Decristan, Jasmin; Naumann, Alexander; Fauth, Benjamin; Rieser, Svenja; Büttner, Gerhard; Klieme, Eckhard

Heterogenität von Schülerleistungen in der Grundschule. Bedeutung unterschiedlicher Leistungsindikatoren und Bedingungsfaktoren für die Einschätzung durch Lehrkräfte

formal und inhaltlich überarbeitete Version der Originalveröffentlichung in:
formally and content revised edition of the original source in:
Zeitschrift für Entwicklungspsychologie und pädagogische Psychologie 46 (2014) 4, S. 181-190

Decristian, J., Naumann, A., Fauth, B., Rieser, S., Büttner, G. & Klieme, E. (2014). Heterogenität von Schülerleistungen in der Grundschule. Bedeutung unterschiedlicher Leistungsindikatoren und Bedingungsfaktoren für die Einschätzung durch Lehrkräfte. Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 46 (4). doi: 10.1026/0049-8637/a000115

Heterogenität von Schülerleistungen in der Grundschule: Bedeutung unterschiedlicher Leistungsindikatoren und Bedingungsfaktoren für die Einschätzung durch Lehrkräfte

Jasmin Decristan[1], Alexander Naumann[2], Benjamin Caspar Fauth[3], Svenja Rieser[3], Gerhard Büttner[3] und Eckhard Klieme[2]

[1]Deutsches Institut für Internationale Pädagogische Forschung, BIQUA

[2]Deutsches Institut für Internationale Pädagogische Forschung, IDeA-Forschungszentrum

[3]Goethe-Universität Frankfurt

Zusammenfassung. Das Thema Heterogenität hat in den letzten Jahren eine zunehmende Präsenz sowohl in empirischer Schul- und Unterrichtsforschung als auch in Schulpraxis und bildungspolitischen Diskussionen erfahren. In diesem Aufsatz wird untersucht, welche Zusammenhänge zwischen verschiedenen objektiven Indikatoren der Heterogenität auf Klassenebene und einer globalen Lehrereinschätzung von Leistungsheterogenität bestehen. Die Ergebnisse zeigen zunächst, dass im Grundschulbereich das Ausmaß von Heterogenität (gemessen an der klasseninternen Streuung) bei verschiedenen Leistungsindikatoren (kognitive Grundfähigkeiten, Leseverständnis, naturwissenschaftliche Kompetenz) unkorreliert ist, so dass sich nicht einheitlich von „der“ Leistungsheterogenität sprechen lässt. Die Einschätzung von 49 Lehrkräften zur Leistungsheterogenität ihrer Klasse stimmte mit objektiven Leistungsstreuungen im Leseverständnis, in naturwissenschaftlicher Kompetenz sowie in einem kombinierten Leistungsindex gut überein. Bei Nicht-Passung tendierten die Lehrkräfte dazu, die Leistungsheterogenität eher zu überschätzen; hierfür erwies sich der Anteil an Kindern mit Migrationshintergrund als bedeutsam. Zusammengenommen bietet dieser Beitrag eine weitere empirische Grundlage für den gegenwärtigen Heterogenitätsdiskurs.

Schlüsselwörter: Heterogenität, Grundschulleistungen, Leseleistungen, naturwissenschaftliche Kompetenz, kognitive Grundfähigkeiten

Abstract. The heterogeneity of students is a prominent topic in empirical research and in educational practice and discourse. The present study uses three achievement indicators

Die Erstellung dieses Beitrags wurde gefördert durch die LOEWE-Initiative der Hessischen Landesregierung.

(cognitive abilities, reading comprehension, and science competency) to describe the quantity of individual differences within primary school classes. Furthermore, this data is combined with ratings of 49 teachers on the heterogeneity in achievement of their entire classes. The results demonstrate that the intra-class standard deviations of the three achievement indicators do not correlate with each other. The teacher's ratings of heterogeneity in achievement fit the student data regarding reading comprehension and scientific literacy. However, teachers tend to give rather high ratings of heterogeneity in class, and particularly variables that are connected with language (intra-class standard deviation of reading comprehension, percentage of children from immigrant families) are connected with these teacher ratings. In summary, these results provide a further empirical basis for the topic of heterogeneity.

Key words: Heterogeneity of achievement, primary school students, reading comprehension, cognitive abilities, scientific literacy

Das Thema Heterogenität hat in Deutschland in den letzten Jahren eine zunehmende Präsenz sowohl in der empirischen Schul- und Unterrichtsforschung (z. B. Bellin, 2009; Gröhlich, Scharenberg & Bos, 2009; Künsting, Post, Greb, Faust & Lipowsky, 2010) als auch in Schulpraxis und bildungspolitischen Diskussionen erfahren (z. B. Reh, 2005; Solzbacher, 2008; Terhart, 2006). Leitend sind hier auch Fragen zum tatsächlichen und wahrgenommenen Ausmaß von Heterogenität in Schulklassen. Entwicklungs- und Pädagogische Psychologie bieten zu diesem Thema wichtige Grundlagen. Mit psychologischen Testinstrumenten lassen sich interindividuelle Unterschiede in den Lernvoraussetzungen präzise beschreiben. Studien zur diagnostischen Kompetenz von Lehrkräften liefern Hinweise zur Güte von Urteilen zu interindividuellen Leistungsunterschieden. Der vorliegende Beitrag greift die Heterogenitäts-Debatte in der pädagogisch-psychologischen Forschung auf. Anhand von Schülerdaten und Lehrerurteilen wird untersucht, inwieweit verschiedene objektive und subjektive Kriterien von Leistungsheterogenität übereinstimmen. Damit soll zu der Frage beigetragen werden, inwieweit Forschung und Praxis dasselbe meinen, wenn sie von Heterogenität sprechen.

Spezifizierung des Begriffs (Leistungs-)Heterogenität

Heterogenität wird oftmals als ein feststehender und vergleichsweise undifferenzierter Begriff („die Heterogenität“) verwendet, der sich auf Verschiedenheit bezieht. Um den Gegenstand der Heterogenität zu spezifizieren, wurden verschiedene Klassifikationssysteme aufgestellt (z. B. Heinzel, 2008). Kernelement dieser Systeme ist das Leistungsvermögen, das in der Psychologie ursprünglich im Kontext der Intelligenz- und Leistungsdiagnostik näher betrachtet wurde. Traditionell wurde dazu zwischen einem generellen („G“-)Faktor der Intelligenz, den kognitiven Grundfähigkeiten, sowie untergeordneten, spezifischen Faktoren unterschieden. Korrelations- und Faktorenanalysen belegen diese Struktur (z. B. Gustafsson & Balke, 1993). Entsprechend werden in der pädagogisch-psychologischen Diagnostik sowie in Studien zu Heterogenität verschiedene Leistungsindikatoren verwendet.

Methodische Operationalisierung von Leistungsheterogenität von Schülerinnen und Schülern

Bei größeren Erhebungen im Schulkontext lassen sich drei „Betrachtungsweisen“ von Heterogenität unterscheiden (vgl. Kluczniok, Große & Roßbach, 2011): Heterogenität (a) in der Gesamtstichprobe, (b) innerhalb von Analyseeinheiten (z. B. Schulen, Klassen) und (c) zwischen Analyseeinheiten. Je nach Betrachtungsweise können die Werte deutlich variieren, wobei naturgemäß die Leistungsheterogenität in Gesamtstichproben am deutlichsten ausfällt. Wird Heterogenität im Sinne von interindividuellen Differenzen (z. B. Unterschiede zwischen Personen in der Stichprobe oder in einer Klasse) untersucht, findet die Analyse auf Ebene I (individuelle Ebene) statt. Wird Heterogenität im Sinne von Unterschieden zwischen aggregierten Einheiten (z. B. Klassen, Schulen) verstanden, wird die den individuellen Werten (Ebene I) hierarchisch übergeordnete, aggregierte Ebene II fokussiert. Ausprägung, Interpretation und Verwendung von Variabilitätsmaßen für Testleistungen können sich somit je nach betrachteter Ebene unterscheiden.

