Schmid, Christine; Paasch, Daniel; Katstaller, Michaela

Kompositionseffekte bei der Notenvergabe in Mathematik auf der 4. Schulstufe der österreichischen Volksschule

formal und inhaltlich überarbeitete Version der Originalveröffentlichung in:
formally and content revised edition of the original source in:
Zeitschrift für Bildungsforschung 6 (2016) 3, S. 265-283

Kompositionseffekte bei der Notenvergabe in Mathematik auf der 4. Schulstufe der österreichischen Volksschule

Christine Schmid, Daniel Paasch, Michaela Katstaller

PD Dr. C. Schmid
Abteilung Bildungsqualität und Evaluation, Deutsches Institut für Internationale Pädagogische Forschung (DIPF), Schloßstr. 29, 60486 Frankfurt a.M., Deutschland
E-Mail: christine.schmid@dipf.de

Dr. D. Paasch
Department Bildungsstandards und Internationale Assessments, Bundesinstitut für Bildungsforschung, Innovation und Entwicklung des österreichischen Schulwesens (BIFIE), Alpenstraße 121, 5020 Salzburg, Österreich

M. Katstaller, MA
School of Education, Universität Salzburg, Erzabt-Klotz-Str. 1, 5020 Salzburg, Österreich

Zusammenfassung Der Beitrag beschäftigt sich mit dem Thema leistungsangemessener Benotung durch Lehrkräfte am Ende der 4. Schulstufe der österreichischen Volksschule anhand einer mehrebenenanalytischen Betrachtung von Effekten der Leistungszusammensetzung von Klassen auf individuelle Noten. Die Analysen erfolgten mit Daten aus der 2013 vom Bundesinstitut für Bildungsforschung, Innovation und Entwicklung des österreichischen Schulwesens (BIFIE) durchgeführten Bildungsstandardüberprüfung Mathematik 4 (N = 73.655). Neben der individuellen Testleistung und der Leistungszusammensetzung der Klasse wurden das Geschlecht, der sozioökonomische Status (SES) und der Migrationshintergrund der Schüler/ innen berücksichtigt. Insgesamt ergab sich ein mittels logistischer Regressionen berechneter Zusammenhang zwischen Testleistung und Note von über 0,60. Demnach gelingt es Lehrkräften über Klassen und Schulen hinweg recht gut, die Leistungsrangfolge von Primarstufenschüler/inne/n mit Noten abzubilden. Darüber hinaus zeigten sich aber auch Referenzgruppeneffekte sowie eine weniger leistungsadäquate Benotungen von besonders leistungsschwachen und besonders leistungsstarken Schüler/inne/n. Zudem fanden sich positive Abweichungen der Noten von den Testleistungen bei Schüler/inne/n aus Elternhäusern mit höherem SES und negative bei jenen mit Migrationshintergrund (nur innerhalb von Klassen). Keine systematischen Abweichungen der Noten von den Testleistungen ergaben sich in Abhängigkeit vom Geschlecht.

Schlüsselwörter Klassenkomposition, Referenzgruppeneffekt, Primarstufe, Benotung, Bildungsstandards, Mehrebenenanalyse

Effects of class composition on marks in mathematics in fourth-grade Austrian elementary schools

Abstract This article addresses the issue of performance-appropriate marks for pupils in Austrian elementary schools at the end of fourth grade, based upon a multilevel analysis including the effect of the performance composition of classes on individual marks. The analyses were performed using data from the 2013 Learning Standards Assessment Mathematics 4 (N = 73,655) conducted by the Federal Institute for Educational Research, Innovation and Development of the Austrian School System (BIFIE). In addition to individual test performance and the performance composition of classes, gender, socio-economic status (SES), and the immigrant backgrounds of pupils were taken into account. Overall, logistic regressions reveal a relationship higher 0,60 between test performance and marks. This means that teachers were fairly successful in matching the performance rankings of primary school pupils with marks across classes and schools. However, the results also indicate the less-than-adequate evaluation of particularly low-performing and highperforming students. In addition, a positive deviation of marks compared to test performance was found for students from families with higher SES, and a negative deviation for students with immigrant backgrounds (within classes only). No systematic deviation of marks compared to test performance was found in relation to gender.

