of missing information on family backgrounds from the parents. Former research could show that secondary-school students without SEN living at home with their parents are reliable proxy reporters for their parents' socioeconomic status and familial background. However, there is no database showing that this conclusion can be generalized to students with SEN-L. Thus, we asked whether the administered student questionnaire validly assessed the social background of these students. In addition to a thorough descriptive analysis of missing data as an indicator of the response behavior of students with SEN-L, the validity of students' answers was also tested by matching the parents' data with the students' responses in order to identify accuracy using a chance-corrected agreement coefficient. Students with SEN-L responded validly and accurately to certain questions, while other items resulted in low completion rates and reduced validity of the students' reports.

© Springer Fachmedien Wiesbaden 2016
Hans-Peter Blossfeld et al. (eds.), *Methodological Issues of Longitudinal Surveys,*
DOI 10.1007/978-3-658-11994-2_15

1 Introduction

When exploring individual educational pathways, as is done in educational panel studies, it is essential to gain a detailed view of the target person and their respective educational contexts. This requires a variety of reliably and validly assessed context and background information about and from the target persons. Within the National Educational Panel Study (NEPS; Blossfeld, Roßbach, & von Maurice, 2011), the collection of context information is accomplished through written questionnaires as well as personal and telephone interviews (Frahm et al., 2011). Next to surveying the target person, further context persons, such as parents, teachers, and principals, are asked to participate in the survey to gain a broad spectrum of relevant information following a multi-informant perspective for several items. In addition to reporting about the target persons' own experiences, appraisals, and further personal information, the participants are asked for statements about third persons. Information about the socioeconomic status and ethnical origin of parents and family, in particular, are retrieved via these proxy-reports, as is also done in several other studies (e. g., PISA 2000; OECD, 2002). These variables are important in case parents do not participate in the study and thus do not provide the relevant information.

Former research indicates that linguistic skills and levels of cognitive development may affect the validity of self-reports. Other pivotal factors for valid data are mental representations of the requested topics as well as the relevance of the question content for the respondent (Fuchs, 2009; Looker, 1989). These aspects raise the question of whether children whose cognitive and linguistic abilities are still developing are able to provide valid and reliable information. However, while it can be assumed that adolescents are generally able to answer a questionnaire, it is to be expected that data quality for children under 14 years of age is comparatively lower (Fuchs, 2009). Until now, whether these findings hold true for the group of students with special educational needs in the area of learning (SEN-L) has remained uncertain. Focusing on this special target population, there is no data to maintain such a conclusion. When considering the validity of responses of students with SEN-L, additional factors, such as reduced attentional resources and delayed cognitive and language development, have to be considered (Schröder, 2000).

This chapter sheds some light on the validity of the survey data that was collected within the NEPS from students with SEN-L.

2 Current State of Research

Surveys are an essential part of research for many scientific disciplines. About 90 % of collected data derive from surveys (Bortz & Döring, 2006). The number of adolescents and children surveyed has been increasing over the past decades. This trend may be accounted for by two facts: On the one hand, research interest has shifted

more and more to the children themselves as autonomous human beings, to their environments, and to their living conditions. On the other hand, children are often used as proxy reporters, for example, to provide details on the socioeconomic status of their parents (Kränzl-Nagl & Wilk, 2000; Scott, 1997).

2.1 Theoretical Context

Written questionnaires, like other forms of surveys, pose certain demands for the respondent and depend on his or her ability and willingness to reply (Scholl, 2003). The respondent has to pass through a cognitive question-answer process that represents a complex interaction between the respondent and the survey instrument (Fuchs, 2004). Tourangeau's (2000) cognitive model of response behavior assumes four stages of answering questions: comprehension of the question, retrieval of the relevant information, judgment regarding the completeness of the information, and editing a response. Based on Tourangeau's model, Krosnick (2000) established a theory identifying two response behaviors. In contrast to an optimal answering process as described by the four steps above, Krosnick specifies an alternative response behavior called *satisficing,* meaning that not all cognitive steps are conducted, and instead, the first acceptable response alternative is chosen. The likelihood of the occurrence of satisficing is related to three factors. Specifically, difficult tasks or items tend to lead to satisficing for respondents with comparatively lower cognitive abilities and less motivation.

Furthermore, each phase of the cognitive-response behavior can be afflicted with stage-specific errors on behalf of the respondent, such as limited comprehension of the question or lacking mental representations that may lead to invalid answers and reduced data quality. There is also evidence that item characteristics, such as the phrasing of questions and items (Benson & Hocevar, 1985; de Leeuw, Borgers, & Smits, 2004), the number of response categories (Borgers & Hox, 2001), the position and order within the questionnaire (Fuchs, 2004), and the salience for the respondent (Looker, 1989; Lipski, 2000) may impact the reliability of data.

2.2 Surveying Children and Adolescents

In general, studies on the validity of students' responses in surveys judge their answers to be predominantly useful. More specifically, adolescents at secondary school who live at home with their parents have been shown to give reliable proxy reports of their parents' socioeconomic status and familial background (Looker, 1989).

Maaz, Kreuter, and Watermann (2006) analyzed the validity of responses collected from 15-year-old adolescents in Germany. The congruency between the students' and their parents' responses to questions regarding parental education and achieved certificates was examined by the agreement coefficient Cohen's kappa (Cohen, 1960).

The results showed a high amount of conformity between the two reports as well as recognizable differences depending on the type of school attended (for mothers' school-leaving qualification: K = .50–.80; for fathers' school-leaving qualification: K = .42–.67). Assuming that attending a certain type of school correlates with the cognitive performance of the students, the results indicate that better cognitive abilities may lead to more valid reports on parental education and that lower cognitive abilities may lead to more difficulties in providing correct answers. These findings suggest that even adolescents are not necessarily able to give correct responses regarding their parents' education. However, West, Sweeting, and Speed (2001) showed that 11-year-old children were able to report correctly on their parents' occupation. They found high or very high levels of agreement (K = .69 for fathers' occupation; K = .82 for mothers' occupation) in addition to low nonresponse rates. Nevertheless, these findings have to be put in context since oral interviews in one-on-one settings were used for the assessments.

Obviously, linguistic demands and complexity of items in a written questionnaire may lead to a challenging answering process. However, compared with item characteristics, child characteristics and abilities seem to play an even more prominent role (Bell, 2007). Borgers, de Leeuw, and Hox (2000), for example, showed that individual differences in reading comprehension of children from the age of 7 to 8 significantly impact on response rates and the consistency of responses. Other results indicate that limited reading competence sometimes influences the response validity of negatively phrased items (Marsh, 1986), thus showing that item and person characteristics interact.

Although research projects in Germany have collected information from students with SEN-L via written questionnaires (Lehmann & Hoffmann, 2009; Wocken, 2005), experiences in surveying this group of students (especially in large-scale assessments) is still rather limited in Germany.

2.3 Students with Special Educational Needs in the Area of Learning

Comprising 40 % of all students with SEN, those with SEN-L constitute by far the largest group of students with SEN in Germany (Autorengruppe Bildungsberichterstattung, 2014). It is a highly heterogeneous group with very heterogeneous competence profiles (Antor & Bleidick, 2001). Students with SEN-L have severe and extensive deficits in the accomplishment of cognitive performance requirements lasting over a period of time. Constraints are primarily found in the acquisition of cognitive-verbal and abstract content (Grünke, 2004). These children's ability to cope with learning requirements can be characterized by using and applying fewer strategies for gathering and processing relevant information (Grünke, 2004; Klauer & Lauth, 1997). Working memory and attention span are expected to be comparatively restricted, which may result in difficulties following instructions (Schmetz, 1999). Children

who are assigned to special schools for students with SEN-L usually have difficulties in reading and writing, which impact on various learning areas (Valtin & Sasse, 2012).

3 Research Question

To gain valid and comparable data within large-scale assessments, standardized administrations of tests and questionnaires are implemented. Considering the characteristics of students with SEN-L, it is worthwhile to ask how they cope with constraints and conditions of standardized surveys. Kränzl-Nagl and Wilk (2000) emphasize the challenges of standardized surveys for children whose cognitive development might be called delayed. Aiming to investigate response validity in written questionnaires for students with SEN-L, the following questions were addressed:

1) Does response validity differ between students with SEN-L at special schools and students without SEN at regular schools?
2) Are there specific differences in the content of the items that are more or less valid for students with SEN-L?
3) Does the validity of the responses depend on students' age?
4) Are there changes in sustained attention across a given questionnaire that might influence response validity?

With respect to partially limited cognitive abilities, it can be expected that students with SEN-L might provide less valid data than general-education students without SEN-L (Borgers & Hox, 2001; Fuchs, 2009). According to previous research, older students—both general-education students and students attending special schools—are anticipated to be more likely to provide valid information compared with their younger peers. To investigate whether students with SEN-L are attentive and able to provide answers throughout a written questionnaire, we observed their performance throughout the advancing questionnaire, expecting a decline in completion rates.

4 Method

4.1 Sample

This study uses data from the NEPS Starting Cohort 3 and NEPS Starting Cohort 4.[1] Within the two cohorts, a series of feasibility studies was conducted including stu-

1 This paper uses data from the National Educational Panel Study (NEPS): Starting Cohort 3–5th Grade, doi:10.5157/NEPS:SC3:1.0.0. From 2008 to 2013, NEPS data were collected as part of the Framework Programme for the Promotion of Empirical Educational Research funded by the Ger-

dents at special schools: Students with SEN-L were oversampled in Grades 5 and 9. N = 587 Grade-5 students with SEN-L were on average M_{age} = 11.44 (SD_{age} = 0.65) years old, and 44.1 % were female. The sample of students with SEN-L in Grade 9 comprised N = 1,186 students with a mean age of M_{age} = 15.55 (SD_{age} = 0.64) years, and 44.4 % were female. As a reference group, data from students attending regular schools were used (Grade 5: N = 5,208, M_{age} = 10.95 [SD_{age} = 0.52]; Grade 9: N = 14,540, M_{age} = 15.19 [SD_{age} = 0.64]).

In addition to the students filling out an extensive student questionnaire, the parents of participating students were asked to take part in the study. The parent interview was implemented as a computer-assisted telephone interview (Frahm et al., 2011). About 51.4 % of parents of Grade-5 students at special schools and 47.0 % of parents of students with SEN-L in Grade 9 participated in the study. The participation rate for parents of general-education students was higher (69.3 % for Grade 5; 54.3 % for Grade 9). Due to varying participation rates, some analyses were restricted to a fraction of the sample.

4.2 Design

With respect to students with SEN-L, the survey follows an experimental design. Specific accommodations for students with SEN-L were implemented to possibly increase and test for aspects of validity. The questionnaire's design is adjusted in terms of (a) length, (b) selected contents, (c) sequence of administrated items, and (d) mode of presentation.

(a) Considering a limited attention span of students with SEN-L, the written questionnaire was reduced in length. Since students with SEN-L attending Grade 5 were surveyed on two days, the written questionnaire was split into two parts. Overall, the amount of items was reduced by 23 % compared with the instrument for regular schools. All students attending Grade 9 were also surveyed on two days: one in fall 2010, and one in spring 2011. Overall, the questionnaires for the sample of students with SEN-L at special schools were shortened by 53 % in comparison with the regular survey instrument (Skopek, Pink, & Bela, 2012a, 2012b). Each shortened survey instrument was designed to require about 20 minutes.

(b) The questionnaires were arranged to cover a broad spectrum of subjects. For instance, general information about the familial background, socioeconomic status, ethnic origin, and language use was surveyed (Kristen et al., 2011; Stocké, Blossfeld, Hoenig, & Sixt, 2011). In addition, the selected content addressed reading engagement as well as nonformal/informal learning environment, school achievement, and

man Federal Ministry of Education and Research (BMBF). As of 2014, the NEPS survey is carried out by the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg in cooperation with a nationwide network.

Table 1 Experimental Design

Experimental group: Forward	BI +	m1 + m2 + m3 (+ m4)
Experimental group: Backward	BI +	(m4 +) m3 + m2 + m1

computer usage (Frahm et al., 2011). The selection and compilation of the items was guided by thematic salience for the students as well as by the linguistic and cognitive requirements of the questions.

(c) Moreover, we anticipated that—in the course of the procedure—the attentiveness of students with SEN-L would decrease substantially so that the validity of individual responses might be affected. To identify and test for potential effects of item position, the design allotted a rotation of content-bound modules (m1–m4) in two experimental versions. It is important to note that the module on *basic information* (BI), including questions concerning the ethnical and social origin, was not touched by this variation. The questionnaire was administered in group settings. All testing groups were randomly assigned to one of the experimental conditions *forward* or *backward,* that is, to the original or a reversed sequence of modules (see Table 1).

(d) With respect to the expected partially limited reading fluency and comprehension, it is questionable whether students with SEN-L were able to answer the provided questionnaire in a straightforward manner without any assistance. To circumvent the effects of reading restrictions, the National Center for Educational Outcomes (NCEO) recommended the use of 'read aloud' as an essential adaptation when evaluating students with SEN-L (Koretz & Barton, 2003). Therefore, the questions and items were presented orally, that is, they were read aloud by the interviewer using a predefined script. The effect of reading aloud on the validity of responses is not addressed in this chapter (see Gresch, Strietholt, Kanders, & Solga in this volume for an analyses and comparison of these data with regular-school students attending *Hauptschule*).

4.3 Measures and Procedures

Several methods were employed to investigate the validity of the data reported by students with SEN-L and to approach the questions raised above. For a direct assessment of the validity of the students' data, the parents' data—which were not always available—were matched with the students' responses in order to identify congruencies and accuracy. Therefore, a coefficient based on the percentage of factual conformity of two reports is calculated. The chance-corrected agreement coefficient Cohen's kappa is a standardized measure of agreement that accounts for the expected proportion of agreements by chance (Wirtz & Caspar, 2002). In general, a kappa value > .75 is suggested to indicate very high agreement, while a kappa between .6 and .75 indi-

cates good agreement (Fleiss & Cohen, 1973). Kappa values between .4 and .6 are regarded as acceptable depending on the specific research subject under study (Wirtz & Caspar, 2002). It is important to note that identical reports of students and their parents do not necessarily imply valid and meaningful data. However, the consistency of independently collected information can be seen as an indication of the plausibility of both the students' as well as the parents' reports.

As further indicators for validity, Bell (2007) suggests inspecting inconsistencies of individual response patterns. Particular focus lies on rates of missing values, such as invalid responses and nonresponses, to detect specific content-related refusals or difficulties. By comparing the observed patterns of nonresponse within and across the two experimental conditions, we analyze positional effects related to decreasing attention. This also provides hints as to whether response behavior is more likely to be related to the specific questions and topics or to the item position within the questionnaire.

5 Results and Analysis

In this section, first results regarding the direct measurement of validity are reported, followed by a description of missing values addressing the question of sustained attention throughout the questionnaire.

The direct measurement of validity operationalized via the agreement coefficient Cohen's kappa is only possible for a subset of all administrated items—namely those requesting facts such as ethnic origin and native language. Looking at the kappa values for these items, a distinctive pattern can be observed (see Table 2).

The agreement coefficients vary between the samples of students at regular schools and students at special schools, as well as between the two age-groups. Altogether, the coefficients follow comparable patterns. Items asking about the country of birth of the students themselves, as well as that of their parents, reach high and very high conformity, respectively. However, values of kappa decline for items regarding the third generation. Not only does the congruency between students' and parents' reports decline, but the completion rate of the items also decreases. The rates of missing values rise from less than 3 % up to almost 40 % for these particular items (see Table 3). However, this increasing rate of item nonresponse corresponds to the administrated order of the items, and it seems rather connected to the content of the questions. The response rates for items regarding the native language of both the child and the parents are higher. The agreement-coefficients over $K = .9$ for the samples in both cohorts indicate high validity for these items. Since these questions permit multiple responses for people growing up multilingually, the chances for congruency are higher and also account for high Kappa values.

Notably, the majority of the agreement coefficients are higher for the sample of general-education students in comparison with students with SEN-L. Exceptions are

Table 2 Agreement Coefficient for Ethnic Origin and Native Language

Variables	Grade 5: Special schools	Grade 5: Regular schools	Grade 9: Special schools	Grade 9: Regular schools
Country of birth	.710	.852	.836	.872
Country of birth: Mother	.668	.871	.889	.884
Country of birth: Father	.695	.853	.717	.879
Country of birth: Maternal grandmother	.620	.581	.388	.547
Country of birth: Maternal grandfather	.458	.506	.264	.511
Country of birth: Paternal grandmother	.230	.557	.075	.408
Country of birth: Paternal grandfather	.376	.555	.17	.421
Native language	.945	.951	.939	.954
Native language: Mother	.949	.969	.963	.963
Native language: Father	.911	.957	.906	.958

Table 3 Proportion of Item Nonresponse for Ethnic Origin and Native Language

Variables	Grade 5: Special schools	Grade 5: Regular schools	Grade 9: Special schools	Grade 9: Regular schools
Country of birth	2.3 %	1.2 %	1.4 %	0.7 %
Country of birth: Mother	7.5 %	3.2 %	7.0 %	2.1 %
Country of birth: Father	14.2 %	5.9 %	11.5 %	4.2 %
Country of birth: Maternal grandmother	21.5 %	11.3 %	17.7 %	5.8 %
Country of birth: Maternal grandfather	29.9 %	16.0 %	22.4 %	8.6 %
Country of birth: Paternal grandmother	32.8 %	14.6 %	26.8 %	9.8 %
Country of birth: Paternal grandfather	39.1 %	18.9 %	28.6 %	11.7 %
Native language	2.1 %	2.7 %	2.8 %	0.9 %
Native language: Mother	6.4 %	3.1 %	4.4 %	1.9 %
Native language: Father	10.3 %	4.2 %	9.0 %	3.8 %

Table 4 Agreement Coefficient for Parental Education and Occupation in Grade 9

Variables	Special schools	Regular schools
Highest education qualification: Mother	.288	.526
Highest education qualification: Father	.348	.466
Employment: Mother	.476	.537
Employment: Father	.604	.483
Vocational position: Mother	.262	.435
Vocational position: Father	.370	.565
Occupation: Mother	.504	.586
Occupation: Father	.501	.511

items with a general high congruency, such as native language and the country of birth of the child and parents. Comparing the two cohorts, the coefficients are nearly identical for the sample of general-education students, while for the sample of students with SEN-L, age seems to have an effect. Students attending Grade 9 at special schools reply less validly to various items regarding ethnic origin compared with students attending Grade 5 at special schools or students attending general-education schools.