Die Spannbreite gibt als Differenz zwischen dem größten und kleinsten Wert den maximalen Werte-Unterschied an. Für metrische Variablen ist die Standardabweichung (*SD*) derjenige Verteilungsparameter, der das Ausmaß an Unterschiedlichkeit in einer Population bzw. Stichprobe beschreibt. Bei normalverteilten Daten streuen auf Ebene I zwei Drittel der individuellen Merkmalsausprägungen im Intervall von ±1 *SD* um den Gesamtmittelwert *M*; auf Ebene II liegen zwei Drittel der Gruppenmittelwerte (z. B. von Klassen) im Bereich $(M \pm SD)/\sqrt{n}$, mit n als Größe der Analyseeinheiten (z. B. Anzahl an Kindern pro Klasse). Welcher Anteil der Gesamtvarianz (Varianz aller individuellen Merkmalsausprägungen) wiederum auf die Variation zwischen Analyseeinheiten zurückgeht, lässt sich anhand der Intra-Klassenkorrelation (ICC) beschreiben. Ob die Variation innerhalb der Analyseeinheiten gleich groß ist, ob also das Ausmaß an „within“-Heterogenität über die Gruppen hinweg („between“) homogen ist, lässt sich schließlich über Tests zur Varianzhomogenität prüfen.

Empirische Befunde zur Leistungsheterogenität in der Grundschule

Vergleichende Schulleistungsstudien geben Aufschluss über interindividuelle Unterschiede in den Leistungen von definierten Populationen. Für die repräsentative Gesamtstichprobe deutscher Grundschulkinder belegte die IGLU-Studie 2011 (Bos, Tarelli, Bremerich-Vos & Schwippert, 2012) eine Standardabweichung von 66 Punkten in den Leseleistungen (M = 541 Punkte); in der TIMS-Studie 2011 (Mullis, Martin, Foy & Arora, 2012) resultierten Standardabweichungen von 61 Punkten in den Mathematikleistungen bzw. von 70 Punkten für die Leistungen in den Naturwissenschaften (jeweilige M = 528 Punkte). Die Tests sind so skaliert, dass der internationale Mittelwert 500 und die Standardabweichung 100 beträgt. Die Leistungsstreuungen in deutschen Grundschulen sind somit im internationalen Vergleich als unterdurchschnittlich zu bewerten.

Die Frage nach einer *Zu- oder Abnahme* von Leistungsheterogenität lässt sich aus gesellschaftlicher (Veränderung als Funktion der Zeit bei konstantem Lebensalter, d.h. Differenzen zwischen Kohorten betreffend) und entwicklungspsychologischer Perspektive (Veränderung als Funktion des Lebensalters) betrachten. Empirische Ergebnisse liegen jeweils bezogen auf Gesamtpopulationen vor. Aus gesellschaftlicher Perspektive deuten Ergebnisse querschnittlicher Vergleichsstudien im Laufe des letzten Jahrzehnts auf eine gewisse Konstanz der Leistungsheterogenität am Ende der Grundschulzeit hin (z. B. Bos et al., 2012). Entwicklungspsychologisch zeigte Ditton (2010) in einer Längsschnittstudie in der Grundschule eine abnehmende Streuung in den Mathematikleistungen und einen schwankenden Verlauf in den Streuungen der Rechtschreib- und Leseleistungen. Für einen zusammengefassten Leistungsindex (mittlere Mathematik- und Deutschleistung) verringerte sich die Standardabweichung über die Grundschulzeit hinweg.

Einschätzungen von Lehrkräften zu Schülerleistungen und Leistungsheterogenität

Für Aussagen zum Ausmaß von Leistungsheterogenität werden neben Leistungstests in der Regel auch Urteile von Lehrkräften herangezogen. Ignorieren Lehrkräfte Leistungsunterschiede, lässt sich davon ausgehen, dass Unterricht nicht optimal an den jeweiligen Bedürfnissen der Kinder ansetzt und ein produktiver Umgang mit Heterogenität kaum gewährleistet werden kann. Werden Leistungsunterschiede dramatisiert, kann dies zu verstärkter Aufmerksamkeit für diese Thematik führen und die Initiierung individueller Fördermaßnahmen voranbringen. Dennoch birgt es die Gefahr, dass Lehrkräfte sich den Anforderungen im Berufsalltag nicht mehr hinreichend gewachsen fühlen (z. B. Reh, 2005).

Im Kontext von Forschung zu „diagnostischer Kompetenz“ wird untersucht, wie gut es Lehrkräften gelingt, individuelle Leistungsstände sowie das Leistungsniveau und interindividuelle Leistungsunterschiede von Kindern in Klassen einzuschätzen (z. B. Lorenz & Artelt, 2009). Die Güte der Beurteilung interindividueller Leistungsunterschiede wird über die Differenzierungs- bzw. Streuungskomponente bestimmt (vgl. Helmke & Schrader, 1987), die als Quotient aus der Streuung von Lehrerurteilen zu individuellen Schülerleistungen und der tatsächlichen Schülerleistungsstreuung bestimmt wird (< 1 = Unterschätzung, = 1 perfekte Übereinstimmung, > 1 Überschätzung von Leistungsheterogenität). Die empirische Befundlage dazu ist uneinheitlich. Zwei Studien mit Sekundarschullehrkräften belegten gute Übereinstimmungen zwischen Lehrerurteilen und Leistungsheterogenität (Brunner, Anders, Hachfeld & Krauss, 2011; Karing, Matthäi & Artelt, 2011). Schrader (1989) und Seeber (2009) verwiesen auf eine Überschätzung, andere auf eine Unterschätzung der Leistungsstreuungen (Spinath, 2005; Südkamp & Möller, 2009).

Lehrerurteile im Kontext diagnostischer Kompetenz beinhalten unterschiedliche Spezifitätsgrade (globale, fach- und aufgabenspezifische Leistungseinschätzungen). Studien zeigen, dass globalere individuelle Leistungen (z. B. in einem Schulfach) in geringerem Ausmaß zutreffend eingeschätzt werden können, während sich für zentrale fachspezifische

Leistungsbereiche (Wortschatz, Textverstehen und Arithmetik) hohe und stabile Übereinstimmungen mit Lehrerurteilen zeigen (z. B. Hoge & Coladarci, 1989; Lorenz & Artelt, 2009). Karst (2012) zeigte zusätzlich, dass die Art der Lehrerurteile auch für die Differenzierungskomponente relevant ist: Schätzten Grundschullehrerinnen individuelle Lösungshäufigkeiten für einzelne Aufgaben ein, wurde die Leistungsheterogenität in der Klasse tendenziell unterschätzt. Wenn konkrete Erfolge bzw. Misserfolge in bestimmten Aufgaben zu prognostizieren waren, wurde das Urteil angemessener.