Keywords Classroom composition, Reference-group effect, Primary school, Grading, Learning standards, Multilevel-analysis

1 Einleitung

Leistungsbeurteilungen in Form von Noten sind ein vielschichtiges Unterfangen, bei dem unterschiedliche Anforderungen an die Lehrkräfte gestellt werden. Zum einen gilt es rechtliche Bestimmungen zu beachten, die aber häufig unkonkrete Handlungsanweisungen enthalten.[1] Zum zweiten spielen pädagogische Gesichtspunkte eine Rolle: Lehrkräfte können durch ihre Benotung Anreize setzen, welche die Schüler/innen zum Lernen motivieren oder sie auch in ihren Bemühungen entmutigen (Köller 2005). Zum dritten wird vor allem von gesellschaftlicher Seite eine leistungsgerechte Notenvergabe gefordert, denn Noten sind besonders in Übergangssituationen mit Berechtigungen verbunden. Um Chancengleichheit zu gewährleisten, sollten Noten ein vergleichbares Leistungsniveau zum Ausdruck bringen (Tent und Birkel 2010).

[1] https://www.ris.bka.gv.at/GeltendeFassung.wxe?Abfrage=Bundesnormen&Gesetzesnummer=10009375 (Zugriff am 11.07.2016).

1.1 Leistungsgerechte Benotung und die soziale Bezugsnorm

In der Literatur werden drei Bezugsnormen diskutiert, an denen sich Lehrkräfte bei der Notengebung orientieren (Rheinberg und Fries 2010). Diese drei Bezugsnormen – die kriterielle, die soziale und die individuelle – erfüllen die genannten Anforderungen in unterschiedlichem Maße. Beispielsweise zeigt die Forschung, dass die Orientierung an einer individuellen Bezugsnorm aus pädagogischer Sicht sinnvoll sein kann, da durch sie die Lernmotivation und das Leistungsselbstkonzept von Schüler/inne/n stärker gefördert werden als durch die Orientierung an einer sozialen Bezugsnorm (Rheinberg und Vollmeyer 2008).
Die Orientierung an der individuellen Leistung der Schüler/innen ist allerdings nicht geeignet, den Anspruch auf eine leistungsgerechte Notenvergabe zu erfüllen, da bei Zugrundelegung dieser Norm nur dann gute Noten gegeben werden, wenn sich individuelle Leistungssteigerungen zeigen. Schüler/innen, die bereits auf hohem Leistungsniveau agieren, haben aber kaum die Möglichkeit, sich noch zu verbessern.
Unter dem Gesichtspunkt einer leistungsgerechten Notenvergabe ist die sachliche Bezugsnorm, die sich an objektivierbaren Leistungskriterien orientiert, die erstrebenswerte (Tent und Birkel 2010). Für Lehrkräfte ist diese Bezugsnorm jedoch kaum umsetzbar, wenn allgemeingültige Standards für eine objektive Leistungsbeurteilung nur unklar formuliert sind oder gänzlich fehlen. Sie behelfen sich daher häufig mit der Anwendung einer sozialen Bezugsnorm, d. h. mit dem Vergleich der Leistungen innerhalb ihrer Klassen.
Innerhalb von Klassen sind Leistungen leichter objektivierbar als über Klassen hinweg, beispielsweise indem bei schriftlichen Leistungsüberprüfungen die Aufgaben mit festen Punktzahlen versehen und den Noten zu erreichende Punktzahlen zugewiesen werden. Nicht selten wird dabei der Notenspiegel dem Leistungsniveau der Klasse angepasst (Sacher 2009). Eine solche Benotungspraxis bedeutet, dass zwar die Rangfolge der Leistungen innerhalb der Klassen berücksichtigt wird – innerhalb von Klassen ist demnach Leistungsgerechtigkeit gewährleistet. In Klassen mit einem niedrigeren Leistungsniveau spiegelt eine gute Note jedoch eine weniger gute Leistung wider als in Klassen mit einem hohen Leistungsniveau – der Anspruch auf Leistungsgerechtigkeit über Klassen hinweg kann somit nicht aufrechterhalten werden. In der Literatur wird dieses Phänomen auch als Referenz- oder Bezugsgruppeneffekt diskutiert (Ingenkamp 1995; Lintorf 2012). Gleichzeitig handelt es sich hierbei um einen Kompositionseffekt, denn neben der individuellen Leistung beeinflusst zusätzlich die leistungsmäßige Zusammensetzung der Klasse die Noten der Schüler/innen. Im Schulalltag kommt in der Regel wohl eine Mischung aus allen drei Bezugsnormen zur Anwendung, wobei der sozialen Bezugsnorm ein hoher Stellenwert zugeschrieben wird (Sacher 2009).