Students in Grade 9 were also asked about their parents' educational qualifications as well as their employment status and occupation (see Table 4). Overall, these items show lower but partially acceptable congruency according to Wirtz and Caspar (2002). These items seem to cause more difficulties for students to respond validly. For students with SEN-L, particularly low agreement coefficients are found for the questions of the highest education qualification and the vocational position of both parents. With one exception (item: current employment of father), the students at regular schools achieve higher congruency with their parents' reports. About one fourth of the fathers of students attending special schools are reported to be without employment. However, less than 10 % of students attending general educational schools report that their fathers are unemployed.

The challenges that questions regarding the educational careers of parents may produce can also be detected by looking at the item nonresponse. For the item of the father's highest educational qualification, rates of missing values are up to 70 % for students with SEN-L (see Table 5). Low completion rates reduce the sample considerably so that results are only meaningful for a subsample of the students at special schools.

Regarding the length and volume of the questionnaires, positional effects of the items were observed. As can be seen in Table 6 and Table 7, the amount of item-non-

Table 5 Proportion of Item Nonresponse for Parental Education and Occupation

Variables	Special schools	Regular schools
Highest education qualification: Mother	63.1 %	24.0 %
Highest education qualification: Father	70.8 %	30.0 %
Employment: Mother	10.7 %	6.7 %
Employment: Father	18.5 %	9.1 %
Vocational position: Mother	45.3 %	26.5 %
Vocational position: Father	41.7 %	27.0 %
Occupation: Mother	56.2 %	34.8 %
Occupation: Father	56.7 %	37.6 %

response varies between the two experimental conditions *forward* and *backward*. The anticipated gradual increase of item-nonresponse in the progress of the survey instrument is not observed. The response patterns indicate that missing values are not directly associated with the position of items within the questionnaire. In fact, the occurrence of reduced completion rates does not seem to depend on the position of the item within the questionnaire, but rather on item content. Since less completion rates occur in both groups for the identical items, subject-specific causes can be assumed.

Thus, subjects dealing with reading engagement and nonformal/informal learning environment tend to lead to item-nonresponse for fifth graders. Modules 2 and 3, which are concerned with familial learning environment and school achievement as well as the quality of instruction, show the lowest rates of missing values in both experimental groups.

Students attending Grade 9 at special schools show lower completion rates in comparison with students at general-education schools. However, there are few dif-

Table 6 Proportion of Item Nonresponse for Modules 1–4 in Grade 5

	Forward	Backward	Number of items
Module 1: Reading engagement	8.2 %	10.3 %	17
Module 2: Familial learning environment, School achievement	5.3 %	7.3 %	19
Module 3: Quality of instruction	5.0 %	7.7 %	15
Module 4: Nonformal/informal learning environment	10.4 %	11.3 %	9

Table 7 Proportion of Item Nonresponse for Modules 1–3 in Grade 9

	Forward	Backward	Number of items
Module 1: Reading engagement	10.0 %	10.8 %	11
Module 2: School achievement, Nonformal/informal learning environment	10.6 %	11.9 %	15
Module 3: Computer usage	7.1 %	10.1 %	22

ferences between modules and topics, regardless of item position. Both groups show the lowest rates of missing values for the subject of computer usage.

6 Discussion

In this chapter, the question of whether the use of a survey questionnaire for students with SEN-L in Grade 5 and Grade 9 is valid was addressed. The results on congruency between the students' and parents' reports as well as the completion rates have revealed some challenges regarding surveying students with SEN-L. Analyses have shown that the response data can be considered valid in respect to particular questions. For some subjects, students with SEN-L are capable of responding validly and accurately. Other items lead to difficulties that can result in low completion rates and reduced validity of the students' reports. The absence of mental representations, for example, those regarding the place of birth of grandparents (especially for students who have a background of migration) may yield problems. However, questions concerning native language—a salient feature of daily communication within the family—lead to higher agreement between the students' and parents' reports. In contrast, agreement coefficients for the socioeconomic status of parents illustrate the constraints of using students with SEN-L as proxy reporters.

However, the matching child's and parent's data does not necessarily indicate validity, and the accuracy of parents' reports is not automatically evident. Nevertheless, congruency can be taken as important evidence regarding the value of the students' data. Additionally, parents' information is not available for all students, and it is therefore only possible to gain information for a fraction of the sample. It is rather important to collect proxy reports from the children themselves, particularly for students whose parents did not take part in the survey. The question of validity is especially relevant for this subsample.

In general, students with SEN-L produce higher rates of missing values, which restricts the calculation of agreement coefficients to a subsample of students. Hence, the interpretation of the results is limited. Reduced completion rates also circumvent a valid and comprehensive description of the entire sample. Reporting on the ethnic

origin of students at special schools and their familial background based on the students' questionnaire alone is not feasible. Systemic variations regarding selective nonresponses may have effects on further analyses (Kreuter, Maaz, & Watermann, 2004).

Considering the length of the administrated questionnaires, they seem to be suitable for students with SEN-L since the occurrence of item nonresponse has proven to be primarily linked to the content of the questions. The mode of reading aloud may support a continuous response behavior. However, even this administration mode does not lead to complete item response for these questions, which may create difficulties because of content and wording. The mechanism behind the reduced completion rate for specific items needs to be addressed in further analyses.

The approach of post-testing as described in this chapter can only take some aspects into account. The response behavior of editing an answer or not is an important and obvious indicator. However, it is not possible to examine cognitive processes that may lead to certain response behaviors. To understand more about the challenging issues regarding the validity of collected data, further aspects need to be considered. The effects of the respondents' cognitive abilities and the interaction with item characteristics as stated in various studies need to be examined in more detail.

Although the validity of the collected data from students with SEN-L seems evident for various items, caution should still be taken when working with the survey data.

References

Antor, G., & Bleidick, U. (Eds.). (2001). *Handlexikon der Behindertenpädagogik: Schlüsselbegriffe aus Theorie und Praxis.* Stuttgart: Kohlhammer.

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

Bell, A. (2007). Designing and testing questionnaires for children. *Journal of Research in Nursing, 12*(5), 461–469.

Benson, J., & Hocevar, D. (1985). The impact of item phrasing on the validity of attitude scales for elementary school children. *Journal of Educational Measurement, 22*(3), 231–240.

Blossfeld, H.-P., Roßbach, H.-G., & von Maurice, J. (Eds.) (2011). Education as a lifelong process: The German National Educational Panel Study (NEPS) [Special issue]. *Zeitschrift für Erziehungswissenschaft, 14.* Wiesbaden: VS Verlag für Sozialwissenschaften.

Borgers, N., de Leeuw, E., & Hox, J. (2000). Children as respondents in survey research: Cognitive developmental and response quality. *Bulletin de Méthodologie Sociologique, 66*(1), 60–75.

Borgers, N., & Hox, J. (2001). Item nonresponse in questionnaire research with children. *Journal of Official Statistics, 17*(2), 321–335.

Bortz, J., & Döring, N. (2006). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler.* Heidelberg: Springer Medizin Verlag.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46.

De Leeuw, E., Borgers, N., & Smits, A. (2004). Pretesting questionnaires for children and adolescents. In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, & E. Singer (Eds.), *Methods for testing and evaluating survey questionnaires* (pp. 409–429). New York: John Wiley & Sons.

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement, 33*(3), 613–619.

Frahm, S., Goy, M., Kowalski, K., Sixt, M., Strietholt, R., Blatt, I., … Kanders, M. (2011). Transition and development from lower secondary to upper secondary school. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft, 14. Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 217–232). Wiesbaden: VS Verlag für Sozialwissenschaften.

Fuchs, M. (2004). Kinder und Jugendliche als Befragte. *ZUMA-Nachrichten, 28*(54), 60–88.

Fuchs, M. (2009). The reliability of children's survey responses. The impact of cognitive functioning on respondent behavior. In Statistics Canada (Ed.), *Symposium 2008: Data collection: Challenges, achievements and new directions* (pp. 1–8). Ottawa: StatCan.

Grünke, M. (2004). Lernbehinderung. In G. W. Lauth, M. Grünke, & J. Brunstein (Eds.), *Interventionen bei Lernstörungen* (pp. 65–77). Göttingen: Hogrefe.

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.

Klauer, K. J., & Lauth, G. W. (1997). Lernbehinderungen und Leistungsschwierigkeiten bei Schülern. In F. E. Weinert (Ed.), *Psychologie des Unterrichts und der Schule* (Enzyklopädie der Psychologie, Serie Pädagogische Psychologie, Bd. 3, pp. 701–738). Göttingen: Hogrefe.

Koretz, D. M., & Barton, K. E. (2003). *Assessing students with disabilities: Issues and evidence* (CSE Technical Report No. 587). Los Angeles, CA: University of California, Center for the Study of Evaluation.

Kränzl-Nagl, R., & Wilk, L. (2000). Möglichkeiten und Grenzen standardisierter Befragungen unter besonderer Berücksichtigung der Faktoren soziale und personale Wünschbarkeit. In F. Heinzel (Ed.), *Methoden der Kindheitsforschung* (pp. 59–76). München: Weinheim.

Kreuter, F., Maaz, K., & Watermann, R. (2004). Der Zusammenhang zwischen Qualität von Schülerangaben zur sozialen Herkunft und den Schulleistungen. In K.-S. Rehberg

(Ed.), *Soziale Ungleichheit—Kulturelle Unterschiede, Verhandlungen des 32. Kongresses der Deutschen Gesellschaft für Soziologie in München 2004* (pp. 3465–3478). Frankfurt: Campus.

Kristen, C., Edele, A., Kalter, F., Kogan, I., Schulz, B., Stanat, P., & Will, G. (2011). The education of migrants and their children across the life course. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft, 14. Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 121–137). Wiesbaden: VS Verlag für Sozialwissenschaften.

Krosnick, J. (2000). The threat of satisficing in surveys: The shortcuts respondents take in answering questions. *Survey Methods Newsletter, 20*(1), 4–8.

Lehmann, R., & Hoffmann, E. (2009). Anlage und Durchführung der Untersuchung. In R. Lehmann, & E. Hoffmann (Eds.), *BELLA. Berliner Erhebung arbeitsrelevanter Basiskompetenzen von Schülerinnen und Schülern mit Förderbedarf "Lernen"* (pp. 17–29). Münster: Waxmann.

Lipski, J. (2000). Zur Verläßlichkeit der Angaben von Kindern bei standardisierten Befragungen. In F. Heinzel (Ed.), *Methoden der Kindheitsforschung. Ein Überblick über Forschungszugänge zur kindlichen Perspektive* (pp. 77–86). München: Weinheim.

Looker, E. D. (1989). Accuracy of proxy reports of parental status characteristics. *Sociology of Education, 62*(4), 257–276.

Maaz, K., Kreuter, F., & Watermann, R. (2006). Schüler als Informanten? Die Qualität von Schülerangaben zum sozialen Hintergrund. In J. Baumert, P. Stanat, & R. Watermann (Eds.), *Herkunftsbedingte Disparitäten im Bildungswesen: Differenzielle Bildungsprozesse und Probleme der Verteilungsgerechtigkeit. Vertiefende Analysen im Rahmen von PISA 2000* (pp. 31–59). Weinheim: VS Verlag für Sozialwissenschaften.

Marsh, H. W. (1986). Negative item bias in rating scales for preadolescent children: A cognitive developmental phenomenon. *Developmental Psychology, 22*(1), 37–49.

OECD (2002). *PISA 2000 technical report.* Retrieved from http://www.pisa.oecd.org/dataoecd/53/19/33688233.pdf

Schmetz, D. (1999). Förderschwerpunkt Lernen. *Zeitschrift für Heilpädagogik, 4,* 134–143.

Scholl, A. (2003). *Die Befragung.* Konstanz: UVK-Verlag.

Schröder, U. (2000). *Lernbehindertenpädagogik: Grundlagen und Perspektiven sonderpädagogischer Lernhilfe.* Stuttgart: Kohlhammer.

Scott, J. (1997). Children as respondents: Methods for improving data quality. In L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, & D. Trewin (Eds.), *Survey measurement and process quality* (pp. 331–350). New York: Wiley.

Skopek, J., Pink, S., & Bela, D. (2012a). *Data manual. Starting Cohort 3—From lower to upper secondary school. NEPS SC 3 1.0.0* (NEPS Research Data Paper). Bamberg: University of Bamberg, National Educational Panel Study.

Skopek, J., Pink, S., & Bela, D. (2012b). *Starting Cohort 4: 9th grade (SC4). SUF-Version 1.0.0. Data Manual* (NEPS Research Data Paper). Bamberg: University of Bamberg, National Educational Panel Study.

Stocké, V., Blossfeld, H.-P., Hoenig, K., & Sixt, M. (2011). Social inequality and educational decisions in the life course. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft, 14. Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 103–119). Wiesbaden: VS Verlag für Sozialwissenschaften.

Tourangeau, R. (2000). *The psychology of survey response.* Cambridge: University Press.

Valtin, R., & Sasse, A. (2012). Schriftspracherwerb. In W. Schrader, & F. B. Wember (Eds.), *Didaktik des Unterrichts im Förderschwerpunkt Lernen* (pp. 179–190). Stuttgart: Kohlhammer.

West, P., Sweeting, H., & Speed, E. (2001). We really do know what you do: A comparison of reports from 11 year olds and their parents in respect of parental economic activity and occupation. *Sociology, 35*(2), 539–559.

Wirtz, M., & Caspar, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität.* Göttingen: Hogrefe.

Wocken, H. (2005). *Andere Länder, andere Schüler? Vergleichende Untersuchungen von Förderschülern in den Bundesländern Brandenburg, Hamburg und Niedersachen* (Forschungsbericht). Retrieved from http://www.mbjs.brandenburg.de/sixcms/media.php/5527/wocken_ergebnis-heft.pdf

About the authors

C. Artelt
Department of Educational Research,
University of Bamberg, Bamberg.

C. H. Carstensen
Psychology and Methods of Educational Research,
University of Bamberg, Bamberg.
Leibniz Institute for Educational Trajectories (LIfBi), Bamberg.

J. Heydrich
University of Bamberg, Bamberg.

L. Nusser
Department of Psychology I: Developmental Psychology,
University of Bamberg, Bamberg
e-mail: lena.nusser@uni-bamberg.de

S. Weinert
Department of Psychology I: Developmental Psychology,
University of Bamberg, Bamberg

Beitrag 3

Nusser, L., Carstensen, C. H. & Artelt, C. (2015). Befragung von Schülerinnen und Schülern mit sonderpädagogischem Förderbedarf Lernen: Ergebnisse zur Messinvarianz. *Empirische Sonderpädagogik, 7* (2), 99-116.

Empirische Sonderpädagogik, 2015, Nr. 2, S. 99-116
ISSN 1869-4845 (Print) · ISSN 1869-4934 (Internet)

Befragung von Schülerinnen und Schülern mit sonderpädagogischem Förderbedarf Lernen: Ergebnisse zur Messinvarianz

Lena Nusser[1], Claus H. Carstensen[1,2] & Cordula Artelt[2]

[1] *Leibniz-Institut für Bildungsverläufe e.V. (LIfBi)*
[2] *Otto-Friedrich-Universität Bamberg*

Zusammenfassung

Dieser Artikel betrachtet die messinvariante Erfassung bildungsrelevanter Konstrukte mit Hilfe schriftlicher Befragungen bei Schülerinnen und Schülern an Förderschulen und Hauptschulen in der 5. Jahrgangsstufe. Um optimale Administrationsbedingungen für Schülerinnen und Schüler mit sonderpädagogischem Förderbedarf Lernen zu erforschen, wurde ein experimentelles Design implementiert. Inwieweit angepasste Erhebungsinstrumente und unterschiedliche Beschulungsformen sowie anzunehmende Kompetenzunterschiede der Befragten eine messäquivalente Erfassung der Lesemotivation und des akademischen Selbstkonzeptes ermöglichen, wird durch Mehrgruppenvergleiche konfirmatorischer Faktorenanalysen untersucht. Die Ergebnisse deuten darauf hin, dass vergleichende Analysen zwischen Schülergruppen an Förderschulen und Hauptschulen für bestimmte Konstrukte und Faktoren sinnvoll interpretierbar sind.

Schlüsselwörter: Messinvarianz, Mehrgruppenvergleich, sonderpädagogischer Förderbedarf, Lesemotivation, akademisches Selbstkonzept

Questionnaires for Students with Special Educational Needs in the Area of Learning: Results from Multi-Group Analysis

Abstract

This article focuses on measurement invariance of the assessment of educationally relevant constructs via written questionnaires for students at special schools and at low track schools attending 5th grade. To examine optimal conditions of administration for students with special educational needs in the area of learning an experimental design was implemented. If accommodated questionnaires, different school enrollments as well as competence differences allow equivalent assessment of reading motivation and academic self-concepts will be investigated with multi-group comparison of confirmatory factor analysis. The results indicate that comparisons between groups of students at special schools and low track schools are meaningful for certain constructs.