In der hier vorliegenden Untersuchung geht es jedoch nicht um die Güte von Lehrerurteilen zu einzelnen Schülern, sondern um globale Urteile zur Leistungsheterogenität in der unterrichteten Klasse. Abweichend von den bisher referierten wissenschaftlichen Arbeiten zur „diagnostischen Kompetenz“ stehen in der bildungspolitischen Debatte sowie in Lehrerbefragungen zum Thema Heterogenität keine individuellem Fähigkeitsurteile im Fokus, sondern vielmehr absolute oder relative Einschätzungen zur (Leistungs-)Heterogenität von Lerngruppen (z. B. Reh, 2005; Solzbacher, 2008). In der COACTIV-Studie (Kunter et al., 2007) wurden im Jahr 2003/04 Lehrkräfte gebeten, solche Heterogenitätseinschätzungen ganzer Klassen („Leistungsstreuung der Klasse im Vergleich zu einer durchschnittlichen Klasse derselben Schulform für das Fach Mathematik“) anhand einer fünfstufigen Skala (von deutlich unter- bis deutlich überdurchschnittlich) abzugeben. Folglich handelt es sich um Urteile auf einer den individuellen Werten hierarchisch übergeordneten Analyseebene, der Heterogenität zwischen Klassen. Die Hälfte der Sekundarschullehrkräfte antworteten mit „durchschnittlich“, die beiden Extrembereiche waren kaum vertreten (Brunner et al., 2011). Insbesondere an der Grundschule als gemeinsamer Regelschule ist jedoch von relativ großen Leistungsunterschieden innerhalb von Klassen auszugehen. Zudem legen bildungspolitische Diskussionen nahe, dass die Heterogenität der Schülerinnen und Schüler zunimmt. So wird die „Fähigkeit zum konstruktiven Umgang mit der wachsenden Heterogenität der

Grundschüler“ als wichtige Anforderung im Berufsalltag gesehen (Terhart, 2006, S. 234). Unklar ist bislang jedoch, inwieweit diese globale Lehrereinschätzung zur Heterogenität mit empirisch erhobenen Leistungsunterschieden von Grundschulkindern übereinstimmt.

Zusammenhänge zwischen Leistungsheterogenität und Klassenkontextmerkmalen

Da die Heterogenitätsdebatte breit geführt wird, können sich entsprechende Aussagen auf die Leistungsfähigkeit von Schülerinnen und Schülern, aber auch auf weitere Heterogenitätsmerkmale, wie sozio-ökonomischer Status (SES) und Ethnizität/Kultur, beziehen. Für die Klasse als Analyseeinheit lassen sich die angeführten Merkmale als Klassenkontextmerkmale beschreiben (Marsh et al., 2012). Studien aus dem Bereich diagnostischer Kompetenz konnten belegen, dass besonders Klassenkontextmerkmale die Güte von Leistungseinschätzungen durch Lehrkräfte systematisch beeinflussen: In leistungs- und sozioökonomisch schwachen Klassen sowie Klassen mit hohem Migrationsanteil unterschätzten Lehrkräfte die individuellen Schülerfähigkeiten (z. B. Hauser-Cram, Sirin & Stipek, 2003; Ready & Wright, 2011). Lehrermerkmale waren hingegen nicht von Bedeutung. Dabei ist von einer gewissen Abhängigkeit bzw. Überschneidung dieser sogenannten Heterogenitätsdimensionen (vgl. Heinzel, 2008) auszugehen und entsprechend zu vermuten, dass Urteile zu Leistungsheterogenität abhängig von weiteren Heterogenitätsmerkmalen, wie dem Migrationshintergrund der Kinder, getroffen werden. Auch die Klassengröße wird im Kontext von Heterogenität genannt und von Lehrkräften als Herausforderung beim Umgang mit Heterogenität angesehen (Winheller, Müller, Hüpping, Rendtorff & Büker, 2012).

Fragestellungen und Hypothesen

In der vorliegenden Arbeit werden die in der bildungspolitischen und pädagogischen Debatte oftmals gegebenen Äußerungen zur Heterogenität von Schülerinnen und Schülern aufgegriffen und mit Leistungstestdaten in Beziehung gesetzt. Ziel ist es, eine breitere empirische Grundlage für den gegenwärtigen Heterogenitätsdiskurs zu bieten. So ist bislang

nicht hinreichend erforscht, inwieweit verschiedene (globale und fachspezifische) Leistungsindikatoren zu vergleichbaren Rangreihen der Klassen hinsichtlich des Ausmaßes an Heterogenität führen. Daher werden folgende Fragestellungen und Hypothesen untersucht:

1. Welche Zusammenhänge zeigen sich zwischen klasseninternen Streuungen verschiedener Leistungsindikatoren von Grundschulkindern?

2. Wie gut gelingt es Grundschullehrkräften die Leistungsheterogenität ihrer Klasse zu beurteilen?

2.1 Ausgehend von Befunden zur diagnostischen Kompetenz von Lehrkräften wird eine moderate Übereinstimmung zwischen Einschätzungen von Grundschullehrkräften zu Leistungsheterogenität gesamter Klassen und fachspezifischen Leistungsstreuungen sowie eine geringe Übereinstimmung mit globalen Leistungsteststreuungen in Klassen erwartet.

2.2 Bildungspolitische und pädagogische Diskussionen legen eine zunehmende Heterogenität der Schülerinnen und Schüler nahe, während Ergebnisse aus Schulleistungsstudien auf eine konstante oder abnehmende Leistungsheterogenität in der Grundschule hindeuten. Erwartet wird daher, dass im Falle einer Nicht-Übereinstimmung von Lehrerurteil und Schülerdaten das absolute Ausmaß der Leistungsheterogenität einer Klasse eher über- als unterschätzt wird.

3. Welche Klassenkontextmerkmale tragen dazu bei, dass Grundschullehrkräfte ihre Klasse als überdurchschnittlich leistungsheterogen einschätzen? Erwartet wird, dass Klassenkontextmerkmale, die aus weiteren Heterogenitätsmerkmalen (SES, Ethnizität/Kultur) gebildet wurden, sowie die Klassengröße dazu beitragen, dass Lehrkräfte eine vergleichsweise große Leistungsheterogenität einer Klasse angeben.

Methode

Stichprobe

Die Daten stammen aus einer umfangreicheren Interventionsstudie an 54 hessischen Grundschulen zu Wirkungen verschiedener Unterrichtsmethoden und beziehen sich auf zwei

jeweils 90-minütige Erhebungen zum Anfang des Schuljahres 2010/11, vor Beginn der Unterrichtsintervention. Zwischen beiden Erhebungen lagen im Mittel 3.5 Tage (*SD* = 5.1). Die Teilnahme war für Lehrkräfte sowie Schülerinnen und Schüler freiwillig. 979 Kinder aus 54 Klassen der dritten Jahrgangstufe nahmen an beiden Tagen teil (*N* = 999 Tag 1; *N* = 1021 Tag 2). Das mittlere Alter der Schülerinnen und Schüler betrug 8.78 Jahre (*SD* = 0.48), etwa die Hälfte waren Mädchen (49 %). Etwa ein Drittel (38 %) der befragten Kinder gab an, dass ihr Vater und/oder ihre Mutter einen Migrationshintergrund haben. Von 49 der 54 Lehrkräfte lagen Einschätzungen zur Leistungsheterogenität ihrer jeweiligen Klasse vor. Von diesen 49 Lehrkräften waren 92 % weiblich. Die Lehrkräfte waren im Mittel 43 Jahre alt (*SD* = 9.19) und durchschnittlich 14.58 Jahre im Schuldienst (*SD* = 8.39). Sie unterrichteten zum Zeitpunkt der Befragung eine dritte Klasse und waren meist(88 %) Klassenlehrerinnen oder -lehrer der teilnehmenden Klassen. 75 % der Lehrkräfte hatten Deutsch und 19 % ein naturwissenschaftliches Fach im Rahmen ihrer Lehramtsausbildung studiert.