1.2 Zum Zusammenhang von Note und Leistung

Wie gut Noten Leistungen widerspiegeln, hängt also stark von der Benotungspraxis von Lehrkräften ab. Mit Blick auf die Schüler/innenschaft eines gesamten Jahrgangs spielen aber auch systemische Bedingungen des Bildungserwerbs eine Rolle. In leistungsdifferenzierten Schulsystemen bilden die unterschiedlichen Leistungsgruppen eigene Referenzrahmen hinsichtlich der zu erwartenden Leistungen.
Auswertungen auf der Grundlage von PISA 2000 zeigen entsprechend, dass in Deutschland Noten innerhalb von Schulformen die Testleistungen besser repräsentieren als über Schulformen hinweg (Baumert et al. 2003). Innerhalb der Schulformen lag die Korrelation der Leistung in einem curriculumsnahen Mathematiktest mit der Mathematiknote bei $r = –0{,}46$, über alle Schulformen

hinweg dagegen nur bei $r = –0,32$. Zudem erklärte die Schulform 37% der Varianz des Mathematikleistungstests, aber nur 20% der Varianz der Mathematiknoten. Die Noten waren sich demnach über die Bildungsgänge hinweg ähnlicher als die Testleistungen.
Entsprechendes lässt sich für Österreich anführen (Eder und Dämon 2010). Auf Grundlage der Pilotuntersuchung zu den Bildungsstandardüberprüfungen in der 8. Schulstufe[2] ergaben sich für die verschiedenen Leistungsniveaus (AHS, 1., 2. und 3. Leistungsgruppe HS)[3] nahezu deckungsgleiche Notenverteilungen. Ein- und dieselbe Note drückte demnach ein anderes Leistungsniveau aus.
In der Primarstufe wird nicht leistungsdifferenziert unterrichtet, weshalb die Zusammenhänge zwischen Noten und Testleistungen hier höher ausfallen sollten als in der Sekundarstufe. Tatsächlich fanden sich in der IGLU-Studie (Bos et al. 2004) höhere Korrelationen zwischen dem Mathematiktest und der Mathematiknote ($r = –0,55$) sowie zwischen dem Lesetest und der Deutschnote ($r = –0,56$) als in der PISA 2000-Studie.
Ebenfalls hoch fiel der Zusammenhang zwischen Mathematiknoten und den entsprechenden Testleistungen auf der Grundlage einer Stichprobe von 841 österreichischen Schüler/inne/n der 7. und 8. Klassen aus ($r = –0,69$; Eder 2003). Dieser hohe Wert erklärt sich vor allem dadurch, dass die Noten nach gängiger österreichischer Praxis in eine achtstufige Skala umgerechnet wurden, wodurch wiederum den unterschiedlichen Referenzrahmen in den jeweiligen Bildungsgängen und Leistungsgruppen Rechnung getragen wurde.[4]
In der angeführten Untersuchung zeigten sich zudem systematische Abweichungen der Noten von den Testleistungen: Schüler/innen der AHS wiesen im Vergleich zu jenen der 1. Leistungsgruppe HS im Schnitt die besseren Testleistungen auf, erhielten aber schlechtere Noten. Diese ungünstigere Benotung der AHS-Schüler/innen wurde im Sinne eines Referenzgruppeneffekts interpretiert.