Keywords: measurement invariance, multi group comparison, special educational needs, reading motivation, academic self-concept

Persönlichkeitseigenschaften, sozioökonomische oder ethnische sowie weitere personenbezogene Merkmale werden in der sozialwissenschaftlichen und psychologischen Forschung häufig mit Hilfe von Fragebögen erfasst (Scholl, 2003). Der Einsatz standardisierter Instrumente ist in diesem Kontext die häufigste Form, um den Anforderungen einer objektiven Erfassung gerecht zu werden. Ziel solcher Befragungen sind nicht zuletzt Gruppenvergleiche und differenzierte Analysen, um Disparitäten in Ausprägung und auch Veränderungen dieser Merkmale zu untersuchen. Voraussetzung hierfür ist die messäquivalente oder auch messinvariante Erfassung der zugrundeliegenden Konstrukte. Andernfalls würden verschiedene Ausprägungen eines Konstruktes, z.B. in Abhängigkeit von Geschlecht, nicht ein und dasselbe Konstrukt in beiden Gruppen in äquivalenter Art und Weise abbilden. Mittelwertsvergleiche wären in einem solchen Fall unzulässig, da sie zu verzerrten Ergebnissen führen würden. Gerade aber das hohe Maß an Standardisierung zur objektiven Erfassung von Konstrukten kann bei bestimmten Personengruppen einer validen und messinvarianten Erfassung entgegenstehen (Lipski, 2000). Dies betrifft vor allem Kinder mit Entwicklungsverzögerungen (Kränzl-Nagl & Wilk, 2000), wie sie auch bei Schülerinnen und Schülern mit einer Lernbehinderung zu finden sind.

Zudem gibt es empirische Hinweise, dass das Leistungsniveau und das Leseverständnis der Befragten Effekte auf die Vergleichbarkeit der Angaben haben können, so dass Daten von Schülerinnen und Schülern unterschiedlicher Schultypen nicht zwingend äquivalent sein müssen (Byrne, Shavelson & Muthén, 1989; Steinmetz, Schmidt, Tina-Booh, Wieczorek & Schwartz, 2009). Möchte man jedoch vergleichende Analysen verschiedener Schülergruppen durchführen, kommt der Messinvarianz eine zentrale Bedeutung zu.

Dies ist insbesondere nicht trivial, wenn Daten von Schülerinnen und Schülern an Förderschulen mit anderen Schülergruppen verglichen werden. Eine schriftliche Befragung von Förderschülerinnen und -schülern stellt eine besondere Herausforderung dar, da sie heterogene Kompetenzprofile und häufig Einschränkungen in der Lesefertigkeit, im sprachlichen Verständnis und in der Aufmerksamkeitsspanne aufweisen (Grünke, 2004). Es stellt sich deshalb die Frage, wie Fragebögen für Schülerinnen und Schüler mit kognitiven Beeinträchtigungen bzw. einem sonderpädagogischen Förderbedarf Lernen (SPF-L) gestaltet werden können, um zum Beispiel bildungsrelevante Faktoren messinvariant zu anderen Schülergruppen ohne Beeinträchtigung zu erheben und somit vergleichende Analysen sinnvoll interpretierbar sind.

Hierfür wurde für Schülerinnen und Schüler an Förderschulen mit dem Schwerpunkt Lernen ein experimentelles Design zur Gestaltung von schriftlichen Befragungen entwickelt und ein angepasster Fragebogen konzipiert, der die Besonderheiten dieser Gruppe berücksichtigen soll. Ob bildungsrelevante Konstrukte trotz dieser Anpassungen bei Schülerinnen und Schülern an Förderschulen im Vergleich zu jenen an Hauptschulen messäquivalent erhoben werden können, wird im Folgenden untersucht.

Schülerinnen und Schüler an Förderschulen Lernen

Schülerinnen und Schüler, bei denen ein sonderpädagogischer Förderbedarf diagnostiziert wird, werden in Deutschland nicht nur integrativ, sondern überwiegend im stark ausdifferenzierten Förderschulsystem in kleineren Lerngruppen unterrichtet. Die Zuweisung zu einer sonderpädagogischen Fördereinrichtung erfolgt basierend auf der Empfehlung der Ständigen Kultusministerkonferenz zu einer der acht spezifizierten Förderschwerpunkte, wie beispielsweise Lernen, Sprache, Sehen oder Hören (Ständige Konferenz der Kultusminister der Länder, 1994). Die größte Gruppe der Schülerinnen und Schüler stellt mit anteilig fast 40% jene mit einem Förderbedarf Lernen dar (Bil-

dungsbericht, 2014). Da insbesondere für den deutschsprachigen Raum eine einheitlich anzuwendende Definition für dieses Phänomen noch aussteht, existieren bisher keine konsistenten Standards zur Diagnostik des sonderpädagogischen Förderbedarfs Lernen (Bildungsbericht, 2014; Bos, Müller & Stubbe, 2010). Aus diesem Grund ist an Förderschulen mit dem Schwerpunkt Lernen eine äußerst heterogene Schülerschaft mit einer stark ausgeprägten Leistungsspanne zu finden, die sich hinsichtlich ihrer Kompetenzprofile stark unterscheiden können (Bos et al., 2010; Gebhardt, Oelkrug & Tretter, 2013).

In Annäherung an das Erscheinungsbild wird das Vorliegen eines SPF-L als überdauernde und weitreichende Einschränkung bei der Bewältigung schulischer Anforderungen beschrieben (Klauer & Lauth, 1997), die sich im Besonderen bei dem „Erwerb kognitiv-verbaler und abstrakter Inhalte (z. B. Lesen, Rechnen)“ (Grünke, 2004, S. 65) zeigt. Als weitere Charakteristika werden diesen Schülerinnen und Schülern schwächere sprachliche Fähigkeiten, geringe Lernumfänge als auch reduzierte Aufmerksamkeit zugeschrieben.

Bisher wurde diese Gruppe von Schülerinnen und Schülern selten bzw. lediglich mit geringen Stichprobenumfängen in large-scale-assessments integriert. Diese Tatsache bedingt sich auch dadurch, dass keine Standards existieren, wie diese Schülergruppe sinnvoll und vergleichbar in standardisierte Befragungen und Kompetenztestungen einbezogen werden kann (Hoermann, 2007). Um den Zugang zum Aufgabenmaterial für bestimmte Schülergruppen mit Beeinträchtigungen zu erleichtern, sowie konstrukt-irrelevante und behinderungsbezogene Einschränkungen zu verringern, werden häufig verschiedene Akkommodationen im Rahmen von Kompetenztests implementiert (Cormier, Altman, Shyyan & Thurlow, 2010). Hierzu gehören zum Beispiel die Darbietung der Aufgaben in Brailleschrift für Kinder und Jugendliche mit visuellen Einschränkungen oder aber eine Reduktion der Aufgaben, mehr Testzeit oder auch Read-Aloud-Akkommodationen für Schülergruppen mit kognitiven Einschränkungen (Koretz & Barton, 2004). Insgesamt sind Forschungsergebnisse zu Akkommodationen und die Vergleichbarkeit von Testergebnissen von Schülerinnen und Schülern mit und ohne Behinderung sehr heterogen, so dass diskutiert wird, ob bestimmte Anpassungen die Instrumente möglicherweise so stark verändern, dass angepasste und nicht angepasste Aufgaben nicht mehr vergleichbar und messinvariant sind (Cormier et al., 2010).

Messinvarianz

Messinvarianz oder auch Messäquivalenz fragt danach, „whether or not, under different conditions of observing and studying phenomena, measurement operations yield measures of the same attribute“ (Horn & McArdle, 1992, S. 117). Die Messinvarianzforschung findet vielseitig Beachtung in verschiedenen Disziplinen und Forschungskontexten.

Um zu prüfen, ob in verschiedenen Substichproben die zu untersuchenden Konstrukte messäquivalent erhoben werden und somit die strukturellen Beziehungen zwischen den Indikatoren und latenten Variablen invariant und vergleichbar sind, bietet sich die Mehrgruppenanalyse konfirmatorischer Faktorenmodelle als vielseitige und robuste Methode an (Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000). Dieses Vorgehen ermöglicht, dass die verschiedenen Modellparameter schrittweise hinsichtlich ihrer Messinvarianz untersucht werden (Chen, 2008; Sass, 2011).

Hierfür werden für die zu betrachtenden Gruppen die Faktorenmodelle simultan geschätzt und schrittweise Restriktionen in Form von gleichgesetzten Parametern über die Gruppen hinweg eingeführt. Starke faktorielle Invarianz herrscht dann, wenn die (a) konfigurale, (b) metrische und (c) skalare Invarianz gegeben ist (Byrne et al., 1989; Sass, 2011).

Zur Prüfung dieser Formen der Invarianz wird zunächst ein konfirmatorisches Modell, welches identische Messbeziehungen zwischen Indikatoren und Faktoren annimmt, für beide interessierenden Gruppen separat geschätzt. Diese Spezifikation dient in den nachfolgenden Analysen zur Messinvarianz mit simultanen Schätzungen als Baseline-Modell. In einem Modell zur (a) konfiguralen Invarianz wird die gleiche Anzahl an freien und fixierten Parametern in beiden Gruppen geschätzt. Die (b) metrische Invarianz nimmt gleiche Faktorladungen für die beobachteten Gruppen an; somit kann von einer identischen Beziehung zwischen den Indikatoren und ihren latenten Faktoren in allen Gruppen ausgegangen werden. Um die Annahme der (c) skalaren Invarianz zu erfüllen, werden zusätzlich die Intercepts der Indikatoren restringiert und über die Gruppen hinweg gleichgesetzt. Personen mit gleichen latenten Faktorwerten geben demnach ähnliche Antworten für die jeweiligen Items an (Sass, 2011). Auf dieser Stufe der Messäquivalenz sind Gruppenvergleiche der Faktormittelwerte sinnvoll interpretierbar (Brown, 2006; Dimitrov, 2010).

In der Praxis hat sich herausgestellt, dass die Hypothese der Messinvarianz für alle Parameter über verschiedene Gruppen hinweg sehr streng und selten realistisch ist. Aus diesem Grund haben Byrne et al. (1989) das Konzept der partiellen Invarianz eingeführt, das davon ausgeht, dass nicht alle Parameter invariant sind. Führen zusätzliche Modellrestriktionen zu einer signifikanten Verschlechterung des Modells, können auf Grund theoretischer Überlegungen und auch explorativer Vorgehensweise einzelne Indikatoren identifiziert werden, deren Parameter in den Gruppen frei geschätzt werden (Vandenberg & Lance, 2000). Um trotz eingeschränkter Messinvarianz sinnvolle Gruppenvergleiche für die gemessenen Konstrukte durchführen zu können, sollte jeder Faktor mit mindestens zwei Indikatoren, die metrische und skalare Invarianz aufweisen, repräsentiert sein (Byrne et al., 1989; Steenkamp & Baumgartner, 1998).

Fragestellung

Bezogen auf die dargelegten Herausforderungen bei der Befragung von Schülerinnen und Schülern mit SPF-L soll im Folgenden die messinvariante Erfassung der Konstrukte Lesemotivation und akademisches Selbstkonzept für die Gruppen der Schülerinnen und Schüler an Förderschulen und Hauptschulen untersucht werden. Grundsätzlich stellt sich die Frage, ob die Konstrukte in diesen beiden Gruppen die gleiche faktorielle Struktur aufweisen und somit die gleichen Indikatoren auf den entsprechenden latenten Faktoren laden. Ist diese Voraussetzung erfüllt, kann die Messinvarianz und somit die Vergleichbarkeit der Faktorladungen und der Intercepts für die beiden Schülergruppen untersucht werden.

Es ist anzunehmen, dass Schülerinnen und Schüler an Hauptschulen eher als die Gruppen an Förderschulen Lernen mit abstraktem sprachlichem Material arbeiten und gegebenenfalls vertrauter mit Test- und Befragungssituationen sind. Daher sollten sie vergleichsweise besser - sowohl dispositionell als auch situativ - in der Lage sein, sich mit dem Befragungsinstrument und den Items auseinanderzusetzen. Ihr Antwortverhalten würde mehr valide und konsistente Antworten liefern, welches die Faktorladungen und Intercepts beeinflusst.

Auf Grund reduzierter Aufmerksamkeitsspannen bei Schülerinnen und Schülern mit SPF-L könnten für diese Zielgruppe Positionseffekte erwartet werden, wenn Items nicht zu Beginn, sondern erst am Ende nach etwa 20- bis 25-minütiger Befragung präsentiert werden. Auf Grund gesunkener Konzentration und Bereitschaft zur Bearbeitung ist eine erhöhte Tendenz zu Response-Sets zu erwarten, die einer validen und messäquivalenten Erfassung entgegenstehen. Um diese Effekte zu untersuchen, erfolgte die Erfassung der Konstrukte bei dieser Schülergruppe mittels Fragebögen in zwei Varianten mit rotierter Itemreihenfolge. Deshalb wird auch hier die Messäquivalenz der Daten geprüft.

Methode

Stichprobe

Die zugrundeliegenden Daten stammen aus der Startkohorte 3 des Nationalen Bildungspanels[1] (NEPS; Blossfeld, Roßbach & von Maurice, 2011). Insgesamt nahmen in der 5. Jahrgangsstufe 57 Förderschulen mit dem Schwerpunkt Lernen aus dem ganzen Bundesgebiet teil; 587 Schülerinnen und Schüler beteiligten sich an der Erhebung, nachdem von ihren Erziehungsberechtigten eine Einverständniserklärung zur Teilnahme an der NEPS-Studie eingeholt worden war. Das durchschnittliche Alter der Schülergruppe betrug M = 11.44 Jahre (SD = 0.65), 44.1% der Teilnehmenden war weiblich. Als Referenz zur Stichprobe der Förderschülerinnen und -schüler dient die Schülergruppe an Hauptschulen. Insgesamt nahmen in dieser Schulform 745 Schülerinnen und Schüler aus 42 Institutionen teil. Das Durchschnittsalter lag bei M = 11.30 Jahre (SD = 0.67) und 45.5% der Schülergruppe war weiblich. Die Daten enthalten nicht für alle diese Schülerinnen und Schüler Informationen über einen vorhandenen sonderpädagogischen Förderbedarf. Auf Grund der Angaben des Statistischen Bundesamtes kann davon ausgegangen werden, dass nur eine geringe Anzahl von Personen mit SPF-L an Hauptschulen zu finden sind (Integrationsrate von 1.59% für das Schuljahr 2010/2011 [eigene Berechnung]; Statistisches Bundesamt, 2011).

Design

Die Erhebungen fanden an einem Schulvormittag bzw. an zwei Schulvormittagen im Klassenkontext statt. Die schriftliche Befragung der Schülerinnen und Schüler erfolgte jeweils am Ende einer Testsitzung. Schülerinnen und Schüler an Hauptschulen erhielten einen umfangreichen Fragebogen, der neben der sozialen und ethnischen Herkunft auch weitere bildungsrelevante Aspekte erfragt (Frahm et al., 2011). Insgesamt dauerte die papierbasierte Befragung 40 Minuten. Nach einer kurzen Instruktion wurde der Fragebogen von den Schülerinnen und Schülern selbstständig ausgefüllt. Die Erhebung an Förderschulen unterlag einem experimentellen Design, um optimale Bedingungen für Befragungen und die Administrationsbedingungen zu erforschen (Heydrich, Weinert, Nusser, Artelt & Carstensen, 2013). So wurde das Instrument für diese Gruppe hinsichtlich (a) der Länge, (b) des Inhalts, (c) der Reihenfolge und (d) des Darbietungsmodus angepasst, um den spezifischen Bedürfnissen dieser Schülergruppe gerecht zu werden. Im Mittel dauerte die Befragung 36 am ersten bzw. 30 Minuten am zweiten Erhebungstag.

a) Die Fragebogeninhalte wurden um etwa ein Drittel reduziert (von 185 auf 125 Items) und auf zwei Erhebungstage aufgeteilt. So konnten die inhaltlichen, zeitlichen und kognitiven Anforderungen an die Zielpersonen verringert werden, ohne auf zentrale Inhaltsbereiche verzichten zu müssen.
b) Die Auswahl der Inhalte orientierte sich u. a. an der Bedeutsamkeit der Fragen für die besondere Zielgruppe. Beispielsweise wurden Fragen zum idealistischen und realistischen Schulabschluss entfernt, da die vorgegebenen Antwortmöglichkeiten nicht die Realität von Förderschülerinnen und -schüler widerspiegeln.
c) Zur Identifizierung potentieller Positionseffekte durch abnehmende Aufmerksamkeit wurde das Instrument am zwei-

[1] Diese Arbeit nutzt Daten des Nationalen Bildungspanels (NEPS) Startkohorte 3 (Klasse 5), doi: 10.5157/NEPS:SC3:1.0.0. Die Daten des NEPS wurden von 2008 bis 2013 als Teil des Rahmenprogramms zur Förderung der empirischen Bildungsforschung erhoben, welches vom Bundesministerium für Bildung und Forschung (BMBF) finanziert wurde. Seit 2014 wird NEPS vom Leibniz-Institut für Bildungsverläufe e.V. (LIfBi) an der Otto-Friedrich-Universität Bamberg in Kooperation mit einem deutschlandweiten Netzwerk weitergeführt.

ten Erhebungstag in zwei Varianten mit rotierter Itemreihenfolge dargeboten. Die Zuweisung der Instrumentenversionen 1 (vorwärts) und 2 (rückwärts) erfolgte zufällig auf Klassenebene.

d) Um eine Strukturierungshilfe zu bieten und eine mögliche schwächere Lesefertigkeit als Hürde bei der validen Beantwortung zu mindern, wurden alle Items von geschulten Personen mit Hilfe eines standardisierten Skriptes vorgelesen.

Messung und Instrumente

Die Forschungsfragen zur Messinvarianz werden beispielhaft anhand verschiedener (Sub-)Skalen aus den Konstruktbereichen Lesemotivation und akademisches Selbstkonzept untersucht, deren theoretische Grundlage und Relevanz kurz vorgestellt werden.