Instrumente und Durchführung

Die kognitive Grundfähigkeit der Kinder wurden über den CFT 20-R (Weiß, 2006; 56 Items, Cronbachs α: .72) erfasst. Zehn Prozent der Gesamtvariation der CFT-Leistungen ging auf die Variation zwischen Klassen zurück (ICC = .10). Das Leseverständnis wurde über den ELFE 1-6 (Lenhard & Schneider, 2006; 120 Items, drei Testteile, Cronbachs α pro Testteil > .91, ICC = .08), die naturwissenschaftliche Kompetenz über einen aus TIMSS (Martin, Mullis & Foy, 2008) adaptierten und erweiterten Test erhoben, der Wissen, Schlussfolgern und Anwenden in den Inhaltsbereichen Chemie, Physik und Erdkunde erforderte. Die zwölf Aufgaben wurden mittels des dichotomen Raschmodells skaliert; es wurden Weighted Likelihood Estimates (WLE, Warm, 1989) als Personenparameter geschätzt (EAP/PV Reliabilität = .70, ICC = .14). Sozioökonomischer Status (SES) und Migrationshintergrund von Kindern wurden per Schülerfragebogen erhoben. Wenn das Kind angab, dass beide

Elternteile in Deutschland geboren worden sind, wurde kein Migrationshintergrund kodiert (= 0), andernfalls wurde eine 1 für Migrationshintergrund vergeben. Als SES-Indikator wurde die Frage nach der Anzahl an Büchern im Haushalt gestellt und die in Schulleistungsstudien gängige Dichotomisierung der Antwortkategorien vorgenommen (> 100 Bücher (hoher SES) = 0, ≤ 100 Bücher (niedriger SES) = 1; vgl. Bos et al., 2005). Die Anzahl der Kinder in den Klassen wurde über die Klassenlisten zu Beginn des Schuljahres ermittelt. Die Erhebungen fanden im Klassenverbund mit geschulten Testleiterinnen und Testleitern statt. Instruktionen und Items wurden laut vorgelesen. Beispielaufgaben zu CFT und ELFE sowie der Test zur naturwissenschaftlichen Kompetenz wurden zusätzlich an die Wand projiziert.

Die Lehrkräfte wurden in Anlehnung an die COACTIV-Studie (Kunter et al., 2007) nach einem Urteil zur gesamten Klasse gebeten. Die Instruktion lautete: „*Bitte vergleichen Sie Ihre aktuelle Klasse der Jahrgangsstufe 3, mit der Sie am Projekt teilnehmen, mit einer durchschnittlichen Klasse dieser Jahrgangsstufe. Ihre aktuelle Klasse 3 ist im Hinblick auf...*“. Es folgten mehrere Unterpunkte, die jeweils anhand eines dreistufigen Antwortformats beantwortet werden sollten (–1 = *geringer*, 0 = *vergleichbar*, 1 = *höher/größer*). Der hier im Fokus stehende Unterpunkt lautete „*Leistungsunterschiede zwischen den Schülern*“. Den Lehrkräften wurde dabei bewusst nicht vorgegeben, auf welche Leistungsmerkmale sie sich beziehen sollten oder was als „gering“ usw. zu gelten habe.

Datenaufbereitung und -analysen

Den Ausgangspunkt der Analyse von Leistungsdaten (Tabelle 1) bildeten die individuellen Werte der drei Leistungstests (Summen- bzw. WLE-Werte), die standardisiert wurden ($M = 0$, $SD = 1$). Ergänzend wurde ein kombinierter Leistungsindex als Mittelwert aus den drei individuellen standardisierten Leistungswerten gebildet und erneut standardisiert (ICC = .15; Korrelationen zwischen individuellen Leistungstestwerten zwischen $r = .28$ und $r = .38$). Für die Analyse auf Klassenebene wurde für die individuellen standardisierten Werte

in WinBUGS 1.4 (Spiegelhalter, Thomas, Best & Lunn, 2003) ein Zwei-Ebenen-Nullmodell mit heterogenen klasseninternen Varianzen spezifiziert, um Kennwerte auf Klassenebene latent zu schätzen und somit Messfehler auf Stichproben-Ebene zu berücksichtigen (vgl. Marsh et al., 2012). Für jede Klasse wurden Mittelwert und Standardabweichung als Heterogenitätsmaß latent geschätzt. Als Kontextmerkmale wurden der prozentuale Anteil an Kindern mit Migrationshintergrund bzw. niedrigem SES in Klassen berechnet. Um die Anteile an Über- oder Unterschätzung der Leistungsstreuung durch Lehrkräfte zu bestimmen, wurden die klasseninternen Standardabweichungen der Testwerte standardisiert ($M = 0$, $SD = 1$ auf Klassenebene) und drei Kategorien zugeteilt (kleiner als –1, d. h. mehr als eine *SD* unter dem Mittelwert = homogen; zwischen –1 und +1 = durchschnittlich; größer als + 1, d. h. mehr als eine *SD* über dem Mittelwert = heterogen). Diese Dreiteilung wurde mit dem dreistufigen Antwortformat der Lehrereinschätzungen verglichen. Die Konsistenz der Ergebnisse wurde mit zwei weiteren Einteilungskriterien der Schülerdaten geprüft, nämlich einer gleichhäufigen Zuteilung (je 33 %) der Klassen zu den drei Kategorien sowie das Herausschneiden von „Unsicherheits-"Bereichen (kleiner als –1.2 *SD*, d. h. mehr als 1.2 *SD* unter dem mittleren Wert = homogen; zwischen –0.8 und + 0.8 *SD* = durchschnittlich; größer als +1.2 *SD* = heterogen; Intervalle von –1.2 bis –0.8 *SD* sowie von +0.8 bis +1.2 *SD* gelten als indifferent und bleiben für die Analyse unberücksichtigt).

Selektion und fehlende Werte

In den 54 Klassen waren zu Beginn des Schuljahres laut Klassenlisten 1.113 Kinder. Von den Eltern lagen Einverständniserklärungen für 1.070 Kinder vor (96 %). Der Anteil fehlender Werte lag bei 7–8 % (Testleistungen) und 15 % (Migrationshintergrund). Ein wesentlicher Teil war darauf zurückzuführen, dass Kinder die Erhebungen zu Beginn des Schuljahres verpasst hatten (7 % Tag 1, 5 % Tag 2; z. B. wegen Krankheit oder Schul-/ Klassenwechsel). Von den 54 Lehrkräften gaben 5 keine Antwort zur Leistungsheterogenität

der Klasse. Die 5 Klassen unterschieden sich nicht von den anderen 49 Klassen hinsichtlich der betrachteten Heterogenitäts- und Klassenkontextmerkmale (Effektstärken $d < \pm 0.06$).