1.3 Zur Verteilung von Noten

Die schulrechtlichen Bestimmungen in Österreich sehen eine kriteriumsbezogene Notenvergabe durch die Lehrkräfte vor. In der Leistungsbeurteilungsverordnung ist festgehalten:

> „Der Lehrer hat die Leistungen der Schüler sachlich und gerecht zu beurteilen, dabei die verschiedenen fachlichen Aspekte und Beurteilungskriterien der Leistung zu berücksichtigen und so eine größtmögliche Objektivierung der Leistungsbeurteilung anzustreben.“ (LB-VO § 11 Abs. 2).

Grundlage für die Beurteilung sollen die Mitarbeit im Unterricht sowie mündliche und schriftliche Leistungsfeststellungen sein. Der Maßstab für die Beurteilung leitet sich aus den Forderungen des Lehrplanes unter Bedachtnahme auf den jeweiligen Stand des Unterrichts ab. Explizit nicht in die Beurteilung einfließen sollen das Verhalten (Benehmen) der Schüler/innen in Schule und Öffentlichkeit, die äußere Form der Arbeit (außer in geregelten Ausnahmefällen) sowie von der

[2] Bildungsstandardüberprüfungen werden in Österreich landesweit und nicht differenziert nach angestrebten Schulabschlüssen durchgeführt.
[3] Die Allgemeine Höhere Schule (AHS) entspricht in etwa den deutschen Gymnasien, die Hauptschule (HS) bildete bis vor kurzem die einzige weitere Schulform in Österreich. Aktuell werden die HS sukzessive durch Neue Mittelschulen (NMS) ersetzt, in denen die Leistungsgruppendifferenzierung zu Gunsten eines binnendifferenzierten Unterrichts aufgegeben wird.
[4] In der AHS und der 1. Leistungsgruppe HS werden vergleichbare Leistungen erbracht. In der 2. Leistungsgruppe
HS wird der Wert 2 und in der 3. Leistungsgruppe der Wert 3 zur Note hinzuaddiert (Eder 2003).