Lesemotivation. Lesemotivation beschreibt die Absicht und das Bedürfnis einer Person, in einer Situation bestimmte Texte zu lesen (Schiefele, 1996). Die Lesemotivation wird als prädiktiv für die Lesekompetenz angesehen, da sich auch unter der Kontrolle kognitiver und sozialer Einflussvariablen signifikant positive Effekte zeigen (Artelt, Naumann & Schneider, 2010). Eine gering ausgeprägte Lesemotivation kann somit auch als eine relevante Ursache für eine niedrige Lesekompetenz verstanden werden. Zusammenhänge sind vor allem für den Faktor der intrinsischen Lesemotivation zu erwarten, der das Leseverhalten und die Häufigkeit von Leseaktivitäten und somit wiederum die Lesekompetenz positiv beeinflusst (McElvany, Kortenbruck & Becker, 2008; Möller & Schiefele, 2004).

Für die Erfassung der Lesemotivation wurden in die Instrumente des NEPS je drei Items für die Subskalen Leselust, Lesen aus Interesse und Selbstkonzept Lesen aus dem Fragebogen zur habituellen Lesemotivation (Möller & Bonerad, 2007) integriert (Frahm et al., 2011; s. Tabelle 1). Das administrierte Antwortformat weist eine vierstufige Likert-Skala auf (1 = *stimme gar nicht zu*; 2 = *stimme eher nicht zu*; 3 = *stimme eher zu*; 4 = *stimme völlig zu*). Die Subskalen Leselust und Lesen aus Interesse weisen eine gute interne Konsistenz auf (Förderschule: Crα = .839 und .760; Hauptschule: Crα = .900 und .846), während die Werte für die Subskala Selbstkonzept Lesen nicht mehr akzeptable sind (Crα =.524 bzw. .546).

Akademisches Selbstkonzept. Das Selbstkonzept wird als die Selbstwahrnehmung einer Person aufgefasst, die durch Erfahrungen unterschiedlichster Art und deren Interpretation geformt wird (Shavelson, Hubner & Stanton, 1976). Für den schulischen Bereich ist besonders das akademische Selbstkonzept relevant, welches nach fachspezifischen Facetten differenziert erfasst werden kann (Byrne, 1996; Marsh, 1990). Es wird angenommen, dass das akademische Selbstkonzept als Fähigkeitskognition einer Person Zusammenhänge mit Leistungen, hier vor allem der schulischen Performanz, aufweist (Möller & Köller, 2004). In der NEPS-Schülerbefragung wurden drei Subskalen zum akademischen Selbstkonzept in Anlehnung an die PISA-Erhebung 2000 implementiert (Kunter et al., 2002), wobei zwischen dem verbalen, dem mathematischen und dem schulischen Selbstkonzept differenziert wird. Jede der drei Skalen wurde jeweils mit drei Items erfragt (Wohlkinger, Ditton, von Maurice, Haugwitz & Blossfeld, 2011; s. Tabelle 1). Das administrierte Antwortformat ist eine vierstufige Likert-Skala (1 = *trifft gar nicht zu*; 2 = *trifft eher nicht zu*; 3 = *trifft eher zu*; 4 = *trifft völlig zu*). Wie von Marsh (1986) vorgeschlagen, wird ein negativ-formuliertes Item aus den Analysen ausgeschlossen, da dieses zur Vermeidung einseitiger Antwortmuster dient und auf Grund unzureichender Validität nicht in die Skalenbildung einbezogen wird. Die Werte der internen Konsistenz der Subskalen betragen für die Stichprobe der Schülerinnen und Schüler mit SPF-L Crα =.707, .813 und .806 und liegen damit in einem

Tabelle 1: Items zur Erfassung der Lesemotivation und des akademischen Selbstkonzeptes

Itemformulierung	Förderschule			Hauptschule		
	N	M	SD	N	M	SD
Lesemotivation						
a) Es macht mir Spaß, Bücher zu lesen. (LL)	480	2.91	1.19	614	2.77	1.13
b) Ich finde Lesen interessant. (LL)	453	2.93	1.16	603	2.73	1.11
c) Wenn ich genügend Zeit hätte, würde ich noch mehr lesen. (LL)	463	2.58	1.24	605	2.53	1.17
d) Ich lese gern etwas über neue Dinge. (LI)	467	2.91	1.17	605	2.75	1.14
e) Ich bin überzeugt, dass ich beim Lesen eine Menge lernen kann. (LI)	467	3.15	1.10	605	2.91	1.11
f) Lesen ist wichtig, um Dinge richtig zu verstehen. (LI)	469	3.35	0.98	606	3.09	1.06
g) Ich habe manchmal Schwierigkeiten, einen Text wirklich gut zu verstehen. (r) (SK)	457	2.51	1.17	598	2.68	1.09
h) Ich kann Texte sehr gut und schnell verstehen. (SK)	465	2.89	1.13	599	2.78	1.01
i) Ich muss vieles erst mehrmals lesen, bevor ich es richtig verstanden habe. (r) (SK)	456	2.35	1.21	592	2.58	1.12
Akademisches Selbstkonzept						
a) Im Fach Deutsch bin ich ein hoffnungsloser Fall. (r) (V)	465	2.97	1.14	652	2.86	0.99
b) Im Fach Deutsch lerne ich schnell. (V)	475	3.02	1.06	661	2.83	0.91
c) Im Fach Deutsch bekomme ich gute Noten. (V)	458	2.97	1.07	649	2.72	0.86
d) Im Fach Mathematik bekomme ich gute Noten. (M)	469	3.14	1.06	658	2.99	0.93
e) Mathematik ist eines meiner besten Fächer. (M)	468	3.05	1.16	659	2.73	1.15
f) Ich war schon immer gut in Mathematik. (M)	466	2.89	1.13	649	2.59	1.09
g) In den meisten Schulfächern lerne ich schnell. (S)	469	3.06	1.01	657	3.06	0.81
h) In den meisten Schulfächern schneide ich in Klassenarbeiten gut ab. (S)	469	2.97	1.04	652	2.94	0.83
i) Ich bin in den meisten Schulfächern gut. (S)	473	3.27	0.94	660	3.08	0.81

Anmerkungen. LL = Leselust. LI = Lesen aus Interesse. SK = Selbstkonzept Lesen. V = Verbales Selbstkonzept. M = Mathematisches Selbstkonzept. S = Schulisches Selbstkonzept. (r) = rekodiert.

akzeptablen bis guten Bereich. An Hauptschulen liegen die Werte bei Crα = .728, .861 und .820.

Implementation in die Erhebungsinstrumente. Das Regelschulinstrument enthält die Items zu den drei Skalen des akademischen Selbstkonzepts im mittleren Teil an den Positionen 84-92 der insgesamt 185 Items. Die Items zur Lesemotivation folgen an den Positionen 137-145. Alle Schülerinnen und Schüler an Förderschulen erhielten die Fragen zu diesen Themenblöcken am zweiten Erhebungstag, an dem insgesamt 63 Items administriert wurden. Für die Vorwärts- und Rückwärts-Varianten der Instrumente wurden die Fragenblöcke modulweise rotiert, damit die inhaltlichen Zusammenhänge und logischen Reihenfolgen bestimmter Themengebiete erhalten blieben. In Instrument 1 wurden die Items zur Lesemotivation zu Beginn an den Positionen 9-17 prä-

sentiert, die Fragen zum akademischen Selbstkonzept folgten an den Positionen 18-26. Im rotierten Instrument 2 standen die Items zum akademischen Selbstkonzept an Position 28-36, während die Fragen zur Lesemotivation am Ende an den Positionen 55-63 folgten.

Methodisches Vorgehen

Zunächst wird die theoretisch postulierte Faktorstruktur in einer konfirmatorischen Faktorenanalyse (KFA) spezifiziert und hinsichtlich etwaiger Fehlspezifikationen mit Hilfe der Fitindices geprüft. Nach der Bestätigung eines theoriegeleiteten Modells, dient dieses in den anschließenden Analysen des Mehrgruppenvergleiches als Baseline-Modell (Brown, 2006; Byrne & van de Vijver, 2010).

Das weitere Vorgehen folgt den im Theorieteil vorgestellten Analysenschritten zur Überprüfung der Messinvarianz. Bei diesem Bottom-Up-Ansatz werden zunächst alle Parameter frei geschätzt. Erst nach und nach werden Restriktionen in das Modell eingeführt, die Faktorladungen und Intercepts gleichsetzen. Da die Invarianz der Faktorladungen geprüft werden soll, werden diese Parameter zunächst alle frei geschätzt. Zu Identifikationszwecken des Modells wird die Varianz der latenten Faktoren auf 1 fixiert.

Die Beurteilung der Güte der Modelle erfolgt anhand der gängigen Fitindizes. Neben dem χ^2-Wert, der stark abhängig von der Stichprobengröße ist (Cheung & Rensvold, 2002; Dimitrov, 2010), werden weitere Fitmaße herangezogen, die verschieden sensitiv gegenüber unterschiedlichen Aspekten der Modellpassung sind (Brown, 2006). Diese sind der *Root Mean Square Error of Approximation* (RMSEA), der *Standardized Root Mean Square Residual* (SRMR), der *Comparative Fit Index* (CFI) sowie der *Tucker Lewis Index* (TLI). Etabliert haben sich in der Forschungspraxis die Anwendungsregeln, dass RMSEA $<.05$, SRMR $<.06$ sowie TLI und CFI $>.95$ auf eine gute Modellgüte hindeuten (Hu & Bentler, 1999).

Für die Durchführung der KFA und Mehrgruppenvergleiche gilt die Voraussetzung der Normalverteilung der Variablen (Dimitrov, 2010), die im vorliegenden Fall nicht gegeben ist. Aus diesem Grund wird die robustere Schätzmethode *Restricted Maximum Likelihood* (MLR) gewählt. Fehlende Werte werden als *missing at random* behandelt. Somit werden im Sinne des *Full Information Maximum Liklihood*-Ansatzes auch unvollständige Datenreihen einbezogen (Sass, 2011). Auch wenn keine Variablen unterschiedlicher Ebenen simultan analysiert werden, wird die genestete Datenstruktur mit zwei Ebenen (Klassen, Schülerinnen und Schüler) in den Analysen berücksichtigt und eine Korrektur des Standardfehlers vorgenommen.

Die Beurteilung der Veränderungen der Passung bei schrittweise eingeführten Restriktionen und Gleichsetzungen der Parameter über die Substichproben hinweg erfolgt mittels des Satorra-Bentler-χ^2-Differenztests (Satorra & Bentler, 2001). Die Invarianz der Parameter ist gegeben, wenn dieser nicht signifikant ausfällt.

Für alle Analysen wurde die Software Mplus Version 7 (Muthén & Muthén, 2012) genutzt.

Ergebnisse

Messinvarianz zu Skalen der Lesemotivation

Baseline-Modell. Wird mittels einer KFA die anzunehmende dreifaktorielle Struktur erzwungen, ergibt sich keine ausreichend gute Modellpassung für die beiden Stichproben. Item h lädt bei hohen Residuen auf allen drei Dimensionen. Dieses Item entstammt der Subskala Selbstkonzept Lesen, welche bereits durch eine unzureichende interne Konsistenz auffiel. Nach Ausschluss dieses Items sowie der Spezifizierung einer Korrelation der Fehlerterme zweier Items

der Subskala Lesen aus Interesse, erreichen die Modelle zufriedenstellende Fitindices für beide Schülergruppen (Förderschulen: $\chi^2(16) = 20.451$, $p = .201$, TLI = .992, CFI = .995, RMSEA = .024; SRMR = .023; Hauptschulen: $\chi^2(16) = 14.712$, $p = .546$, TLI = 1.001, CFI = 1.000, RMSEA = .000, SRMR = .016; s. Tabelle 2). Die Fehlerkovarianzen deuten darauf hin, dass in diesem Modell bestimmte Zusammenhänge nicht ausreichend berücksichtigt sind. Da die Daten beider Subgruppen für das spezifizierte Baseline-Modell jedoch eine hinreichende Passung aufweisen, wird mit der Prüfung der Messinvarianz fortgefahren.

Positionseffekte. In einem ersten Schritt wird geprüft, ob die beiden eingesetzten Versionen mit rotierten Itemreihenfolgen für die Stichprobe der Förderschülerinnen und -schüler zu einer messinvarianten Messung der Konstrukte geführt haben. Es könnte vermutet werden, dass die Position der Items am Anfang bzw. am Ende des Fragebogens einen Effekt auf die Erfassung der Faktoren zur Lesemotivation hat. In einem Mehrgruppenvergleich wird daher die Messäquivalenz der Daten der beiden rotierten Instrumente untersucht.

Das aufgestellte dreifaktorielle Modell führt auch in den beiden Subgruppen zu einer sehr guten Passung (s. Tabelle 3), so dass es als Baseline-Modell für die Prüfung der Messinvarianz zugrunde gelegt werden kann (Instrument 1: $\chi^2(16) = 20.707$, $p = .190$, TLI = .982, CFI = .990, RMSEA = .033, SRMR = .029; Instrument 2: $\chi^2(16) = 19.195$, $p = .259$, TLI = .989, CFI = .993, RMSEA = .029, SRMR = .025). Das Modell der konfiguralen Messinvarianz für beide Instrumentengruppen, für die die gleiche Anzahl an Faktoren sowie das gleiche Muster an freien und fixierten Parametern spezifiziert ist, ergibt ebenso eine sehr

Tabelle 2: Faktorladungen, Faktorkorrelationen und Modellpassungen zur Lesemotivation

Förderschule ($n = 493$)							**Hauptschule ($n = 622$)**						
Faktorladungen							***Faktorladungen***						
	LL		LI		SK			LL		LI		SK	
Item a	.823***						Item a	.899***					
Item b	.877***						Item b	.927***					
Item c	.686***						Item c	.777***					
Item d			.718***				Item d			.752***			
Item e			.650***				Item e			.862***			
Item f			.657***				Item f			.745***			
Item g					.822***		Item g					.568***	
Item i					.587***		Item i					.994***	
Faktorkorrelationen							***Faktorkorrelationen***						
LL	1.00						LL	1.00					
LI	.903***		1.00				LI	.944***		1.00			
SK	-.150*		-.242**		1.00		SK	-.150**		-.233***		1.00	
Fitindices							***Fitindices***						
χ^2	df	p	TLI	CFI	RMSEA	SRMR	χ^2	df	p	TLI	CFI	RMSEA	SRMR
20.451	16	.201	.992	.995	.024	.023	14.712	16	.546	1.001	1.000	.000	.016

Anmerkungen. LL = Leselust. LI = Lesen aus Interesse. SK = Selbstkonzept Lesen.
*** $p < .001$; ** $p < .01$; * $p < .05$. Angaben in standardisierten Werten.

Tabelle 3: Indikatoren zur Messinvarianz für die Skalen zur Lesemotivation in den verschiedenen Instrumenten für Förderschülerinnen und -schüler

Modell	χ^2	df	p	TLI	CFI	RMSEA	SRMR	$\Delta\chi^2(\Delta df)$	p
Instrument 1 (n = 263)	20.707	16	.190	.982	.990	.033	.029		
Instrument 2 (n = 230)	19.195	16	.259	.989	.993	.029	.025		
Konfigurale Invarianz	40.316	32	.149	.985	.991	.032	.027		
Metrische Invarianz	46.962	40	.209	.990	.993	.027	.044	5.312 (8)	.724
Skalare Invarianz	49.580	45	.296	.994	.995	.020	.044	2.172 (5)	.825

gute Passung ($\chi^2(32) = 40.316$, $p = .149$, TLI = .985, CFI = .991, RMSEA = .032, SRMR = .027). Wird die zusätzliche Restriktion der gleichen Faktorladungen zur Prüfung der metrischen Invarianz in beiden Gruppen hinzugefügt, führt dies nicht zu einer signifikanten Verschlechterung des Modellfits ($\Delta\chi^2(8) = 5.312$, $p = .724$). Die Annahme gleicher Intercepts in beiden Gruppen zur Prüfung der skalaren Invarianz bringt ebenfalls keine bedeutsame Verschlechterung des χ^2-Wertes ($\Delta\chi^2(5) = 2.172$, $p = .825$) oder der Fitindices mit sich.

Für das Konstrukt der Lesemotivation ist somit sowohl konfigurale, metrische als auch skalare Messinvarianz gegeben. Somit kann davon ausgegangen werden, dass das Konstrukt der Lesemotivation mit drei latenten Faktoren äquivalent mit den beiden rotierten Befragungsinstrumenten bei Schülerinnen und Schülern mit SPF-L erhoben wird.

Vergleich mit der Hauptschule. Als weitere Forschungsfrage ist zu untersuchen, ob die Annahme der starken faktoriellen Invarianz auch für beide Gruppen der Schülerinnen und Schüler an Förderschulen und an Hauptschulen gelten kann. Die eingesetzten Instrumente unterscheiden sich sowohl hinsichtlich der Länge als auch im Darbietungsmodus. In einem analogen Vorgehen wird ein Mehrgruppenvergleich für die Gruppen der Schülerschaft an Förderschulen und Hauptschulen vorgenommen. Basierend auf dem bestätigten Baseline-Modell werden konfigurale, metrische und skalare Messinvarianz geprüft (s. Tabelle 4).