Ergebnisse

Deskriptive Befunde

Tabelle 1 enthält Daten zur Beschreibung von Leistungsheterogenität nach den drei Betrachtungsweisen (a) Gesamtstichprobe, (b) zwischen und (c) innerhalb von Klassen. In allen drei Leistungsbereichen zeigten sich deutliche interindividuelle Unterschiede. Für die kognitiven Grundfähigkeiten und das Leseverständnis lagen die Streuungen der hier einbezogenen hessenspezifischen Stichprobe deskriptiv geringfügig unter den Streuungen der bundesweiten Normierungsstichproben (vgl. Lenhard & Schneider, 2006; Weiß, 2006) und somit im erwartbaren Bereich. Bezogen auf die Gesamtstichprobe (erster Tabellen-Abschnitt) variierten niedrigste und höchste individuelle Werte zwischen –3.27 Standardabweichungen (CFT) und 3.70 Standardabweichungen (ELFE) um den Mittelwert der Gesamtstichprobe. Auch gab es deutliche Unterschiede zwischen Klassen im mittleren Leistungsniveau (zweiter Tabellen-Abschnitt). Die mittleren Streuungen in Klassen (dritter Tabellen-Abschnitt) geben die durchschnittlichen interindividuellen Unterschiede von Kindern in Klassen an; diese sind erwartungsgemäß geringer als die Variation in der Gesamtstichprobe. Die Spannbreite demonstriert deutliche Unterschiede in der Leistungsheterogenität zwischen den Klassen. Im vierten Tabellen-Abschnitt stehen klassenspezifische Variationen in sozio-demographischen (Anteil an Kindern mit Migrationshintergrund bzw. geringem SES) und dem strukturell-organisatorischen Kontextmerkmal (Anzahl Kinder in der Klasse).

Forschungsfrage 1

Zuerst wurde der Frage nachgegangen, inwieweit die klasseninternen Streuungen verschiedener Leistungsmaße (kognitive Grundfähigkeiten, Lesefähigkeit und naturwissenschaftliche Kompetenz) korrelieren und sich somit konsistent von „der“

Leistungsheterogenität einer Klasse sprechen lässt. Für Streuungen im Leseverständnis und in naturwissenschaftlicher Kompetenz bestätigten sich Normalverteilungen (Shapiro-Wilk-Test für kleine Stichproben: $W(54) = .98$, $p > .05$), für die kognitiven Grundfähigkeiten resultierte eine signifikante Abweichung ($W(54) = .95$, $p < .05$). Folglich wurden in Hinblick auf die Forschungsfrage verschiedene Korrelationskoeffizienten herangezogen, die für das jeweilige Skalenniveau der Variablen adäquat sind. Im Ergebnis fanden sich weder zwischen den jeweiligen klasseninternen Streuungen in kognitiven Grundfähigkeiten und im Leseverständnis (Spearmans $\rho = .15$, $p > .05$), noch zwischen den Streuungen in kognitiven Grundfähigkeiten und in naturwissenschaftlicher Kompetenz ($\rho = –.11$, $p > .05$) oder denjenigen in Leseverständnis und in naturwissenschaftlicher Kompetenz ($r = .07$, $p > .05$) signifikante Übereinstimmungen. Verschiedene Leistungsindikatoren kamen somit zu unterschiedlichen Rangreihen der Klassen zum Ausmaß von Leistungsheterogenität.

Hypothesen 2.1 und 2.2

Als nächstes wurde die Übereinstimmung zwischen Schülerdaten und Einschätzungen von Lehrkräften zur Leistungsheterogenität gesamter Klassen geprüft. Zunächst zeigte sich, dass die Hälfte der Lehrkräfte ($n = 25$, 51 %) für die jeweilige Klasse eine durchschnittliche Leistungsheterogenität angaben, ein kleiner Anteil eine unterdurchschnittliche ($n = 5$, 10 %) und ein größerer Anteil ($n = 19$, 39 %) eine überdurchschnittliche Heterogenität. Diese Einschätzungen folgten keiner Normalverteilung ($W(49) = .77$, $p < .01$).

Globale Lehrereinschätzungen und Schülerdaten zur Leistungsheterogenität korrelierten auf Klassenebene erwartungsgemäß moderat, wenn sich die Schülerdaten auf Leseverständnis ($\rho = .27$, $p < .05$) oder naturwissenschaftliche Kompetenz bezogen ($\rho = .25$, $p < .05$). Keine Übereinstimmung gab es, wenn kognitive Grundfähigkeit als Leistungsindikator verwendet wurde ($\rho = .01$ $p > .05$). Wider Erwarten fand sich auch eine moderate Übereinstimmung zwischen Lehrerurteil und dem kombinierten Leistungsindex ($\rho = .42$, $p < .01$).

Um Lehrer- und Schülerdaten klassenweise und direkt aufeinander zu beziehen, wurden die klasseninternen Streuungen anhand verschiedener Einteilungskriterien drei Kategorien (homogen, durchschnittlich, heterogen) zugewiesen und mit dem dreistufigen Antwortformat der Lehrereinschätzungen verglichen (s. Tabelle 2). Unabhängig vom Einteilungskriterium zeigte sich deskriptiv mehrheitlich eine Übereinstimmung zwischen Lehrereinschätzung und Schülerdaten. Der Binomialtest zur Prüfung empirischer versus erwarteter Häufigkeiten (50:50) der beiden Randkategorien (Über- oder Unterschätzung) zeigte hypothesenkonform – für aller vier Leistungsindikatoren unter allen drei Einteilungskriterien – dass Lehrkräfte eher zu einer Über- als zu einer Unterschätzung von Leistungsheterogenität tendierten ($p < .05$).

Hypothese 3

In binär-logistischen Regressionsanalysen (Tabelle 3) wurde die relative Bedeutung von Leistungsheterogenitätsmaßen und weiteren Klassenkontextmerkmalen (kognitiv: mittlere Leistungen, sozio-demographisch: Anteil an Kindern mit Migrationshintergrund, strukturell-organisatorisch: Klassengröße) für die Lehrereinschätzung der Klasse als überdurchschnittlich leistungsheterogen geprüft. Um die Konsistenz der Ergebnisse zwischen Einzeltestleistungen sowie dem kombinierten Leistungsindex zu zeigen, gingen in Modell 1 die drei Leistungstests als separate Prädiktoren bzw. in Modell 2 der kombinierte Leistungsindex als Prädiktor ein. Für beide Modelle zeigten sich jeweils mit Sprache verbundene Merkmale als bedeutsame Prädiktoren der Lehrereinschätzung einer Klasse als deutlich leistungsheterogen: Zusätzlich zur klasseninternen Streuung im Leseverständnis (Modell 1) und im kombinierten Leistungsfaktor (Modell 2) übte jeweils der Anteil an Kindern mit Migrationshintergrund in Klassen einen positiven Effekt auf das Urteil aus. Je höher der Migrationsanteil in Klassen, desto eher nahmen Lehrkräfte – nach Kontrolle tatsächlicher Leistungsstreuung in Klassen und weiterer Kontextmerkmale – die jeweilige Klasse als deutlich leistungsheterogen wahr.

Wider Erwarten ließen sich für SES und Klassengröße keine zusätzlichen Effekte auf das Lehrerurteil einer Klasse als überdurchschnittlich leistungsheterogen nachweisen.

Diskussion

Das Thema Leistungsheterogenität hat in Deutschland sowohl in bildungspolitischen Diskussionen als auch in der Schul- und Unterrichtsforschung einen hohen Stellenwert. Daraus abgeleitete Schlussfolgerungen und Maßnahmen können weitreichende Konsequenzen, bis zur Umstrukturierung von Schulsystemen, haben. Die empirische Basis dafür, insbesondere an Grundschulen, ist jedoch aktuell als unzureichend zu bewerten (vgl. Kluczniok et al., 2011). Ziel der vorliegenden Analysen war es, eine breitere empirische Grundlage für den gegenwärtigen Diskurs zum Thema (Leistungs-)Heterogenität zu bieten.