Meinung der Lehrkraft abweichende, sachlich vertretbareMeinungsäußerungen der Schüler/innen (Informationsblatt zum Schulrecht, Teil 3, BMUKK 2007).[5]
Über die Noten ist festgehalten, dass mit „Sehr gut“ Leistungen bewertet werden sollen, welche die Anforderungen des Lehrplans „in weit über das Wesentliche hinausgehendem Ausmaß“ erfüllen sowie eine „deutliche Eigenständigkeit beziehungsweise die Fähigkeit zur selbständigen Anwendung“ des Wissens und Könnens auf neuartige Aufgaben erkennen lassen. Für die Note „Gut“ sind Leistungen gefordert, welche die Anforderungen des Lehrplans „in über das Wesentliche hinausgehendem Ausmaß erfüllen“ sowie „merkliche Ansätze zur Eigenständigkeit beziehungsweise bei entsprechender Anleitung die Fähigkeit zur Anwendung“ des Wissens und Könnens auf neuartige Aufgaben zeigen. Mit der Note „Befriedigend“ sind Leistungen zu beurteilen, welche die im Lehrplan gestellten Anforderungen „in den wesentlichen Bereichen zur Gänze“ erfüllen, „dabei werden Mängel in der Durchführung durch merkliche Ansätze zur Eigenständigkeit ausgeglichen“. Mit „Genügend“ sind Leistungen zu bewerten, welche die Anforderungen „in den wesentlichen Bereichen überwiegend“ erfüllen und die Note „Nicht genügend“ ist den Leistungen vorbehalten, welche „nicht einmal alle Erfordernisse“ für die Beurteilung mit „Genügend“ erfüllen (LB-VO § 14 Abs. 2–6). Den Anforderungen des Lehrplans wird demnach im Wesentlichen bereits bei einer Beurteilung mit „Befriedigend“ genüge getan.
Österreichische Untersuchungen (Eder und Dämon 2010) zeigen, dass in der Sekundarstufe I tatsächlich „Befriedigend“ die am häufigsten vergebene Note darstellt. Nur sehr selten wird ein „Nicht genügend“ vergeben, die Note „Sehr gut“ kommt in nennenswertem Umfang nur in der AHS und der 1. Leistungsgruppe HS vor. Zudem gleicht die Notenverteilung sowohl über die gesamte Schüler/innenschaft hinweg als auch innerhalb von Leistungsgruppen am ehesten einer Normalverteilung.
Deutlich anders sieht die Notenverteilung am Ende der Primarstufe aus, wenn es um den Übergang in weiterführende Schulen geht. In einer Untersuchung von Eder (2007) lag etwa ein Viertel der Schüler/innen am Ende der 4. Klasse im Schnitt (Deutsch, Sachunterricht und Mathematik) auf der Note 1, über die Noten 2 bis 5 hinweg nahmen die Anteile kontinuierlich ab. Eine ähnlich rechtsschiefe Verteilung berichtet Wallner-Paschon (2009) auf Grundlage von PIRLS 2006 für die Mathematiknote in der 4. Klasse Volksschule: Hier wurden 40% Einsen, 35% Zweien, 18% Dreien, 6% Vieren und 1% Fünfen vergeben. Etwas strenger fiel die Deutschnote aus mit 29% Einsen, 37% Zweien, 26% Dreien, 8% Vieren und unter 1% Fünfen.

1.4 Zur Reliabilität und Validität von Noten und Testleistungen

Die Bildungsstandards für Mathematik in der Volksschule legen diejenigen Kompetenzen fest, die Schüler/innen bis zum Ende der 4. Schulstufe erworben haben sollen. Sie wurden von Lehrer/inne/n und Fachdidaktiker/inne/n in Form von konkret formulierten Lernergebnissen aus den Lehrplänen abgeleitet (Schreiner und Breit 2014). In Österreich wurden die Bildungsstandards in Form von Regelstandards formuliert und in Kompetenzstufen unterteilt. Testwerte, die der Kompetenzstufe 3 zugeordnet sind, indizieren, dass die Bildungsstandards nicht nur erreicht, sondern übertroffen wurden. Testwerte auf der Stufe 2 dokumentieren das Erreichen der Bildungsstandards. Testwerte auf der Stufe 1 besagen, dass die Bildungsstandards nur teilweise erreicht wurden, und Testwerte unter der Kompetenzstufe 1 bedeuten das Nichterreichen der Standards (Schreiner und Breit 2014). Eine direkte Zuordnung der Noten zu Kompetenzstufen ist weder möglich (fünf Noten, aber nur vier Kompetenzstufen), noch war sie intendiert. Unter messtheoretischen Gesichtspunkten haben beide

[5] https://www.bmbf.gv.at/schulen/recht/info/schulrecht_info_3_5822.pdf?4dzi3h (Zugriff am11.07.2016).