Die simultane Schätzung des Modells bei nicht fixierten Parametern führt zu einer guten Passung ($\chi^2(32) = 35.054$, $p = .325$, TLI = .998, CFI = .999, RMSEA = .013, SRMR = .019). Die Restriktion gleicher Faktorladungen in beiden Gruppen führt zu einer signifikanten Steigerung des χ^2-Wertes ($\Delta\chi^2(8) = 23.292$, $p = .003$). Erlaubt man dem Modell, die Ladung eines Indikators auf

Tabelle 4: Indikatoren zur Messinvarianz für die Skalen zur Lesemotivation bei Schülerinnen und Schülern an Hauptschulen und Förderschulen

Modell (n = 1115)	χ^2	df	p	TLI	CFI	RMSEA	SRMR	$\Delta\chi^2(\Delta df)$	p
Konfigurale Invarianz	35.054	32	.325	.998	.999	.013	.019		
Metrische Invarianz	53.441	40	.076	.993	.995	.025	.051	23.292 (8)	.003
Partielle metrische Invarianz	45.388	39	.223	.997	.998	.017	.037	11.673 (7)	.112
Skalare Invarianz*	52.720	44	.173	.996	.997	.019	.040	7.719 (5)	.172

* Modell basiert auf partieller metrischer Invarianz.

den latenten Faktor Leseinteresse (Item e) in beiden Gruppen frei zu schätzen, wird der Anstieg des χ^2-Wertes ($\Delta\chi^2(7) = 11.673$, $p = .112$) nicht signifikant. Dies ist ein Fall von partieller metrischer Invarianz. Da jedoch für den Faktor Leseinteresse zwei Items messäquivalente Faktorladungen aufweisen, kann dennoch von einer invarianten Erfassung des Faktors ausgegangen werden (Brown, 2006; Byrne et al., 1989). Die zusätzliche Annahme gleicher Intercepts aller acht Indikatoren kann auf Grund des χ^2-Differenz-Tests ($\Delta\chi^2(5) = 7.719$, $p = .172$) beibehalten werden.

Trotz der partiellen metrischen Invarianz kann somit auf Grundlage der Ergebnisse von einer starken Messinvarianz ausgegangen werden, die zeigt, dass die Erfassung an Förderschulen und Hauptschulen für das Konstrukt der Lesemotivation messinvariant erfolgt. Ist starke Messäquivalenz gegeben, können Differenzen zwischen den Faktormittelwerten beider Gruppen sinnvoll interpretiert werden, da sie sich auf existierende Gruppenunterschiede beziehen und nicht auf nicht invariante Messungen zurückzuführen sind.

Messinvarianz zu Skalen des akademischen Selbstkonzeptes

Baseline-Modell. Ein konfirmatorisches Modell mit der theoretisch zugrundeliegenden Struktur ergibt zunächst lediglich eine akzeptable Modellpassung (Förderschule: $\chi^2(17) = 45.990$, $p < .001$, TLI = .942, CFI = .965, RMSEA = .059; SRMR = .038; Hauptschule: $\chi^2(17) = 136.536$, $p < .001$, TLI = .878, CFI = .926, RMSEA = .102, SRMR = .054). Da das Item d der Subskala mathematisches Selbstkonzept zusätzlich auf dem Faktor des verbalen Selbstkonzeptes lädt und somit nicht dem theoretisch postulierten Ladungsmuster entspricht, wird dieses Item aus den weiteren Analysen ausgeschlossen. Das nachfolgende spezifizierte Modell führt zu einer hinreichenden Modellpassung für beide Stichproben (Förderschule: $\chi^2(11) = 20.989$, $p = .034$, TLI = .974, CFI = .986, RMSEA = .043, SRMR = .024; Hauptschule: $\chi^2(11) = 38.677$, $p < .001$, TLI = .954, CFI = .976, RMSEA = .061, SRMR = .027; s. Tabelle 5).

Positionseffekte. Betrachtet man für die Stichprobe der Schülerinnen und Schüler mit SPF-L an Förderschulen die Messinvarianz der beiden Instrumentenvarianten, zeigen sich folgende Modellpassungen im Mehrgruppenvergleich (s. Tabelle 6).

Das zuvor aufgestellte Baseline-Modell weist auch in den beiden Subgruppen der beiden Förderschulinstrumente eine gute Passung auf (Instrument 1: $\chi^2(11) = 20.700$, $p = .037$, TLI = .953, CFI = .976, RMSEA = .058, SRMR = .034; Instrument 2: $\chi^2(11) = 10.015$, $p = .529$, TLI = 1.006, CFI = 1.000, RMSEA = .000, SRMR = .023). Die weitere Prüfung der konfiguralen, metrischen und skalaren Messinvarianz zeigt, dass zusätzliche Restriktionen keine signifikante Verschlechterung der Passung oder des χ^2-Wertes ergeben, so dass auch für das Konstrukt des akademischen Selbstkonzepts von einer invarianten Messung mit den beiden rotierten Instrumentenversionen ausgegangen werden kann.

Vergleich mit der Hauptschule. Der Vergleich mit der Stichprobe an Hauptschulen zeigt jedoch, dass bei Gleichsetzung der Parameter die Modellpassung teilweise signifikant abnimmt (s. Tabelle 7).

Die Gleichsetzung der Faktorladungen aller Indikatoren führt zu einem signifikanten χ^2-Differenz-Test ($\Delta\chi^2(7) = 17.688$, $p = .013$). Für die Faktoren des verbalen und des schulischen Selbstkonzeptes muss jeweils die Ladung eines Indikators (Item c und h) in beiden Gruppen frei geschätzt werden, um keine signifikante Modellverschlechterung zu erreichen ($\Delta\chi^2(5) = 9.830$, $p = .080$). Die Gleichsetzung der Intercepts aller Variablen führt ebenso zu einer signifikanten Verschlechterung des χ^2-Wertes ($\Delta\chi^2(4) = 14.555$, $p < .001$). Ein Intercept des schulischen Selbstkonzeptes (Item i) muss über beide Gruppen frei geschätzt wer-

den, um keinen signifikanten χ^2-Differenz-Test zu erhalten ($\Delta\chi^2(3) = 1.459$, $p = .692$). Dieses Item weist somit keinen invarianten Intercept für beide Schülergruppen auf.

Für das Konstrukt des akademischen Selbstkonzeptes gibt es Fälle partieller metrischer und partieller skalarer Messinvarianz. Bezüglich der metrischen Messinvarianz sind zwei Ladungen der insgesamt sieben Items als nicht invariant einzustufen. Zusätzlich weist ein weiterer Indikator einen nicht messäquivalenten Intercept auf. Somit ist die Voraussetzung nach Byrne et al. (1989) zweier messinvarianter Indikatoren pro latentem Faktor nicht gegeben.

Tabelle 5: Faktorladungen, Faktorkorrelationen und Modellpassungen zum akademischen Selbstkonzept

Förderschule ($n = 490$)							**Hauptschule ($n = 674$)**						
Faktorladungen							***Faktorladungen***						
	V	M	S					V	M	S			
Item b	0.718***						Item b	.745***					
Item c	0.774***						Item c	.765***					
Item e		0.820***					Item e		.842***				
Item f		0.775***					Item f		.849***				
Item g			0.770***				Item g			.786***			
Item h			0.754***				Item h			.756***			
Item i			0.784***				Item i			.792***			
Faktorkorrelationen							***Faktorkorrelationen***						
V	1.00						V	1.00					
M	0.353***	1.00					M	.078	1.00				
S	0.807***	0.575***	1.00				S	.704***	.423***	1.00			
Fitindices							***Fitindices***						
χ^2	df	p	TLI	CFI	RMSEA	SRMR	χ^2	df	p	TLI	CFI	RMSEA	SRMR
20.989	11	.034	.974	.986	.043	.024	38.677	11	<.001	.954	.976	.061	.027

Anmerkungen. V = Verbales Selbstkonzept. M = Mathematisches Selbstkonzept. S = Schulisches Selbstkonzept.
*** p<.001; ** p<.01; * p<.05. Angaben in standardisierten Werten.

Tabelle 6: Indikatoren zur Messinvarianz für die Skalen des akademischen Selbstkonzepts in den verschiedenen Instrumenten für Förderschülerinnen und -schüler

Modell	**χ^2**	**df**	**p**	**TLI**	**CFI**	**RMSEA**	**SRMR**	**$\Delta\chi^2(\Delta df)$**	**p**
Instrument 1 (n = 261)	20.700	11	.037	.953	.976	.058	.034		
Instrument 2 (n = 229)	10.015	11	.529	1.006	1.000	.000	.023		
Konfigurale Invarianz	32.353	22	.072	.973	.986	.044	.030		
Metrische Invarianz	42.171	29	.054	.974	.982	.043	.057	9.747 (7)	0.203
Skalare Invarianz	48.635	33	.039	.973	.979	.044	.063	6.777 (4)	0.148

Diskussion

In diesem Artikel wurde die Faktorstruktur der Konstrukte Lesemotivation und des akademischen Selbstkonzeptes betrachtet sowie die Homogenität dieser bei einem rotierten Erhebungsinstrument und bei verschiedenen Schülergruppen untersucht. Zur Prüfung der Messinvarianz wurden konfirmatorische Mehrgruppenvergleiche mit den Restriktionen gleicher Parameter (Faktorstruktur, Faktorladungen, Intercepts) angewendet.

Modellergebnisse

Für die drei Subskalen der Lesemotivation konnten durch Ausschluss eines Items und die Spezifikation einer Fehlertermkorrelation hinreichende Modellpassungen erreicht werden. Das ausgeschlossene Item h der Subskala Selbstkonzept Lesen könnte möglicherweise auf Grund der anderen Polung in der Formulierung nicht zur Faktorstruktur passen bzw. bei einer etwas oberflächlichen Bearbeitung der Fragen zu nicht validen Antworten führen. Die spezifizierte Residuenkorrelation zwischen den Items e und f lässt vermuten, dass weitere Zusammenhänge (noch) nicht adäquat im Modell berücksichtigt sind. Wenn man den Inhalt dieser beiden Items betrachtet, zielen sie augenscheinlich auf das Thema (schulisches) Lernen, wohingegen das Item d derselben Subskala die Freude an der Tätigkeit Lesen betont. Somit kann ein systematischer Messfehler vermutet werden.

Auch für die Skala des akademischen Selbstkonzeptes wurde ein Item aus den Analysen ausgeschlossen, um einen akzeptablen Modellfit zu erhalten. Die KFA zeigt, dass das nachfolgend ausgeschlossene Item d des mathematischen Selbstkonzepts zusätzlich eine signifikante Faktorladung für die verbale Komponente aufweist. Dieses Befundmuster einer Kreuzladung ist in beiden Stichproben zu finden und führt dazu, dass dieses Item nicht in der intendierten Form zu interpretieren ist. Möglicherweise führt die einheitliche Formulierung der aufeinanderfolgenden Items (c: „Im Fach Deutsch bekomme ich gute Noten“ und d: „Im Fach Mathe bekomme ich gute Noten“) zu nicht nach Schulfächern differenzierten Angaben.

Positionseffekte

Die Modelltests konnten zeigen, dass die rotierten Instrumente für die Schülerinnen und Schüler mit SPF-L an Förderschulen zu einer messinvarianten Erfassung der Subskalen zur Lesemotivation und zum akademischen Selbstkonzept führten. Die Annahme, dass eine abnehmende Aufmerksamkeit im Verlauf der Befragung zu weniger validen Antworten führen könnte, hat sich nicht bestätigt. Der administrierte Umfang der Befragung stellt sich als angemessen für diese Schülergruppe dar. Zudem ist festzustellen,

Tabelle 7: Indikatoren zur Messinvarianz für die Skalen des akademischen Selbstkonzepts bei Schülerinnen und Schülern an Hauptschulen und Förderschulen

Modell (n = 1164)	χ^2	df	p	TLI	CFI	RMSEA	SRMR	$\Delta\chi^2(\Delta df)$	p-value
Konfigurale Invarianz	59.313	22	< .001	.962	.980	.054	.026		
Metrische Invarianz	77.216	29	< .001	.962	.974	.053	.081	17.688 (7)	.013
Partielle Metrische Invarianz	69.792	27	< .001	.964	.977	.052	.061	9.830 (5)	.080
Skalare Invarianz*	83.516	31	< .001	.962	.972	.054	.062	14.555 (4)	< .001
Partielle Skalare Invarianz*	72.391	30	< .001	.968	.977	.049	.061	1.455 (3)	.692

* = Modell basiert auf partieller metrischer Invarianz.

dass es nicht zu vorzeitigem Abbrechen der Befragung seitens der Schülerinnen und Schüler kam. Auf Grund der vorstrukturierten Bearbeitung mit Hilfe des Vorleseleitfadens wurden alle Teilnehmenden durch das gesamte Instrument geführt. Eine eigenständige Bearbeitung des Fragebogens hätte möglicherweise auf Grund der hohen Leselast einen erhöhten Anteil fehlender Angaben hervorrufen können (Gresch, Strietholt, Kanders & Solga, 2014).

Vergleiche mit der Hauptschule

Die Ergebnisse zur Messinvarianz im Vergleich mit der Gruppe der Hauptschülerinnen und -schüler sind weniger eindeutig zu beurteilen. Betreffend der Skalen zur Lesemotivation kann trotz partieller Messinvarianz von einer starken faktoriellen Invarianz ausgegangen werden.

Die Ladung des Item e auf den Faktor Lesen aus Interesse ist über die Schülergruppen hinweg nicht invariant. Für die Hauptschulgruppe stellt dieses Item den Indikator mit der stärksten Ladung dar, während es für die Schülergruppe an Förderschulen den schwächsten Indikator darstellt (s. Tabelle 2). Dieses Item fragt nach der Überzeugung, durch die Tätigkeit Lesen viel lernen zu können. Diese Formulierung scheint von beiden Gruppen konzeptuell unterschiedlich aufgefasst zu werden, so dass es zu nicht äquivalenten Faktorladungen kommt (Chen, 2008). Da die beiden anderen Indikatoren für den Faktor Lesen aus Interesse invariant sind, könnte man für vergleichende Analysen in Erwägung ziehen, dieses Item auszuschließen (Cheung & Rensvold, 1999).

Das Modell zum akademischen Selbstkonzept weist bei den Schülerinnen und Schüler mit SPF-L eine bessere Passung auf als in der Hauptschulstichprobe. Insgesamt fließen nur sieben Indikatoren in das Modell ein. Daher ist das Ergebnis zweier nicht äquivalenter Faktorladungen (Item c und h) und einem nicht äquivalenten Intercept (Item i) nicht trivial für vergleichende Analysen. Bei Ausschluss der drei genannten nicht messinvarianten Indikatoren würden die Faktoren des verbalen und des schulischen Selbstkonzeptes lediglich jeweils mit einem Item repräsentiert. Unauffälliger ist diesbezüglich die Facette des mathematischen Selbstkonzeptes, das zwei Items mit messinvarianten Faktorladungen und Mittelwerten aufweist. Somit ist der hier vorgestellte Konstruktbereich in seiner Gesamtheit nicht für schulformübergreifende Analysen geeignet.

Zwei Items weisen invariante Faktorladungen auf. Diese beiden sind sich inhaltlich ähnlich, als dass sie jeweils nach Noten im Fach Deutsch bzw. in Klassenarbeiten fragen. Die Problematik dieser Fragestellung liegt darin begründet, dass die Bewertung an Förderschulen anders erfolgt als an Hauptschulen: Häufig gibt es an Förderschulen individuelle Berichterstattungen über Lernfortschritte anstelle von Noten. So geben 21.5 bzw. 23.9 % der Schülerinnen und Schüler an Förderschulen an, dass sie in ihrem letzten Zeugnis keine Note für das Fach Mathe bzw. Deutsch erhalten haben. In Hauptschulen machen nur 2.2 bzw. 4.0% der Zielpersonen diese Angabe.

Zusätzlich zeigten die Mehrgruppenvergleiche, dass der Intercept des Items i („Ich bin in den meisten Schulfächern gut.“) nicht messinvariant in den beiden Schülergruppen ist. Die Schülerinnen und Schüler an Förderschulen weisen hier einen höheren Mittelwert auf als die Schülergruppe an Hauptschulen. Diese Tatsache allein erklärt nicht die Nicht-Äquivalenz des Parameters. Grund für die mangelnde Messinvarianz könnte jedoch sein, dass Schülerinnen und Schüler, die einen ähnlichen Faktormittelwert für die Skala des schulischen Selbstkonzeptes aufweisen, in den beiden Stichproben nicht mit dem gleichen Ausmaß an Zustimmung auf dieses Item reagieren (Chen, 2008).

Wird das akademische Selbstkonzept dieser verschiedenen Schülerpopulationen betrachtet und verglichen, sind Referenzgruppeneffekte zu bedenken. In Förder-

schulen ist eine Art „Schonraum“ geschaffen worden, in dem auch besonderer Wert auf soziale Entwicklung sowie schulische und persönliche Erfolgserlebnisse gelegt wird. Da das akademische Selbstkonzept von schulischen Erfahrungen, Erfolgs- und Misserfolgserlebnissen, sozialen Vergleichen und der Interpretation dieser geformt wird, spiegeln sich pädagogische und didaktische Maßnahmen des differenzierten Unterrichts und der segregierten Schulsettings darin wider (Bos et al., 2010). Auf Grund der unterschiedlichen Rahmenbedingungen könnte sich das akademische Selbstkonzept bei diesen Schülergruppen nicht nur in Bezug auf die mittlere Ausprägung anders darstellen, sondern sich auch bezüglich der relativen Beiträge der Einzelitems (d. h. konzeptuell) unterscheiden.

Ausblick

Die Summe der Anpassungen der Administrationsbedingung bei der vorliegenden Studie hat für die Gruppe der Schülerinnen und Schüler mit SPF-L zu aussichtsreichen Ergebnissen geführt. Daher lohnt es sich, diese Akkommodationen bei schriftlichen Befragungen dieser und auch weiterer Schülergruppen ausführlicher zu untersuchen sowie weiterführende Untersuchungen zu spezifischen Effekten der Anpassungen, wie z. B. dem Vorlesen, durchzuführen. Die dargestellten Ergebnisse sind beispielhafte Resultate, die sich nicht zwingend auf weitere Konstrukte und Skalen übertragen lassen. Hierzu wären weitere Studien wünschenswert.