Zunächst wurde geprüft, ob die anhand verschiedener Leistungsindikatoren berechneten klasseninternen Streuungen zu vergleichbaren Ergebnissen zum Ausmaß an Heterogenität von Klassen kommen. In dieser Stichprobe fanden sich keine positiven Zusammenhänge zwischen den klasseninternen Streuungen in Leseverständnis, naturwissenschaftlicher Kompetenz und kognitiven Grundfähigkeiten. Verschiedene Leistungsindikatoren kamen somit zu unterschiedlichen Rangreihen der Klassen bezogen auf das Ausmaß von Heterogenität. Von „der“ Leistungsheterogenität einer Klasse lässt sich also nicht sprechen; vielmehr sollte der jeweilige Leistungsbereich spezifiziert werden. Dieser Befund lässt vermuten, dass Modelle zur Vorhersage von individuellen Schülerleistungen mit verschiedenen Indikatoren von Leistungsheterogenität zu unterschiedlichen Ergebnissen führen, was mit zur uneinheitlichen Befundlage bei der Vorhersage von Schülerleistungen durch Leistungsheterogenitätsmaße beigetragen haben könnte (vgl. Bellin, 2009; Gröhlich et al., 2009; Künsting et al., 2010).

Anknüpfend an Befunde zur diagnostischen Kompetenz von Lehrkräften (z. B. Hoge & Coladarci, 1989) wurde geprüft, ob Lehrereinschätzungen zur Leistungsheterogenität mit Schülerdaten übereinstimmen – und wenn ja, welche Leistungsmaße die höchste

Übereinstimmung ergeben. Konsistent mit Befunden aus COACTIV (Kunter et al., 2007) gaben etwa die Hälfte aller Lehrkräfte eine durchschnittliche Leistungsheterogenität ihrer Klasse an. Erwartungsgemäß zeigten sich moderate Zusammenhänge zwischen dem global erhobenen Lehrerurteil einerseits und der objektiv gemessenen Heterogenität in Leseverständnis und naturwissenschaftlichen Kompetenzen andererseits. Wider Erwarten gab es auch eine signifikante Übereinstimmung zwischen Lehrerurteil und einem aus verschiedenen Leistungsbereichen kombinierten Leistungsindex (wie bspw. verwendet von Ditton, 2010). Schließlich deutet die mangelnde Übereinstimmung zwischen Lehrerurteil und klasseninternen Streuungen in kognitiven Grundfähigkeiten auf eine Diskrepanz zwischen der Schul- und Unterrichtsforschung und der Perspektive der Schulpraxis hin. Während kognitive Grundfähigkeiten als eine der zentralen Variablen für schulischen (Miss-)Erfolg angesehen werden können und oft (auch) zur Operationalisierung von Leistungsheterogenität verwendet werden (z. B. Bellin, 2009; Gröhlich et al., 2009), fokussierten Grundschullehrkräfte diesen Bereich nicht direkt oder nicht ausschließlich, wenn sie entsprechende Einschätzungen vornahmen. Allerdings sind kognitive Grundfähigkeiten im Schulalltag auch weit weniger sichtbar als bspw. das Leseverständnis. Somit könnten interindividuelle Unterschiede in den kognitiven Grundfähigkeiten eher indirekt, vermittelt über die fachspezifischen Leistungsunterschiede, für die Einschätzung von Leistungsheterogenität durch Lehrkräfte bedeutsam sein und/oder im kombinierten Leistungsindex zum Tragen kommen.

Darüber hinaus zeigen Ergebnisse dieser Stichprobe, dass Grundschullehrkräfte ihre Klassen mehrheitlich als durchschnittlich leistungsheterogen wahrnahmen und die Urteile mit den Schülerdaten mehrheitlich übereinstimmten. Im Falle einer Nicht-Passung tendierten sie jedoch erwartungsgemäß zu einer Überschätzung der Leistungsheterogenität ihrer Klasse. Inwieweit diese Überschätzung mit unterrichtlichem Handeln (z.B. vermehrter Einsatz von Differenzierungsmaßnahmen) und persönlichem Belastungserleben der Lehrkräfte in

Beziehung steht (Reh, 2005), kann mit dieser Studie nicht beantwortet werden. Die Ergebnisse dieser Studie legen jedoch nahe, dass im Zuge der Heterogenitätsdebatte und der steigenden Aufmerksamkeit für dieses Thema immer auch empirisch werden sollte, inwieweit wahrgenommene und tatsächliche Leistungsheterogenität übereinstimmen.

Schließlich wurde geprüft, welche Bedeutung Klassenkontextmerkmale für die Einschätzung einer Klasse als überdurchschnittlich leistungsheterogen haben. Studien zu Lehrereinschätzung en von individuellen Schülerleistungen belegen einen mit Klassenkontextmerkmalen verbundenen „Bias“ (z. B. Hauser-Cram et al., 2003; Ready & Wright, 2011). Die hier vorliegenden Befunde zu Lehrereinschätzungen einer gesamten Klasse zeigen zunächst, dass als Leistungsstreuungsmaße das Leseverständnis bzw. der kombinierte Leistungsindex zur Vorhersage des Lehrerurteils beitrug. Ergänzend und über Leistungsstreuungen und weitere Kontextmerkmale hinaus erwies sich der Anteil an Kindern mit Migrationshintergrund in Klassen als bedeutsam. Dagegen geht ein relativ hoher Anteil an Kindern mit niedrigerem SES insbesondere mit einer Unterschätzung des Leistungsniveaus (Ready & Wright, 2011), nicht aber mit der Wahrnehmung hoher Leistungsstreuung, einher. Unterricht ist eng mit Sprachkenntnissen in der Unterrichtssprache verbunden; möglicherweise fokussieren Lehrkräfte daher das Leseverständnis und den Migrationshintergrund von Kindern als besonders gut erfassbare Informationen bei ihrem Urteil zur Leistungsheterogenität von Klassen. Insgesamt sind die verschiedenen Heterogenitätsmerkmale sowohl in Klassifikationssystemen von Heterogenität als auch in der Empirie verwoben. Da der Migrationshintergrund von Kindern besonders eng mit einer Wahrnehmung von Leistungsheterogenität durch Lehrkräfte gekoppelt zu sein scheint, sollte in Lehrerbefragungen explizit auf eine Einschätzung von Leistungsheterogenität unabhängig vom Migrationshintergrund der Kinder hingewiesen und der Leistungsbereich spezifiziert werden. Die Klassengröße als in Befragungen genannte Herausforderung zum Umgang mit

Heterogenität (Winheller et al., 2012) leistete für Lehrereinschätzungen einer Klasse als überdurchschnittlich leistungsheterogen keinen eigenständigen Erklärungsbeitrag.

Grenzen des vorliegenden Beitrags

Lehrkräfte dieser Stichprobe gaben in Anlehnung an COACTIV (Kunter et al., 2007) ein globales Urteil zur Leistungsheterogenität einer gesamten Klasse ab. Dieses Format entspricht nicht den üblicherweise im Kontext von diagnostischer Kompetenz abgegebenen individuellen Fähigkeitseinschätzungen, spiegelt aber die bildungspolitische und pädagogische Heterogenitätsdebatte wider. Inwieweit die aus individuellen Fähigkeitseinschätzungen berechnete Streuungskomponente diagnostischer Kompetenz (vgl. Helmke & Schrader, 1987) mit Urteilen zur Leistungsheterogenität von gesamten Klassen übereinstimmen, war kein Ziel der Studie und bleibt somit zu prüfen.