Formen der Leistungsbeurteilung jeweils spezifische Stärken und Schwächen. Experimentelle Studien etwa lassen erhebliche Zweifel an der Objektivität und Reliabilität von Schulnoten aufkommen. Schulnoten sind nicht nur durch die bereits angesprochenen Referenzgruppeneffekte geprägt, sondern werden beispielsweise auch durch Erwartungen und Vorurteile der Lehrkräfte beeinflusst (zusammenfassend Lintorf 2012; Tent und Birkel 2010). Allerdings stellen auch Leistungstests keine fehlerfreien Messungen dar. Bei einem Test handelt es sich um eine punktuell durchgeführte Messung, deren Ergebnis u. a. durch die Tagesform der Getesteten beeinflusst sein kann. Noten dagegen, insbesondere Zeugnisnoten, basieren auf Leistungsbeurteilungen, die über einen längeren Zeitraum hinweg gesammelt werden. Testleistungen werden zudem überwiegend schriftlich und häufig im Multiple-Choice-Format erhoben. Lehrkräfte dagegen nutzen verschiedene Methoden der Leistungserhebung, insbesondere auch mündliche Formen. Bei der Beurteilung berücksichtigen sie stärker den Kontext der Leistungserbringung.
Immer dann, wenn die Mitarbeit im Unterricht sowie mündliche Leistungen mitbenotet werden sollen, decken Noten und Testleistungen nicht dasselbe ab. Lintorf (2012) spricht in diesem Zusammenhang von einer Asymmetrie von Note und Testleistung. Eine solche Asymmetrie besteht auch dann, wenn das Testmodell, das der Konstruktion eines Leistungstests zugrunde liegt, nicht dem intendierten Curriculum entspricht.
Die Beurteilung der Validität von Noten im Vergleich zu Testleistungen hängt maßgeblich vom herangezogenen Kriterium ab (z. B. Prognosesicherheit späteren Leistungserfolgs, Übereinstimmung mit anderen Testleistungen; zusammenfassend Tent und Birkel 2010). Festgehalten werden kann, dass Noten im Allgemeinen instruktional valider sind, d. h. sich stärker auf das beziehen, was im Unterricht tatsächlich thematisiert wurde. Testleistungen dagegen, insbesondere wenn sie im Rahmen von Bildungsstandardüberprüfungen erhoben werden, sollten im Hinblick auf das intendierte Curriculum valider sein. Sie haben zudem gegenüber Noten den unbestrittenen Vorteil, dass sie einen über Schulen und Klassen hinweg objektivierten Bezugsrahmen darstellen (Wilhelm und Kunina-Habenicht 2015). Der Vergleich von Testleistungen und Noten kann deshalb wertvolle Hinweise auf die Benotungspraxis von Lehrkräften liefern.

1.5 Systematische Abweichungen der Noten von Leistungstestwerten

Um systematische Abweichungen von Noten gegenüber Testwerten handelt es sich, wenn diese in Abhängigkeit von bestimmten Merkmalen wie dem Geschlecht, dem sozioökonomischen Status oder dem Migrationshintergrund auftreten. Solche Abweichungen können ihre Ursache in der mangelnden Objektivität von Lehrkräften haben oder auf eine legitime Berücksichtigung von Leistungsmerkmalen wie z. B. der mündlichen Mitarbeit im Unterricht zurückgehen.
Es liegen Forschungsbefunde vor, die solche systematischen Abweichungen belegen. Maaz et al. (2013) berichten auf der Grundlage von Daten der ELEMENT-Studie über positive Effekte des Geschlechts zu Gunsten der Mädchen sowie der sozialen Herkunft (ISEI) auf die Durchschnittsnote am Ende der Grundschulzeit. Der Migrationshintergrund zeigte keinen über die genannten Faktoren hinausgehenden Effekt. Eine zweite Analyse, die auf der Grundlage von TIMSS-Daten durchgeführt wurde, führte zu ähnlichen Ergebnissen: Bei Kontrolle der Leistungstestwerte ergaben sich positive Effekte für die kognitiven Grundfähigkeiten, für das Geschlecht zu Gunsten der Mädchen, für die soziale Herkunft zu Gunsten eines höheren ISEI, eines höheren elterlichen Bildungshintergrundes und Bücherbesitzes sowie häufigeren kulturellen Aktivitäten der Eltern. Keine Belege fanden sich für die weitergehende Annahme, dass die strengere Benotung von Jungen sowie von Schüler/inne/n aus sozial benachteiligten Elternhäusern auf geschlechts- bzw. sozialschichtspezifische Ausprägung