In der Sonderpädagogik als auch in der Bildungsforschung bleiben die messinvarianten und damit vergleichbaren Messungen von bildungsrelevanten Konstrukten und Kompetenzen weiterhin ein wichtiges Forschungsfeld. Zentrale Fragen bezüglich der Entwicklung von Schülerinnen und Schülern an Förderschulen im Vergleich zu Schülerinnen und Schülern ohne SPF in anderen Schulformen oder auch im Vergleich zu sogenannten Integrationskindern im allgemeinen Schulsystem sind bisher noch unbeantwortet und bedürfen vertiefter Analysen sowie weiterer Forschung. Die experimentellen Machbarkeitsstudien im Rahmen des NEPS und die damit generierten Daten können hierzu einen wichtigen Beitrag leisten.

Literaturverzeichnis

Artelt, C., Naumann, J. & Schneider, W. (2010). Lesemotivation und Lernstrategien. In E. Klieme, C. Artelt, J. Hartig, N. Jude, O. Köller, M. Prenzel, W. Schneider & P. Stanat (Hrsg.), *PISA 2009. Bilanz nach einem Jahrzehnt* (S. 73-112). Münster: Waxmann.

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

Blossfeld, H.-P., Roßbach, H.-G. & von Maurice, J. (Hrsg.) (2011). Education as a lifelong Process – The German National Educational Panel Study (NEPS). *Zeitschrift für Erziehungswissenschaft,* Sonderheft 14.

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.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: Guilford Press.

Byrne, B. M. (1996). Academic self-concept: Its structure, measurement, and relation to academic achievement. In B. A. Bracken (Hrsg.), *Handbook of self-concept: developmental, social, and clinical considerations* (S. 287–316). New York: Wiley.

Byrne, B. M., Shavelson, R. J. & Muthén, B. O. (1989). Testing for the equivalence of factor covariance and mean structures. The issue of partial measurement invari-

ance. *Psychological Bulletin*, *105*(3), 456–466.

Byrne, B. M. & van de Vijver, F. J. R. (2010). Testing for measurement and structural equivalence in large-scale cross-cultural studies: Addressing the issue of nonequivalence. *International Journal of Testing*, *10*(2), 107–132.

Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of Personality and Social Psychology*, *95*(5), 1005–1018.

Cheung, G. W. & Rensvold, R. B. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new method. *Journal of Management*, *25*, 1-27.

Cheung, G. W. & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 9*(2), 233–255.

Cormier, D. C., Altman, J., Shyyan, V. & Thurlow, M. L. (2010). *A summary of the research on the effects of test accommodations: 2007-2008* (Technical Report 56). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Dimitrov, D. (2010). Testing for factorial invariance in the context of construct validation. *Measurement and Evaluation in Counseling and Development*, *43*(2), 121–149.

Frahm, S., Goy, M., Kowalski, K., Sixt, M., Strietholt, R., Blatt, I., Bos, W. & Kanders, M. (2011): Transition and development from lower secondary to upper secondary school. In H.-P. Blossfeld, H.-G. Roßbach & J. Maurice (Hrsg.), *The National Educational Panel Study: need, main features, and research potential* (S. 217–232). Wiesbaden: VS Verlag für Sozialwissenschaften.

Gebhardt, M., Oelkrug, K. & Tretter, T. (2013). Das mathematische Leistungsspektrum bei Schülerinnen und Schülern mit sonderpädagogischem Förderbedarf in der Sekundarstufe. Ein explorativer Querschnitt der fünften bis neunten Klassenstufe in Münchner Förderschulen. *Empirische Sonderpädagogik, 5*(2), 130–143.

Gresch, C., Strietholt, R., Kanders, M. & Solga, H. (2014). Reading-aloud versus self-administrated student questionnaires: An experiment on data quality. Manuscript submitted for publication.

Grünke, M. (2004). Lernbehinderung. In G. W. Lauth, M. Grünke & J. C. Brunstein (Hrsg.), *Interventionen bei Lernstörungen. Förderung, Training und Therapie in der Praxis* (S. 65–77). Göttingen: Hogrefe.

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.

Hoermann, B. (2007). *Die Unsichtbaren in PISA, TIMSS & Co. Kinder mit Lernbehinderungen in nationalen und internationalen Schulleistungsstudien*. Verfügbar unter: http://bildungswissenschaft.univie.ac.at/fe2/fileadmin/user_upload/inst_bildungswissenschaft/Diplomarbeit_Hoermann.pdf.

Horn, J. L. & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*, *18*, 117–144

Hu, L.-T. & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55.

Klauer, K. J. & Lauth, G. W. (1997). Lernbehinderungen und Leistungsschwierigkeiten bei Schülern. In F. E. Weinert (Hrsg.), *Psychologie des Unterrichts und der Schule* (Enzyklopädie der Psychologie, Themenbereich D, Serie I, Pädagogische Psychologie) (S. 701-738). Göttingen: Hogrefe.

Koretz, D. & Barton, K. (2004). Assessing students with disabilities: Issues and evi-

dence. *Educational Assessment*, *9*(1/2), 29-60.

Kunter, M., Schümer, G., Artelt, C., Baumert, J., Klieme, E., Neubrand, M., Prenzel, M., Schiefele, U., Schneider, W., Stanat, P., Tillmann, K.-J. & Weiß, M. (2002). *PISA 2000: Dokumentation der Erhebungsinstrumente*. Berlin: Max-Planck-Institut für Bildungsforschung.

Kränzl-Nagl, R. & Wilk, L. (2000). Möglichkeiten und Grenzen standardisierter Befragung unter besonderer Berücksichtigung der Faktoren soziale und personale Wünschbarkeit. In F. Heinzel (Hrsg.), *Methoden der Kindheitsforschung. Ein Überblick über Forschungszugänge zur kindlichen Perspektive* (S. 59-75). Weinheim: Juventa.

Lipski, J. (2000). Zur Verläßlichkeit der Angaben von Kindern bei standardisierten Befragungen. In F. Heinzel (Hrsg.), *Methoden der Kindheitsforschung. Ein Überblick über Forschungszugänge zur kindlichen Perspektive* (S. 77-86). Weinheim: Juventa.

Marsh, H. W. (1986). Verbal and math self-concepts: An internal/external frame of reference model. *American Educational Research Journal*, *23*, 129–149.

Marsh, H. W. (1990). The Structure of academic self-concept: The Marsh/Shavelson Model. *Journal of Educational Psychology*, *82*(4), 623–636.

McElvany, N., Kortenbruck, M. & Becker, M. (2008). Lesekompetenz und Lesemotivation. Entwicklung und Mediation des Zusammenhangs durch Leseverhalten. *Zeitschrift für Pädagogische Psychologie, 22*(3-4), 207–219.

Möller, J. & Bonerad, E.-M. (2007). Fragebogen zur habituellen Lesemotivation. *Psychologie in Erziehung und Unterricht*, *54*, 259–267.

Möller, J. & Köller, O. (2004). Die Genese akademischer Selbstkonzepte: Effekte dimensionaler und sozialer Vergleiche. *Psychologische Rundschau*, *55*(1), 19-27.

Möller, J. & Schiefele, U. (2004). Motivationale Grundlagen der Lesekompetenz. In U. Schiefele, C. Artelt, W. Schneider & P. Stanat (Hrsg.), *Struktur, Entwicklung und Förderung von Lesekompetenz. Vertiefende Analysen im Rahmen von PISA 2000* (S. 101–124). Wiesbaden: Verlag für Sozialwissenschaften.

Muthén, L. K. & Muthén, B. O. (2012). Mplus Version 7 [Computersoftware]. Los Angeles, CA.

Sass, D. A. (2011). Testing measurement invariance and comparing latent factor means within a confirmatory factor analysis framework. *Journal of Psychoeducational Assessment, 29*(4), 347–363.

Satorra, A. & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, *66*, 507-514.

Schiefele, U. (1996). *Motivation und Lernen mit Texten*. Göttingen: Hogrefe.

Scholl, A. (2003). *Die Befragung. Sozialwissenschaftliche Methode und kommunikationswissenschaftliche Anwendung*. Konstanz: UVK (UTB).

Shavelson, R. J., Hubner, J. J. & Stanton, G. C. (1976). Selfconcept: Validation of construct interpretations. *Review of Educational Research*, *46*, 407–444.

Statistisches Bundesamt (2011). Bildung und Kultur. Allgemeinbildende Schule. Schuljahr 2010/2011. Fachserie 11. Wiesbaden.

Ständige Konferenz der Kultusminister der Länder (1994). Empfehlungen zur sonderpädagogischen Förderung in den Schulen in der Bundesrepublik Deutschland. Verfügbar unter: http://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/1994/1994_05_06-Empfehlung-sonderpaed-Foerderung.pdf

Steenkamp, J.-B. E. M. & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, *25*(1), 78–107.

Steinmetz, H., Schmidt, P., Tina-Booh, A., Wieczorek, S. & Schwartz, S. H. (2009). Testing measurement invariance using multigroup CFA: Differences between

educational groups in human values measurement. *Quality & Quantity*, *43*(4), 559–616.

Vandenberg R. J. & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*(1), 4–69.

Wohlkinger, F., Ditton, H., von Maurice, J., Haugwitz, M. & Blossfeld, H.-P. (2011). Motivational concepts and personality aspects across the life course. In H.-P. Blossfeld, H.-G. Roßbach & J. Maurice (Hrsg.), *The National Educational Panel Study: need, main features, and research potential* (S. 155–168). Wiesbaden: VS Verlag für Sozialwissenschaften.

Lena Nusser
Leibniz-Institut für Bildungsverläufe e.V.
Wilhelmsplatz 3
96047 Bamberg
lena.nusser@lifbi.de

Erstmalig eingereicht: 03.11.2014
Überarbeitung eingereicht: 02.02.2015
Angenommen: 07.02.2015

Beitrag 4

Nusser, L. & Wolter, I. (2016). There's plenty more fish in the sea. Das akademische Selbstkonzept von Schülerinnen und Schülern mit sonderpädagogischem Förderbedarf Lernen in integrativen und segregierten Schulsettings. *Empirische Pädagogik, 30* (1), 130-143.

Empirische Pädagogik
2016, 30. Jahrgang, Heft 1, S. 130-143
© Verlag Empirische Pädagogik

There's plenty more fish in the sea. Das akademische Selbstkonzept von Schülerinnen und Schülern mit sonderpädagogischem Förderbedarf Lernen in integrativen und segregierten Schulsettings

Lena Nusser & Ilka Wolter

Der in vielen Untersuchungen bestätigte Big-Fish-Little-Pond-Effekt (BFLPE) hat gezeigt, dass das akademische Selbstkonzept von Schülerinnen und Schülern nicht nur durch eigene erreichte Leistungen geformt, sondern auch negativ von den Leistungen der Klassengemeinschaft beeinflusst wird. Ausgehend von empirischen Befunden, die zeigen, dass das akademische Selbstkonzept für die schulische Leistungsentwicklung relevant ist und sich im sozialen Vergleich entwickelt (vgl. Möller & Köller, 2004), soll der BFLPE bei der Schülerschaft an Förderschulen Lernen (n = 587 in 91 Klassen) mehrebenenanalytisch überprüft werden. Die Autorinnen nehmen an, dass die Mechanismen des sozialen Vergleiches auch in segregierten Schulsettings ihre Wirkung entfalten und sich ein negativer Effekt des mittleren Leistungsniveaus der Klasse auf das individuelle akademische Selbstkonzept nachweisen lässt. Zudem wurde das Modell des BFLPE vergleichend mit integrativ beschulten Lernenden (n = 148 in 103 Klassen) untersucht. Da es sich bei integrativen Klassen für die Kinder mit einem sonderpädagogischen Förderbedarf Lernen um einen vergleichsweise leistungsstärkeren Kontext handeln sollte, wird angenommen, dass der negative Effekt auf das akademische Selbstkonzept für Integrationskinder (d. h. Kinder mit Förderbedarf) in Regelschulen stärker ausfällt als für Lernende an Förderschulen. Es zeigte sich, dass in Förderschulklassen das mittlere Leistungsniveau der Klasse (d. h. Notendurchschnitt der Klasse) unter Berücksichtigung der individuellen Leistung keinen Einfluss auf das akademische Selbstkonzept hatte, während in Integrationsklassen ein deutlicher Kontexteffekt in erwarteter Richtung zu finden war.

Schlagwörter: akademisches Selbstkonzept – Big-Fish-Little-Pond-Effekt – Förderschule – Integration – sonderpädagogischer Förderbedarf

1 Einleitung

Im Rahmen der Erforschung der Leistungsentwicklung sowie der Selbstkonzepte von Kindern mit sonderpädagogischem Förderbedarf (SPF) gelangen zunehmend Fragestellungen zum Vergleich des Förderschulsettings mit integrativen Settings in Regelschulen in den Fokus. Es lässt sich interessanterweise feststellen, dass, obwohl Kinder in Regelschulen im Mittel bessere Leistungen in standardisierten Tests zeigen als Kinder mit einem sonderpädagogischen Förderbedarf Lernen (SPF-L) in Förderschulen (Kocaj, Kuhl, Kroth, Pant & Stanat, 2014), sich die Selbstkonzepte von Kindern mit SPF-L in Förderschulen von den Kindern mit SPF-L in Integrations-

klassen teilweise deutlich und in umgekehrter Richtung unterscheiden. Dies wird zurückgeführt auf die unterschiedlichen Vergleichskontexte und -gruppen für die Schülerschaft in segregierten und inklusiven Schulklassen (vgl. Möller, 2013). Obwohl die Befunde bisheriger Studien auf das Wirken sozialer Vergleichsprozesse hinweisen, gibt es bis dato unseres Wissens keine Studie, die den tatsächlichen schulischen Leistungskontext empirisch modelliert und diese psychologischen Mechanismen für Lernende mit SPF-L in verschiedenen Schulsettings nachweist. Insgesamt lässt sich aus empirischen Arbeiten ableiten, dass sich die Kinder in Förderschulsettings vor allem in Bezug auf ihre Selbstkonzepte zur Peer-Akzeptanz, dem globalen Selbstwert und der emotionalen Integration von jenen in Integrationsklassen unterscheiden, jedoch nicht in Bezug auf das kognitive Selbstkonzept (Schwab, 2014). Vor dem Hintergrund der Befunde von Kocaj et al. (2014), dass auch die Schülerschaft mit SPF in Regelschulen höhere Kompetenzen in den Bereichen Lesen und Mathematik aufweisen als in Förderschulen, stellen Regelschulen einen systematisch stärkeren Leistungskontext dar als Förderschulen. Der für die vorliegende Arbeit relevante und im Zuge der Selbstkonzeptforschung empirisch bestätigte Einfluss des Kontextes zeigt sich in dem vielzitierten Big-Fish-Little-Pond-Effekt (BFLPE; z. B. Marsh, Trautwein, Lüdtke & Köller, 2008; Seaton, Marsh & Craven, 2010; Trautwein, Lüdtke, Marsh, Köller & Baumert, 2006). Dieser Effekt beschreibt den konsistent empirisch auffindbaren Befund, dass das akademische Selbstkonzept von Lernenden nicht nur durch ihre eigenen erreichten Leistungen bestimmt wird, sondern darüber hinaus auch negativ von der mittleren Leistung der Klasse der Schülerin oder des Schülers beeinflusst wird.

Die Leistungsstärke des Kontextes kann demnach für die individuelle Ausprägung des akademischen Selbstkonzeptes eine entscheidende Rolle spielen. Diese Befunde sind unseres Erachtens in der Diskussion um Beschulungsformen im Sinne der Integration und der Segregation der Lernenden mit SPF-L nicht zu vernachlässigen. Dieser Schülergruppe wurden bereits sehr früh Hilfs- und Sonderschulen zugewiesen, um ihnen dort in kleineren Lerngruppen eine bedarfsgerechte Förderung zukommen zu lassen (Pfahl, 2008). Die Schaffung eines angenommenen leistungshomogeneren (Lern-)Umfeldes an Förderschulen sollte nach der Ratifikation der UN-Behindertenrechtskonvention 2009 der Vergangenheit angehören. Ein inklusives Bildungsangebot für alle Kinder bedeutet mehr Heterogenität hinsichtlich unterschiedlicher Variablen, aber vor allem auch hinsichtlich der schulischen Leistungen und Performanz der Kinder. Die Umgebung und Klassenzusammensetzung unterscheiden sich bei segregierten und integrativen Schulformen daher deutlich, so dass sich diese Umstände auch auf die Ausformung des akademischen Selbstkonzeptes auswirken können (z. B. Gans, Kenny & Ghandy, 2003).

In dieser Arbeit untersuchen wir daher das Fähigkeitsselbstkonzept von Kindern mit SPF-L in fünften Klassen, die entweder in einem segregierten (d. h. Förderschulen) oder einem integrativen Schulsetting (d. h. Integrationsklassen in Regelschulen) lernen.

2 Akademisches Selbstkonzept von Lernenden mit sonderpädagogischem Förderbedarf Lernen

Insgesamt ist der Anteil der Schülerinnen und Schüler mit einem diagnostizierten Förderbedarf an der Gesamtschülerschaft in Deutschland in den letzten Jahrzehnten stetig angestiegen. Dabei bilden Kinder und Jugendliche mit einem sonderpädagogischen Förderbedarf Lernen die größte Gruppe (anteilig fast 40%; Autorengruppe Bildungsbericht, 2014). Sowohl hinsichtlich der Förderquote als auch der diagnostischen Verfahren sind in den einzelnen Bundesländern deutliche Unterschiede festzustellen (Dietze, 2012). Diese Tatsache ist auch eine Konsequenz aus der noch ausstehenden einheitlich anzuwendenden Definition einer Lernbehinderung (Bos, Müller & Stubbe, 2010). SPF-L wird als eine überdauernde und weitreichende Einschränkung bei der Bewältigung schulischer Anforderungen beschrieben (Klauer & Lauth, 1997), die sich im Besonderen bei dem „Erwerb kognitiv-verbaler und abstrakter Inhalte“ (Grünke, 2004, S. 65) zeigt. Zudem werden der Schülerschaft mit SPF-L schwächere sprachliche Fähigkeiten, geringe Lernumfänge sowie eine reduzierte Aufmerksamkeitsspanne zugeschrieben (Grünke, 2004).