Gleichermaßen bleibt zu prüfen, welche Referenzeinheit für „durchschnittlich“ Lehrkräfte zugrunde legen, wenn sie die Leistungsheterogenität einer durchschnittlichen Klasse der Jahrgangsstufe 3 einschätzen und inwieweit Merkmale auf Schulebene (Kooperation im Kollegium, Leistungserfassung durch Vergleichsarbeiten oder normierte Schulleistungstests) sowie der Lehrkraft selbst (z. B. Berufserfahrung) die Urteile mit beeinflussen. Zwar belegen Ready und Wright (2011), dass Lehrermerkmale im Vergleich zu Klassenkontextmerkmalen eine geringe Bedeutung für die Adäquatheit individueller Leistungseinschätzungen haben, jedoch können schulische Rahmenbedingungen und Lehrermerkmale für Urteile von Leistungsstreuungen gesamter Klassen bedeutsam sein.

Schließlich muss in Hinblick auf die Übereinstimmung von Schüler- und Lehrerdaten eingeschränkt werden, dass in den vorliegenden Analysen Mathematik als zentraler fachspezifischer Bereich in der Grundschule nicht berücksichtigt werden konnte. Es lassen sich auch hier substanzielle Zusammenhänge zwischen Schülerdaten und Lehrerurteilen erwarten; die empirische Prüfung steht jedoch noch aus.

Literatur

Bellin, N. (2009). *Klassenkomposition, Migrationshintergrund und Leistung. Mehrebenenanalyse zum Sprach- und Leseverständnis von Grundschülern*. Wiesbaden: VS.

Bos, W., Lankes, E.-M., Prenzel, M., Schwippert, K., Valtin, R., Voss, A. & Walther, G. (2005). *IGLU. Skalenhandbuch zur Dokumentation der Erhebungsinstrumente*. Münster: Waxmann.

Bos, W., Tarelli, I., Bremerich-Vos, A. & Schwippert, K. (2012). *IGLU 2011. Lesekompetenzen von Grundschulkindern im internationalen Bereich*. Münster: Waxmann.

Brunner, M., Anders, Y., Hachfeld, A. & Krauss, S. (2011). Diagnostische Fähigkeiten von Mathematiklehrkräften. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss & M. Neubrand (Hrsg.), *Professionelle Kompetenz von Lehrkräften – Ergebnisse des Forschungsprogramms COACTIV* (S. 215–234). Münster: Waxmann.

Ditton, H. (2010). Differentielle Leistungsentwicklung in der zweiten Grundschulzeit. *Zeitschrift für Grundschulforschung, 3*, 83–98.

Gröhlich, C., Scharenberg, K. & Bos, W. (2009). Wirkt sich Leistungsheterogenität in Schulklassen auf den individuellen Lernerfolg in der Sekundarstufe aus? *Journal for Educational Research Online, 1*, 86–105.

Gustafsson, J.-E. & Balke, G. (1993). General and specific abilities as predictors of school achievement. *Multivariate Behavioral Research, 28*, 407–434. doi:10.1207/s15327906mbr2804_2

Hauser-Cram, P., Sirin, S. R. & Stipek, D. (2003). When teachers' and parents' values differ: Teachers' ratings of academic competence in children from low-income families. *Journal of Educational Psychology, 95*, 813–820. doi:10.1037/0022-0663.95.4.813

Heinzel, F. (2008). Umgang mit Heterogenität in der Grundschule. In J. Ramseger & M. Wagener (Hrsg.), *Chancenungleichheit in der Grundschule. Ursachen und Wege aus der Krise* (S. 133–138). Wiesbaden: VS.

Helmke, A. & Schrader, F-W. (1987). Interactional effects of instructional quality and teacher judgment accuracy on achievement. *Teaching and Teacher Education, 3*, 91–98. doi:10.1016/0742–051X(87)90010-2

Hoge, R. D. & Coladarci, T. (1989). Teacher-based judgments of academic achievement: A review of literature. *Review of Educational Research, 59*, 297–313. doi:10.3102/00346543059003297

Karing, C., Matthäi, J. & Artelt, C. (2011). Genauigkeit von Lehrerurteilen über die Lesekompetenz ihrer Schülerinnen und Schüler in der Sekundarstufe I – Eine Frage der Spezifität? *Zeitschrift für Pädagogische Psychologie, 25*, 159–172. doi:10.1024/1010-0652/a000041

Karst, K. (2012). *Kompetenzmodellierung des diagnostischen Urteils von Grundschullehrern*. Münster: Waxmann.

Kluczniok, K., Große, C. & Roßbach, H.-G. (2011). Heterogene Lerngruppen in der Grundschule. In W. Einsiedler, M. Götz, A. Hartinger, F. Heinzel, J. Kahlert & U. Sandfuchs (Hrsg.), *Handbuch Grundschulpädagogik und Grundschuldidaktik* (S. 180–186). Bad Heilbrunn: Klinkhardt.

Kunter, M., Klusmann, U., Dubberke, T., Baumert, J., Blum, W., Brunner, M., Jordan, A., Krauss, S., Löwen, K., Neubrand, M. & Tsai, Y.-M. (2007). Linking aspects of teacher competence to their instruction: Results from the COACTIV project. In M. Prenzel (Hrsg.), *Studies on the educational quality of schools. The final report on the DFG Priority Programme* (S. 39–59). Münster: Waxmann.

Künsting, J., Post, S., Greb, K., Faust, G. & Lipowsky, F. (2010). Leistungsheterogenität im mathematischen Anfangsunterricht – Ein Risiko für die Leistungsentwicklung? *Zeitschrift für Grundschulforschung, 3*, 46–64.

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

Lorenz, C. & Artelt, C. (2009). Fachspezifität und Stabilität diagnostischer Kompetenz von Grundschullehrkräften in den Fächern Deutsch und Mathematik. *Zeitschrift für Pädagogische Psychologie, 23*, 211–222. doi:10.1024/1010-0652.23.34.211

Marsh, H. W., Lüdtke, O., Nagengast, B., Trautwein, U., Morin, A. J. S., Abduljabbar, A. S. & Köller, O. (2012). Classroom climate and contextual effects: conceptual and methodological issues in the evaluation of group-level effects. *Educational Psychologist, 47*, 106–124. doi:10.1080/00461520.2012.670488

Martin, M. O., Mullis, I. V. S. & Foy, P. (with Olson, J. F., Erberber, E., Preuschoff, C. & Galia, J.). (2008). *TIMSS 2007 International Science Report: Findings from IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grades*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V. S., Martin, M. O., Foy, P. & Arora, A. (2012). *TIMSS 2011 international results in science.* Chestnut Hill, MA: Boston College.

Ready, D. D. & Wright, D. L. (2011). Accuracy and inaccuracy in teachers' perceptions of young children's cognitive abilities: The role of child background and classroom context. *American Educational Research Journal, 48*, 335–360. doi:10.3102/0002831210374874

Reh, S. (2005). Warum fällt es Lehrerinnen und Lehrern so schwer, mit Heterogenität umzugehen? *Die Deutsche Schule, 97*, 76–86.

Schrader, F.-W. (1989). *Diagnostische Kompetenzen von Lehrern und ihre Bedeutung für die Gestaltung und Effektivität des Unterrichts*. Frankfurt a. M.: Lang.

Seeber, S. (2009). Urteilsgenauigkeit von Lehrerinnen und Lehrern in der sonderpädagogischen Förderung. In R. Lehman & E. Hoffmann (Hrsg.), *BELLA, Berliner Erhebung arbeitsrelevanter Basiskompetenzen von Schülerinnen und Schülern mit Förderbedarf „Lernen“* (S. 197–208). Münster: Waxmann.