motivationaler Merkmale zurückzuführen sind.
Eine Untersuchung von Ditton und Krüsken (2006) dokumentiert allerdings, dass solche verhaltensbezogenen Merkmale, wenn sie über die Lehrkräfte erhoben werden, durchaus eine Rolle spielen können. In einer Längsschnittstudie unter Einbezug von 30 bayrischen Schulen erwies sich neben den Testleistungen (3. und 4. Klasse) und der Einschätzung der Lesefähigkeit und des sprachlichen Ausdrucks (4. Klasse) die Beurteilung der sozial- und arbeitsstilbezogenen Kompetenzen (4. Klasse) als prädiktiv für die Abschlussnote in der 4. Klasse. Einen ebenfalls positiven, aber auf Grund der Kontrolle der Beurteilung der sozial- und arbeitsstilbezogenen Kompetenzen sowie anderer Größen nur noch schwachen Effekt, hatte ein höherer sozialer Status des Elternhauses.

2 Forschungsfragen und Vorgehensweise

Mit der im Jahre 2013 vom Bundesinstitut für Bildungsforschung, Innovation und Entwicklung des österreichischen Schulwesens (BIFIE) durchgeführten Bildungsstandardüberprüfung Mathematik 4. Schulstufe (Schreiner und Breit 2014) liegt eine Vollerhebung aller 4. Klassen der österreichischen Volksschule vor. Die Vollerhebung ermöglicht Mehrebenenanalysen unter Berücksichtigung von Klassen.
Neben den Testleistungen in Mathematik liegen Angaben zur Halbjahresnote vor, welche über den Schüler/innenfragebogen erhoben wurden. Zudem wurden Angaben zum Geschlecht, zum Beruf von Vater und Mutter und zum Migrationshintergrund erhoben.

2.1 Verteilung der Noten

In einem ersten Schritt soll deskriptiv die Verteilung der Noten mit jener der Testleistungen verglichen werden. Der Mathematiktest zur Überprüfung der Bildungsstandards zielt auf die Abbildung von Regelstandards ab, er ist so konstruiert ist, dass sich für die Testleistungen eine Normalverteilung ergibt (Schreiner und Breit 2014). Für die Noten wird entsprechend den berichteten Befunden (Wallner-Paschon 2009) eine eher rechtsschiefe Verteilung erwartet.

2.2 Der Zusammenhang von Noten und Testleistungen

Die schulrechtlichen Bestimmungen in Österreich sehen eine sachbezogene Leistungsbeurteilung vor. Mit der Überprüfung der Bildungsstandards Mathematik in der 4. Klasse Volksschule liegen erstmals Daten vor, die es aufgrund des einheitlichen Referenzrahmens erlauben, die Leistungen über einzelne Klassen und Schulen hinweg zu vergleichen. Der Vergleich von Testleistungen mit den Halbjahresnoten in Mathematik kann Hinweise liefern, wie gut oder schlecht Noten die Leistungsrangfolge auch über Klassen hinweg abbilden. Dass dies keine einfache Aufgabe darstellt und Noten und Testleistungen zudem Unterschiedliches messen und bezwecken, wurde bereits erwähnt. Erwartet wird deshalb nur ein mittlerer Zusammenhang zwischen Testleistungen und Noten.

2.3 Systematische Abweichungen der Noten von den Testleistungen

Die Höhe des Zusammenhangs zwischen Testleistungen und Noten hängt auch davon ab, inwieweit sich systematische Abweichungen ergeben. Vor dem Hintergrund der dargestellten Befunde kann erwartet werden, dass Mädchen sowie Schüler/innen aus sozioökonomisch besser gestellten

Elternhäusern jeweils bessere Noten erzielen als es ihren Testleistungen entspricht. Für den Migrationshintergrund hingegen sollten sich keine Abweichungen zwischen Noten und Leistungen ergeben.