Aktuelle Untersuchungen deuten darauf hin, dass Kinder mit SPF-L von einer inklusiven Beschulung profitieren und sich die Leistungen in den Domänen Lesen, Zuhören und Mathematik signifikant von jenen Kindern an Förderschulen unterscheiden (Kocaj et al., 2014). Dieser Effekt lässt sich möglicherweise durch die regelmäßige Interaktion mit leistungsstärkeren Klassenkameradinnen und -kameraden erklären, die eine anregendere Lernumwelt darstellen können (Bos, Müller & Stubbe, 2010). Gleichzeitig bedeutet die gemeinsame Beschulung von Lernenden mit und ohne SPF eine besondere Klassenkomposition, welche sich auch in einer breiten Leistungsspanne sowie in gruppenspezifischen Unterschieden hinsichtlich der gezeigten Schulleistungen widerspiegelt. Diese leistungsbezogenen Besonderheiten können sich wiederum auf die Ausbildung des Selbstkonzeptes auswirken (vgl. Schwab, 2014).

Das akademische Selbstkonzept ist definiert als eine Komponente des globalen Selbstkonzeptes einer Person, also der Gesamtheit des selbstbezogenen Wissens eines Individuums (z. B. Brunner, Keller, Dierendonck, Reichert, Ugen, Fischbach & Martin, 2010; Shavelson, Hubner & Stanton, 1976). Ein im Zuge der Selbstkonzeptforschung bekanntes Phänomen ist der BFLPE: Kinder mit einem hohen

Selbstkonzept erleben im sozialen Vergleich mit einer neuen, sehr leistungsstarken Gruppe einen Selbstkonzeptabfall und vice versa (Marsh, 2005).

Oftmals weisen Kinder und Jugendliche an Förderschulen ein höheres akademisches Selbstkonzept auf als jene mit einem SPF an Regelschulen (Venetz, Tarnutzer, Zurbriggen & Sempert, 2010). Diese Befunde werden auf Referenzgruppeneffekte zurückgeführt, welche die verschiedenen Bezugsnormen für die verschiedenen Schülerpopulationen widerspiegeln (z. B. Möller, Streblow & Pohlmann, 2009). Kinder mit SPF-L an Regelschulen befinden sich in einem leistungsstärkeren Umfeld, was sich in einem geringer berichteten Selbstkonzept ausdrückt, während Lernende an Förderschulen in (augenscheinlich) homogeneren und leistungsschwächeren Lerngruppen ihre Kompetenzen als höher wahrnehmen. Diese Annahmen zur Erklärung im Mittel unterschiedlicher Selbstkonzepte von Kindern und Jugendlichen mit SPF-L in segregierten und inklusiven Schulsettings sind allerdings bislang nicht empirisch über die Modellierung von leistungsbezogenen Kontexteffekten nachgewiesen worden. Zu diesem Zweck operationalisieren wir die individuelle und kontextbezogene Leistung über die erreichten Noten der Lernenden, obwohl Trautwein, Lüdtke, Köller und Baumert (2006) argumentieren, dass aufgrund der teilweise im sozialen Vergleich resultierenden Vergabe von Noten durch die Lehrpersonen, die individuelle Leistung (d. h. Note) bereits auf die mittlere Leistung der Referenzgruppe (d. h. Klassendurchschnittsnote) zurückgeht. Somit stellt der BFLPE eine Konsequenz der Notenvergabepraxis dar (vgl. auch Trautwein et al., 2006). Trotzdem zeigt sich in empi-rischen Arbeiten, dass der mittlere Einfluss der Klassenleistung auch nach Kontrolle der Schulnoten und der individuellen Leistung nachgewiesen werden kann (Lüdtke, Köller, Artelt, Stanat & Baumert, 2002). Ausgehend von der Argumentation nach Möller, Zimmermann und Köller (2014), dass Lernende ihr Selbstkonzept auf Noten, welche aus unserer Sicht die eigentlich sichtbaren Rückmeldungen im Klassenkontext sind, basieren und nicht auf Leistungen in standardisierten Testverfahren basieren, welche aus unserer Sicht die eigentlich sichtbaren Rückmeldungen im Klassenkontext sind, wählen wir die individuelle Note sowie den Notendurchschnitt der Klasse als proxy für die jeweiligen Leistungsstände in der Vorhersage der individuellen Selbstkon-zepte der Schülerschaft mit SPF-L.

Anzumerken ist weiterhin, dass in den Studien, in welchen leistungsbezogene Selbstkonzepte untersucht wurden (z. B. Möller et al., 2009), vergleichsweise wenig Forschungsarbeiten zu domänenspezifischen Selbstkonzepten zu finden sind. Dies ist vor allem interessant, da gezeigt werden konnte, dass die Schülerschaft mit SPF-L durchaus eine sehr heterogene Leistungsgruppe beispielsweise in der Ma-thematik darstellt (vgl. Gebhardt, Oelkrug & Trettner, 2013).

3 Fragestellung

Ausgehend von empirischen Befunden, die zeigen, dass das akademische Selbstkonzept für die schulische Leistungsentwicklung relevant ist und sich im sozialen Vergleich entwickelt (vgl. Möller & Köller, 2004), verfolgt dieser Beitrag zwei Ziele.

Erstens soll der BFLPE an Förderschulen Lernen mehrebenenanalytisch untersucht werden. Es wird angenommen, dass die Mechanismen des sozialen Vergleiches auch in segregierten Schulsettings ihre Wirkung entfalten und sich ein negativer Effekt des mittleren Leistungsniveaus der Klasse (Level 2, Kontextebene) auf die individuellen akademischen Selbstkonzepte (Level 1, Individualebene) nachweisen lässt. Zweitens wird der BFLPE vergleichend bei Lernenden mit SPF-L an Förderschulen und in Integrationsklassen an Regelschulen untersucht. Da es sich bei integrativen Klassen für die Kinder mit SPF-L um einen vergleichsweise leistungsstärkeren Kontext handeln sollte, wird angenommen, dass der negative Effekt auf das akademische Selbstkonzept für Integrationskinder mit SPF-L in Regelschulen stärker ausfällt als für Lernende an Förderschulen. Die Fragestellungen werden domänenspezifisch im Bereich der Mathematik untersucht, da in dieser Domäne aufgrund der stärkeren Verankerung im schulischen Lernkontext größere Effekte zu erwarten sind (vgl. Schurtz, Pfost, Nagengast & Artelt, 2014).

4 Methoden

4.1 Stichprobe

Zur Untersuchung der aufgeworfenen Fragestellungen werden Daten des Nationalen Bildungspanels (NEPS; Blossfeld, Roßbach & Maurice, 2011) herange-zogen[1]. Schülerinnen und Schüler (N = 6112) im 5. Jahrgang (Startkohorte 3), deren Eltern ihr Einverständnis erteilt hatten, haben im Herbst 2010 an der Erhebung teilgenommen. Die hier untersuchte Stichprobe besteht aus n = 587 Schülerinnen und Schülern in 91 Klassen an Förderschulen Lernen (vgl. Heydrich, Weinert, Nusser, Artelt & Carstensen, 2013) sowie n = 148 integrativ beschulten Schülerinnen und Schülern mit SPF-L in 103 Regelschulklassen, deren Eltern eine Diagnose des sonderpädagogischen Förderbedarfs Lernen in einem telefonischen Interview berichteten. Diese Schülergruppe besuchte verschiedene Schularten, überwiegend jedoch Haupt- (31.1 %) oder Realschulen (28.4 %), sowie 9.5 % Grundschulen,

[1] Diese Arbeit nutzt Daten des Nationalen Bildungspanels (NEPS) Startkohorte 3 (Klasse 5), doi: 10.5157/NEPS:SC3:2.0.0. Die Daten des NEPS wurden von 2008 bis 2013 als Teil des Rahmenprogramms zur Förderung der empirischen Bildungsforschung erhoben, welches vom Bundesministerium für Bildung und Forschung (BMBF) finanziert wurde. Seit 2014 wird NEPS vom Leibniz-Institut für Bildungsverläufe e.V. (LIfBi) an der Otto-Friedrich-Universität Bamberg in Kooperation mit einem deutschlandweiten Netzwerk weitergeführt.

18.9 % Schulen mit mehreren Bildungsgängen bzw. Gesamtschulen und 12.1 % Gymnasien.

4.2 Instrument

Das akademische Selbstkonzept wurde auf einer vierstufigen Likert-Skala (1 = "trifft gar nicht zu" bis 4 = "trifft völlig zu") differenziert für die Fächer Deutsch und Mathematik sowie fachübergreifend erfragt (Wohlkinger, Ditton, Maurice, Haugwitz & Blossfeld, 2011). Der vorliegende Beitrag bezieht sich auf die Angaben zum Fach Mathematik. Zudem wurden die Kinder nach ihren Noten im letzten Zeugnis für das Fach Mathematik befragt. Tabelle 1 gibt die Formulierungen sowie deskriptiven Statistiken der zugrundeliegenden Items wieder.

Tabelle 1: Deskriptive Statistiken zum akademischen Selbstkonzept und den Noten in Mathematik

Item	Segregiert beschulte Kinder mit SPF-L			Integrativ beschulte Kinder mit SPF-L			Kinder in Integrationsklassen ohne SPF-L		
	N	M	SD	N	M	SD	N	M	SD
1) Im Fach Mathematik bekomme ich gute Noten.	469	3.14	1.06	131	2.89	0.96	1124	3.07	0.82
2) Mathematik ist eines meiner besten Fächer.	468	3.05	1.16	128	2.84	1.12	1128	2.77	1.07
3) Ich war schon immer gut in Mathematik.	466	2.89	1.13	128	2.73	1.07	1116	2.75	1.02
Mathematiknote im letzten Zeugnis	420	2.56	1.23	145	3.00	1.11	1172	2.55	0.93

Anmerkung. Zeugnisnoten: 1 = sehr gut bis 6 = ungenügend.

Das erste Item „Im Fach Mathematik bekomme ich gute Noten." wurde aufgrund einer signifikanten Kreuzladung auf den Faktor des verbalen akademischen Selbstkonzeptes aus den Analysen ausgeschlossen. Die internen Konsistenzen für die Subskala des mathematischen Selbstkonzeptes liegen zwischen Cronbachs $\alpha = .78$ (an Förderschulen) und Cronbachs $\alpha = .86$ (Kinder mit SPF-L in Integrationsklassen)

sowie Cronbachs α = .84 (Kinder ohne SPF-L in Integrationsklassen) und damit in einem akzeptablen bis guten Bereich. Eine aufgrund der differenziellen kognitiven und auch sprachlichen Ausgangsvoraussetzungen leicht unterschiedliche Administration des Fragebogens an Förderschulen (d. h., Fragen und Antworten wurden laut vorgelesen, siehe hierzu Nusser, Heydrich, Carstensen, Artelt & Weinert, in Druck) resultierte basierend auf Analysen zur Messinvarianz im Vergleich zur Schülerschaft an Hauptschulen in einer etwas eingeschränkten Vergleichbarkeit des Konstruktes zum akademischen Selbstkonzept. Da diese Einschränkung vorrangig in den Subskalen zum verbalen und globalen akademischen Selbstkonzept zu finden war, während die Subskala zum mathematischen Selbstkonzept für die hier relevanten Items invariante Faktorladungen und Intercepts aufwies (Nusser, Carstensen & Artelt, 2015), sehen wir die Voraussetzungen für einen Gruppenvergleich in dieser Studie als gegeben.

4.3 Statistische Analysen

Die Untersuchung der hier thematisierten Individual- und Kontexteffekte wurde anhand von mehrebenenanalytischen Modellen mit der Software Mplus 7 (Muthén & Muthén, 2012) durchgeführt. Aufgrund der nicht vorliegenden Normalverteilung der Daten wurde die robustere Schätzmethode restricted maximum likelihood (MLR) gewählt, die auch unvollständige Datenreihen berücksichtigen kann (Sass, 2011).

Die Mathematiknote (nicht rekodiert) der Schülerinnen und Schüler wurde auf Level 1 als grand-mean zentrierte Variable in das Modell aufgenommen. Im Sinne bisheriger BFLPE-Analysen wurde der Prädiktor durchschnittliche Note als auf Klassenebene aggregierte und wiederum grand-mean zentrierte Variable auf Level 2 in das Modell eingeführt, um Gruppendifferenzen zu adjustieren (Huguet et al., 2009; Trautwein, Marsh & Nagy, 2009). In der Substichprobe der integrativ beschulten Schülerschaft mit SPF-L bezieht sich diese aggregierte Variable auf die mittlere Mathematiknote des gesamten Klassenverbandes, während auf Individualebene nur diejenigen Kinder mit denen aus Förderschulen verglichen wurden, die einen diagnostizierten SPF-L aufwiesen. Zur Überprüfung differenzieller Effekte unserer Ergebnisse wurde dieser Effekt auch für Kinder ohne SPF-L in den Integrationsklassen getestet. Der Faktor mathematisches Selbstkonzept wurde auf Level 1 latent modelliert, um Messfehler zu korrigieren und die Faktorladungen für Level 2 entsprechend fixiert (vgl. Marsh et al., 2009).

Zur Überprüfung, ob der Klasseneffekt der durchschnittlichen Leistung auf das mathematische Selbstkonzept über den individuellen Effekt der eigenen Note hinausgeht, wird der BFLPE als zusätzlicher Parameter direkt berechnet. Dies erfolgt über die Differenz zwischen den geschätzten Regressionsgewichten der indi-

viduellen und mittleren Note auf Level 1 und Level 2 (vgl. Marsh et al., 2009). Zur Berechnung der Effektstärke wurde in Anlehnung an Marsh et al. (2009) die Erweiterung der Formel nach Tymms (2004) zur Berücksichtigung der Gesamtvarianz auf Level 1 und Level 2 genutzt, die eine vergleichbare Interpretation mit Cohens d zulässt.

5 Ergebnisse

Für die drei verschiedenen Zielpopulationen wurde das Modell jeweils getrennt geschätzt. Dabei weisen die Modelle eine sehr gute bis akzeptable Passung auf (für Förderschulen: $\chi^2(1) = 1.688$, RMSEA = 0.034, CFI = 0.997, TLI = 0.982, SRMR (between) = 0.092, SRMR (within) = 0.004; für Kinder mit SPF-L in Integrationsklassen: $\chi^2(1) = 0.821$, RMSEA = 0.000, CFI = 1.000, TLI = 1.011, SRMR (between) = 0.029, SRMR (within) = 0.002; für Kinder ohne SPF-L in Integrationsklassen: $\chi^2(1) = 17.540$, RMSEA = 0.115, CFI = 0.983, TLI = 0.900, SRMR (between) = 0.110, SRMR (within) = 0.006). Weitere Modellergebnisse sind in Tabelle 2 dargestellt.

Für die Lernenden an Förderschulen Lernen zeigt sich, dass der Einfluss der individuellen Note auf das mathematische Selbstkonzept signifikant ist (b = -0.27). Unter Berücksichtigung dieser individuellen Note hat das mittlere Notenniveau der Klassengemeinschaft an Förderschulen jedoch keinen Effekt auf das entsprechende Selbstkonzept (b = 0.09). Der viel berichtete BFLPE kann in dieser Stichprobe nicht reproduziert werden, sondern das akademische Selbstkonzept scheint überwiegend von der individuellen Leistung bzw. der individuellen Rückmeldung der Lehrperson in Form von Noten geprägt zu sein.

Anders hingegen zeigt sich für die Schülerschaft mit SPF-L in Integrationsklassen neben dem Individualeffekt, d. h., je besser die eigene Note ist, desto höher ist auch das berichtete mathematische Selbstkonzept (b = -0.62), ein deutlicher Kontexteffekt in erwarteter Richtung. Je höher der Notendurchschnitt der Klasse im Fach Mathematik ausfällt, desto geringer ist das mathematische Selbstkonzept der Schülerinnen und Schüler mit SPF-L in Integrationsklassen ausgeprägt. Kontrolliert um den Effekt der individuellen Note erweist sich das mittlere Notenniveau der Klasse (Kinder mit und ohne SPF-L) als bedeutsamer Einfluss auf das individuelle mathematikbezogene Selbstkonzept (b = 0.59) der Lernenden mit SPF-L in Regelschulklassen. Dieser deutliche Effekt spiegelt sich auch in der Berechnung des BFLPE = 1.21 wider sowie in einer besonders hohen Effektstärke von 1.26.

Auch für die Kinder ohne SPF-L in Regelschulen finden sich die erwarteten Ergebnisse: Der Effekt der individuellen Note auf das mathematische Selbstkonzept (b = -0.58) fällt vergleichbar zu den Kindern mit SPF-L in den Integrationsklassen

an Regelschulen aus. Weiterhin zeigt sich für Kinder ohne SPF-L in Integrationsklassen unter Kontrolle der individuellen Note ein negativer Effekt des mittleren Klassennotendurchschnitts auf das mathematische Selbstkonzept (b = 0.23). Der BLFPE = 0.80 mit einer hohen Effektstärke von 1.07 findet sich auch in der Schülergruppe ohne SPF-L in Integrationsklassen.