Solzbacher, C. (2008). Was denken Lehrerinnen und Lehrer über individuelle Förderung? *Pädagogik, 60*, 38–42.

Spiegelhalter D. J., Thomas A., Best, N. G. & Lunn, D. (2003). *WinBUGS* (V. 1.4) [Computer Software]. Cambridge: MRC Biostatistics Unit. http://www.mrcbsu.cam.ac.uk/bugs/

Spinath, B. (2005). Akkuratheit der Einschätzung von Schülermerkmalen durch Lehrer und das Konstrukt der diagnostischen Kompetenz. *Zeitschrift für Pädagogische Psychologie, 19*, 85–95. doi:10.1024/1010-0652.19.12.85

Südkamp, A. & Möller, J. (2009). Referenzgruppeneffekte im Simulierten Klassenraum. *Zeitschrift für Pädagogische Psychologie, 23*, 161–174. doi:10.1024/1010-0652.23.34.161

Terhart, E. (2006). Kompetenzen von Grundschullehrerinnen und -lehrern. Kontext, Entwicklung, Beurteilung. In P. Hanke (Hrsg.), *Grundschule in Entwicklung* (S. 233–248). Münster: Waxmann.

Warm, T. A. (1989). Weighted likelihood estimation of ability in the item response theory. *Psychometrika, 54*, 427–450. doi:10.1007/BF02294627

Weiß, R. H. (2006). *CFT 20-R. Grundintelligenztest Skala 2. Revision*. Göttingen: Hogrefe.

Winheller, S., Müller, M., Hüpping, B., Rendtorff, B. & Büker, P. (2012). *Dokumentation der Studie ProLEG: Professionalisierung von Lehrkräften für einen reflektierten Umgang mit Ethnizität und Geschlecht in der Grundschule.* Paderborn: PLAZ-Forum.

Dr. Jasmin Decristan
Deutsches Institut für Internationale Pädagogische Forschung, BIQUA
BIQUA
60316 Frankfurt am Main
E-Mail: decristan@dipf.de

Alexander Naumann
Prof. Eckhard Klieme
Deutsches Institut für Internationale Pädagogische Forschung
IDeA-Forschungszentrum
Schloßstraße 29
60486 Frankfurt am Main

Benjamin Fauth
Svenja Rieser
Prof. Gerhard Büttner
Goethe-Universität Frankfurt
IDeA-Forschungszentrum
PEG - Gebäude
Grüneburgplatz 1
60323 Frankfurt am Main

Tabelle 1. Deskriptive Daten der Schülerstichprobe zu verschiedenen Leistungsbereichen sowie zu Klassenkontextmerkmalen

	N	*Min*	*Max*	*M*	*SD*
Gesamtstichprobe					
Kognitive Grundfähigkeiten	991	7	44	26.95	6.10
Leseverständnis	981	7	98	45.15	14.30
Naturwissenschaftliche Kompetenz	997	0	12	6.37	2.31
Klassenmittelwerte					
Kognitive Grundfähigkeiten	54	19.46	33.78	26.77	2.49
Leseverständnis	54	30.83	56.06	44.70	5.62
Naturwissenschaftliche Kompetenz	54	3.74	8.00	6.32	1.01
Streuungen in Klassen					
Kognitive Grundfähigkeiten	54	3.77	8.91	5.68	1.08
Leseverständnis	54	5.83	19.84	13.24	3.02
Naturwissenschaftliche Kompetenz	54	1.33	2.95	2.12	0.33
Soziodemographische und strukturell-organisatorische Klassenkontextmerkmale					
Migrationsanteil (in %)	54	0	83.33	44.07	21.20
Anteil geringer SES (in %)	54	6.67	85.00	44.03	18.16
Anzahl Kinder in Klasse	54	8	27	20.61	3.33

Tabelle 2. Prozent an (Nicht-)Übereinstimmung zwischen Lehrereinschätzung und latent geschätzten Leistungsstreuungen in Klassen bei unterschiedlichen Einteilungskriterien

	Kognitive Grundfähigkeiten	Leseverständnis	Naturwiss. Kompetenz	Kombinierter Leistungsindex
Kriterium A: Dreiteilung der Schülerdaten anhand +/– 1 SD				
Unterschätzung	14.3 %	10.2 %	16.3 %	12.2 %
Übereinstimmung	46.9 %	57.1 %	32.7 %	46.9 %
Überschätzung	38.8 %	32.7 %	51.0 %	40.8 %
Kriterium B: gleichmäßige Dreiteilung der Schülerdaten (jeweils 33 %)				
Unterschätzung	20.4 %	18.4 %	24.5 %	12.2 %
Übereinstimmung	38.8 %	40.8 %	26.5 %	53.1 %
Überschätzung	40.8 %	40.8 %	49.0 %	34.7 %
Kriterium C: drei qualitative Kategorien (> –1.2 SD, ≤ –0.8 SD bis ≤ 0.8 SD, > +1.2 SD)				
Klassenanzahl	*n = 36*	*n = 41*	*n = 37*	*n = 36*
Unterschätzung	13.9 %	12.2 %	18.9 %	8.3 %
Übereinstimmung	50.0 %	56.1 %	29.7 %	50.0 %
Überschätzung	36.1 %	31.7 %	51.4 %	41.7 %

Tabelle 3. Multiple binär-logistische Regressionsanalysen zur Vorhersage der Lehrereinschätzung zur Leistungsheterogenität durch Klassenkontextmerkmale

	Lehrereinschätzung zur Leistungsheterogenität (dichotom)							
	Modell 1				Modell 2			
	B (SE)	95% KI für Exp *B*			*B (SE)*	95% KI für Exp *B*		
		Untere s	Exp *B*	Obere s		Untere s	Exp *B*	Obere s
Latent geschätzte klasseninterne Streuungen								
Kognitive Grundfähigkeiten	0.15 (0.40)	0.53	1.16	2.54				
Leseverständnis	1.09* (0.51)	1.10	2.96	7.99				
Naturwiss. Kompetenz	0.67 (0.43)	0.86	1.99	4.61				
Kombinierter Leistungsindex					1.37* (0.49)	1.51	3.93	10.18
Weitere Klassenkontextmerkmale								
Latente Klassenmittelwerte kognitive Grundfähigkeiten	0.31 (0.48)	0.53	1.36	3.50				
Latente Klassenmittelwerte Leseverständnis	–0.34 (0.50)	0.27	0.71	1.89				
Latente Klassenmittelwerte naturwiss. Kompetenz	–0.05 (0.71)	0.24	0.96	3.86				
Latente Klassenmittelwerte kombinierter Leistungsindex					–0.23 (0.44)	0.33	0.79	1.90
Migrationsanteil	1.23* (0.61)	1.05	3.43	11.23	1.04* (0.49)	1.09	2.84	7.42
Anteil geringer SES	0.14 (0.48)	0.34	0.87	2.22	0.08 (0.44)	0.46	1.09	2.56
Anzahl Kinder in Klasse	0.41 (0.43)	0.65	1.50	3.49	0.52 (0.40)	0.77	1.68	3.67
Nagelkerkes Pseudo-R^2	.402				.418			

Anmerkungen: Lehrereinschätzung wurden dichotomisiert (0 = homogen/durchschnittlich, 1 = heterogen in den Leistungen). KI = Konfidenzintervall.

* $p < .05$.