Tabelle 2: Ergebnisse des BFLPE-Mehrebenenmodells

	Segregiert beschulte Kinder mit SPF-L (n = 583)		Integrativ beschulte Kinder mit SPF-L (n = 148)		Kinder in Integrationsklassen ohne SPF-L (n = 1250)	
Mittlere Gruppengröße	6.20		1.44		12.380	
ICC (Item 2)	.055		.157		.112	
ICC (Item 3)	.029		.177		.052	
	b	SE	b	SE	b	SE
Level 1						
Faktorladungen						
Item 2	1.000	0.000	1.000	0.000	1.000	0.000
Item 3	1.259***	0.302	0.834***	0.093	1.081***	0.044
MSK ON Note	-0.266***	0.061	-0.620***	0.083	-0.578***	0.037
Level 2						
Faktorladungen						
Item 2	1.000	0.000	1.000	0.000	1.000	0.000
Item 3	1.259***	0.302	0.834***	0.093	1.081***	0.044
MSK ON Note	0.093	0.073	0.588***	0.189	0.225***	0.061
R^2						
MSK-Level 1	0.166**	0.0504	0.432***	0.088	0.397***	0.035
MSK-Level 2	0.110	0.181	0.594	0.437	0.256	0.125
Zusätzliche Parameter						
BFLPE	-	-	1.207***	0.231	0.803***	0.088
Effektstärke	-	-	1.260***	0.164	1.068***	0.123

Anmerkung: ICC = Intraklassen Korrelation; MSK = Mathematisches Selbstkonzept; Note = Mathematiknote im letzten Zeugnis; BFLPE = Big-Fish-Little-Pond-Effekt; *** p<.001; ** p<.01; * p<.05.

6 Diskussion

Ziel dieser Studie war es, mittels mehrebenanalytischer Modelle den Big-Fish-Little-Pond-Effekt (BFLPE; z. B. Marsh, Trautwein, Lüdtke & Köller, 2008) im Vergleich von Lernenden mit SPF-L in Förderschulklassen und Integrationsklassen in Regelschulen zueinander zu betrachten. Unsere Annahme, dass es sich bei integrativen Klassen für die Kinder mit SPF-L um einen subjektiv leistungsstärkeren Kontext handeln sollte, und somit der negative Effekt auf das akademische Selbstkonzept für Integrationskinder mit SPF-L in Regelschulen stärker ausgeprägt als für Lernende an Förderschulen mit dem Schwerpunkt Lernen, konnten wir in dieser Studie bestätigen. Es zeigte sich, dass der BFLPE in Integrationsklassen für die Kinder mit SPF-L sehr hoch ausfällt (tendenziell auch stärker als für die Kinder ohne SPF-L in diesen Klassen). Entgegen den Erwartungen zeigte sich in den für die Kinder mit SPF-L Förderschulklassen – also den segregierten Schulsettings – kein Einfluss des Kontextes auf die Einschätzung der eigenen Fähigkeit in Mathematik. Es scheint, als würde der Mechanismus sozialer Vergleiche (z. B. Dijkstra, Kuyper, Van der Werf, Buunk & Zee, 2008; Marsh, 2005) bezogen auf die erreichten Noten in Mathematik für die Schülerschaft an Förderschulen keine bedeutsame Rolle bei der Ausprägung ihres akademischen Selbstkonzeptes spielen, sondern sich auf Basis dieser Analysen vielmehr auf Grund der individuellen Leistung etablieren (vgl. Valentine, DuBois & Cooper, 2004). Alternativ erklärend können Befunde herangezogen werden, die zeigen, dass die Zusammenhänge zwischen individueller Leistung und Selbstkonzept bei leistungsschwächeren Kindern und Jugendlichen geringer ausfallen als bei leistungsstärkeren, welche Möller und Pohlmann (2010) über Motive des Selbstschutzes oder die weniger realistische (d. h. eher optimistische) Selbsteinschätzung der leistungsschwächeren Lernenden erklären. Es ist plausibel anzunehmen, dass auch bei Lernenden mit SPF-L in unterschiedlichen Schulsettings unterschiedliche psychologische Mechanismen aktiviert werden, wenn die eigene akademische Leistung eingeschätzt werden soll.

Im Vergleich dazu spielt für Kinder und Jugendliche mit SPF-L, die eine Integrationsklasse an einer Regelschule besuchen, der Kontext eine weitaus größere Rolle. Die stärkere Klassengemeinschaft dieser Schülergruppe führt zu einem geringer ausgeprägten akademischen Selbstkonzept für das Fach Mathematik (siehe hierzu Bos, Müller & Stubbe, 2010).

Eine Klassenkomposition mit schwächeren und stärkeren Lernenden wird als besondere Stärke der Integration gesehen, die das gemeinsame und voneinander Lernen fördern und darüber zu besseren Leistungen führen soll (vgl. auch Kocaj et al., 2014). Die vergleichsweise höheren Schulleistungen der Klassenkameradinnen und -kameraden und damit in vielen Fällen einhergehenden besseren Benotungen durch die Lehrkraft haben jedoch scheinbar auch negative Folgen für die Integra-

tionskinder mit SPF-L: Im Zuge eines sozialen Vergleiches der vermeintlich schwächeren Kinder und Jugendlichen mit SPF-L wird deren akademisches Selbstkonzept in Mathematik stark beeinflusst - es kommt zu einer geringeren Einschätzung der eigenen Fähigkeiten in Mathematik. Ein solcher Referenzgruppeneffekt könnte zu einem geringeren Interesse und somit zu einem verminderten Wissenszuwachs in der entsprechenden Domäne führen (vgl. Köller, Trautwein, Lüdtke & Baumert, 2006). Die Kompensation einer solchen Entwicklung bedarf eines sinnvollen pädagogischen Konzeptes, in dem der integrative Gedanke sich nicht nachteilig für die betroffenen Schülerinnen und Schüler erweist.

Im Falle der segregiert beschulten Schülerschaft ist in diesen Analysen kein Kontexteffekt nachweisbar. Die ohnehin hohe Ausprägung des akademischen Selbstkonzeptes ist nicht vom Notendurchschnitt der Klassengemeinschaft in Mathematik beeinflusst. Der Mechanismus des sozialen Vergleiches entfaltet sich entgegen unseren Erwartungen in den Förderschulklassen, anders als in Regelschulklassen, nicht. Jedoch ist für diese Schülergruppe an verschiedenen Stellen berichtet worden, dass mit Verlassen der Förderschule als „Schonraum" (Schumann, 2008) und einem damit verbundenen Bezugsgruppenwechsel (vgl. Seaton, Marsh & Craven, 2010) das Selbstkonzept signifikant einbricht (vgl. Gans, Kenny & Ghandy, 2003).

Die vorliegende Studie weist neben den vielfachen Vorteilen einer large-scale Erfassung mit umfangreichen Kontextmaßen in verschiedenen Schulsettings auch Einschränkungen auf, die die Einordnung der Ergebnisse betreffen. So wurden Leistungen sowohl auf Individual- als auch auf Kontextebene über Noten operationalisiert. Trautwein, Lüdtke, Köller und Baumert (2006) argumentieren zu dem Zusammenhang von Noten und Selbstkonzepten, dass aufgrund der Vergabe von Noten durch die Lehrpersonen über klassenspezifische soziale Vergleiche, die individuelle Leistung (gemessen über die Note) bereits auf die mittlere Leistung der Referenzgruppe (d. h. Klasse) zurückgeht und der BFLPE somit eine Konsequenz dieser Notenvergabepraxis ist (vgl. auch Trautwein, Lüdtke, Marsh, Köller & Baumert, 2006). Andererseits weisen Möller, Zimmermann und Köller (2014) daraufhin, dass Lernende ihr Selbstkonzept auf Noten und nicht auf Leistungen in standardisierten Testverfahren basieren. Es ist aus unserer Sicht davon auszugehen, dass für soziale Vergleiche die Sichtbarkeit von Noten im Klassenzimmer im Vergleich zu Kompetenzwerten, gemessen in standardisierten Tests, ausschlaggebender ist.

Literatur

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

Blossfeld, H.-P., Roßbach, H.-G. und Maurice, J. von (Hrsg.), (2011). Education as a Lifelong Process - The German National Educational Panel Study (NEPS). Zeitschrift für Erziehungswissenschaft: Sonderheft 14.

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.

Brunner, M., Keller, U., Dierendonck, Ch., Reichert, M., Ugen, S., Fischbach, A. & Martin, R. (2010). The structure of academic self-concepts revisited: The nested Marsh/Shavelson model. Journal of Educational Psychology, 102, 964-981.

Dietze, T. (2012). Zum Stand der sonderpädagogischen Förderung in Deutschland. Die Schulstatistik 2010/11. Zeitschrift für Heilpädagogik, 63, 26-31.

Dijkstra, P., Kuyper, H., Van der Werf, G., Buunk, A. P. & Zee, Y. G. (2008). Social comparison in the classroom: A review. Review of Educational Research, 78, 828–879.

Gans, A. M., Kenny, M. C. & Ghandy, D. L. (2003). Comparing the self-concept of students with and without learning disabilities. Journal of Child Psychology and Psychiatry, 42, 581-592.

Gebhardt, M., Oelkrug, K. & Tretter, T. (2013). Das mathematische Leistungsspektrum bei Schülerinnen und Schülern mit sonderpädagogischem Förderbedarf in der Sekundarstufe. Ein explorativer Querschnitt der fünften bis neunten Klassenstufe in Münchner Förderschulen. Empirische Sonderpädagogik, 2, 130-143.

Grünke, M. (2004). Lernbehinderung. In G. W. Lauth, M. Grünke & J. C. Brunstein (Hrsg.), Interventionen bei Lernstörungen. Förderung, Training und Therapie in der Praxis (S. 65–77). Göttingen: Hogrefe.

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, 217–240.

Huguet, P., Dumas, F., Marhs, H. W., Régner, I., Wheeler, L., Suls, J., Seaton, M. & Nezlek, J. (2009). Clarifying the Role of Social Comparison in the Big-Fish–Little-Pond Effect (BFLPE): An Integrative Study. Journal of Personality and Social Psychology, 97, 156-170.

Klauer, K. J. & Lauth, G. W. (1997). Lernbehinderungen und Leistungsschwierigkeiten bei Schülern. In F. E. Weinert (Hrsg.), Psychologie des Unterrichts und der Schule (S. 701-738). Göttingen: Hogrefe.

Kocaj, A., Kuhl, P., Kroth, A., 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, 165-191.

Köller, O. Trautwein, U., Lüdtke, O. & Baumert, J. (2006). Zum Zusammenspiel von schulischer Leistung, Selbstkonzept und Interesse in der gymnasialen Oberstufe. Zeitschrift für Pädagogische Psychologie, 20, 27–39.

Lüdtke, O., Köller, O., Artelt, C., Stanat, P. & Baumert, J. (2002). Eine Überprüfung von Modellen zur Genese akademischer Selbstkonzepte: Ergebnisse aus der PISA-Studie. Zeitschrift für Pädagogische Psychologie, 16, 151-164.

Marsh, H. W. (2005). Big-fish-little-pond effect on academic self-concept. Zeitschrift für Pädagogische Psychologie, 19, 119–127.

Marsh, H. W., Lüdtke, O., Robitzsch, A., Trautwein, U., Asparouhov, T., Muthén, B. & Nagengast, B. (2009). Doubly-latent models of school contextual effects: Integrating multilevel and structural equation approaches to control measurement and sampling errors. Multivariate Behavioral Research, 44, 764–802.

Marsh, H. W., Trautwein, U., Lüdtke, O. & Köller, O. (2008). Social comparison and big-fish-little-pond effects on self-concept and other self-belief constructs: Role of generalized and specific others. Journal of Educational Psychology, 100, 510-524.

Muthén, L. K. & Muthén, B. O. (2012). Mplus Version 7 [Computersoftware]. Los Angeles, CA.

Möller, J. (2013). Effekte inklusiver Beschulung aus empirischer Sicht. In J. Baumert, V. Masur, J. Möller, T. Riecke-Baulecke, H.-E. Tenorth & R. Werning (Hrsg.), Inklusion. Forschungsergebnisse und Perspektiven. Schulmanagement Handbuch, S. 15-37. München: Oldenbourg.

Möller, J. & Köller, O. (2004). Die Genese akademischer Selbstkonzepte: Effekte dimensionaler und sozialer Vergleiche. Psychologische Rundschau, 55, 19-27.

Möller, J. & Pohlmann, B. (2010). Achievement differences and self-concept differences: Stronger associations for above or below average students? British Journal of Educational Psychology, 80, 435-450.

Möller, J., Streblow, L. & Pohlmann, B. (2009). Leistung und Selbstkonzept bei lernbehinderten Schülern. Heilpädagogische Forschung, 28, 132-138.

Möller, J., Zimmermann, F. & Köller, O. (2014). The reciprocal internal/external frame of reference model using grades and test scores. British Journal of Educational Psychology, 84, 591-611.

Nusser, L., Carstensen C. H. & Artelt, C. (2015). Befragung von Schülerinnen und Schülern mit sonderpädagogischem Förderbedarf Lernen: Ergebnisse zur Messinvarianz. Empirische Sonderpädagogik, 2, 99-116.

Nusser, L., Heydrich, J., Carstensen, C. H., Artelt, C., & Weinert, S. (in Druck). Validity of Survey Data of Students with Special Educational Needs – Results from the National Educational Panel Study. In: H.-P. Blossfeld, J. v. Maurice, J. Skopek, & M. Bayer (Eds.), Methodological Issues of Longitudinal Surveys.

Pfahl, L. (2008). Die Legitimation der Sonderschule im Lernbehinderungsdiskurs in Deutschland im 20. Jahrhundert. Discussion Paper SP I 2008-504. Wissenschaftszentrum Berlin für Sozialforschung. Verfügbar unter: www.skylla.wz-berlin.de/pdf/2008/i08-504.pdf [24.02.2016].

Sass, D. A. (2011). Testing measurement invariance and comparing latent factor means within a confirmatory factor analysis framework. Journal of Psychoeducational Assessment, 29, 347–363.

Schumann, B. (2008). „Ich schäme mich ja so!": Eine wissenschaftliche Untersuchung zum Selbstkonzept von Schülern und Schülerinnen an der Sonderschule für Lernbehinderte. Heilpädagogik online, 01, 83-92.

Schurtz, I., Pfost, M., Nagengast, B. & Artelt, C. (2014). Impact of social and dimensional comparisons on student's mathematical and English subject-interest at the beginning of secondary school. Learning and Instruction, 34, 32-41.

Schwab, S. (2014). Haben sie wirklich ein anders Selbstkonzept? Ein empirischer Vergleich von Schülern mit und ohne sonderpädagogischen Förderbedarf im Bereich Lernen. Zeitschrift für Heilpädagogik, 65, 116-121.

Seaton, M., Marsh, H. W. & Craven, R. G. (2010). Big-fish-little-pond effect: Generalizability and moderation – two sides of the same coin. American Educational Research Journal, 47, 390-433.

Shavelson, R. J., Hubner, J. J. & Stanton, G. C. (1976). Self-concept: Validation of construct interpretations. Review of Educational Research, 46, 407–444.

Trautwein, U., Lüdtke, O., Marsh, H. W., Köller, O. & Baumert, J. (2006). Tracking, grading, and student motivation: Using group composition and status to predict self-concept and interest in ninth-grade mathematics. Journal of Educational Psychology, 98, 788-806.

Trautwein, U., Lüdtke, O., Köller, O. & Baumert, J. (2006). Self-esteem, academic self-concept, and achievement: How the learning environment moderates the dynamics of self-concept. Journal of Personality and Social Psychology, 90, 334-349.

Trautwein, U., Marsh, H. W., Nagy, G. (2009). Within-School Social Comparison: How Students Perceive the Standing of Their Class Predicts Academic Self-Concept. Journal of Educational Psychology, 101, 853-866.

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

Valentine, J. C., DuBois, D. L. & Cooper, H. (2004). The relation between self-beliefs and academic achievement: A meta-analytic review. Educational Psychologist, 39, 111–133.

Venetz, M., Tarnutzer, R., Zurbriggen, C. & Sempert, W. (2010). Die Qualität des Erlebens von Lernenden in integrativen und separativen Schulformen: Eine Untersuchung mit der Experience Sampling Method (ESM). Zürich, Switzerland: University of Applied Sciences of Special Needs Education.

Wohlkinger, F., Ditton, H., von Maurice, J., Haugwitz, M. & Blossfeld, H.-P. (2011). Motivational concepts and personality aspects across the life course. In H.-P. Blossfeld, H.-G. Roßbach & J. von Maurice (Hrsg.), The National Educational Panel Study: need, main features, and research potential (S. 155–168). Wiesbaden: VS Verlag für Sozialwissenschaften.

There's plenty more fish in the sea. Academic self-concept of students with special educational needs in the area of learning in inclusive and segregated school settings

The empirically proven Big-Fish-Little-Pond-Effect (BFLPE) has shown that the academic self-concept of students is not only influenced by their own achievement, but is also negatively affected by the average achievement of their class. Against the background of empirical studies showing that the academic self-concept is relevant for the development of competencies and that it develops in a process of social comparison (Möller & Köller, 2004), this article examines the BFLPE for students at special schools (n = 587 in 91 classes) in a multilevel analysis. The authors anticipate that the mechanism of social comparison would also unfold in segregated school settings and that a negative effect of the average achievement-level within class on the individual academic self-concept would be confirmed. Moreover, the BFLPE-model was explored in comparison to students enrolled in integrative school settings (n = 148 in 103 classes). Since integrative classes for students with special educational needs in the area of learning form a comparably more powerful context, we assumed that the negative effect on the academic self-concept would be stronger for integrated students at general education schools than for students at special schools. Multilevel analyses showed that the average achievement level (i.e., the grade average of each class) controlled for individual achievements did not have an effect on the academic self-concept of students at special schools. For integrated classes a distinct context effect in the expected direction was found.

Keywords: academic self-concept – Big-Fish-Little-Pond-Effect – inclusion – special educational needs

Autorinnen

Lena Nusser, Otto-Friedrich-Universität Bamberg,
Dr. Ilka Wolter, Leibniz-Institut für Bildungsverläufe e.V. (LIfBi), Bamberg.
Korrespondenz an: lena.nusser@uni-bamberg.de