2.4 Soziale Bezugsnorm

Starke Zusammenhänge zwischen Testleistungen und Noten sprechen dafür, dass Lehrkräfte objektive Kriterien bei der Bewertung der Schüler/innenleistungen zu Grunde legen. Die Stärke des Zusammenhangs wird jedoch bei Anwendung der sozialen Bezugsnorm geschmälert, weil unter diesen Umständen eine gute Leistung in einer leistungsschwachen Klasse besser benotet wird als eine gute Leistung in einer leistungsstarken Klasse.
Ein erster Hinweis auf die Anwendung der sozialen Bezugsnorm durch Lehrkräfte liegt vor, wenn in Mehrebenenanalysen mehr Varianz bei den Testleistungen als bei den Noten durch die Klasse erklärt wird. Bei Anwendung der sozialen Bezugsnorm werden unabhängig vom Leistungsniveau der Klasse jeweils annähernd normalverteilte Noten vergeben, die Varianz zwischen den Klassen wird dadurch minimiert.
Ein zweiter Hinweis ist gegeben, wenn sich in Mehrebenenanalysen ein negativer Effekt des Klassenmittelwerts der Testleistungen auf die Noten als abhängiger Größe ergibt. Ein solcher Effekt indiziert, dass in Klassen mit einem hohen mittleren Leistungsniveau für die gleichen Leistungen schlechtere Noten vergeben werden als in Klassen mit einem niedrigeren mittleren Leistungsniveau. Solche inversen Effekte des mittleren Leistungsniveaus von Klassen auf Noten zeigten sich sowohl in einer mehrebenenanalytischen Untersuchung von Lintorf (2012), die auf Daten einer Erhebung in deutschen Grundschulen basierte, als auch in einer Untersuchung von Hochweber et al. (2014), in der Daten von Achtklässlern der MARKUS-Studie ausgewertet wurden. Nicht mehrebenenanalytische Auswertungen von Daten der österreichischen PIRLS-Studie von 2006 ergaben ebenfalls, dass Noten vom Leistungsniveau der einzelnen Klassen abhängig waren (Wallner- Paschon 2009).

2.5 Die besonders leistungsschwachen und leistungsstarken Schüler/innen in den Klassen

Qualitative Studien weisen darauf hin, dass Lehrkräfte dazu neigen, besonders leistungsschwachen Schüler/inne/n eine nicht leistungsangemessene Note zu geben, um ihnen ein negatives Feedback zu ersparen, welches sie demotivieren und die weitere Kooperation im Unterricht untergraben könnte (Streckeisen et al. 2006). Eine, gemessen an den Leistungen, positivere Benotung von leistungsschwachen Schüler/inne/n hätte einen schwächeren Zusammenhang zwischen Testleistungen und Noten in der unteren Leistungsgruppe zur Folge. Diese Annahme soll überprüft werden, indem klassenspezifisch das leistungsschwächste Quartil identifiziert wird.
Bei einer rechtschiefen Verteilung der Noten liegt der Verdacht nahe, dass auch am anderen Ende der Leistungsskala Noten die Leistungen nicht hinreichend differenziert wiedergeben. Bei einem hohen Anteil an Einsen kann von einem Deckeneffekt bei der Benotung ausgegangen werden. Dieser sollte sich ebenfalls in einem abgeschwächten Zusammenhang zwischen Testleistungen und Noten innerhalb des jeweils obersten Leistungsquartils im Vergleich zu den entsprechenden mittleren Leistungsquartilen in den Klassen niederschlagen.

3 Vorgehensweise

Die erste Frage zielt auf die Verteilung der Noten im Vergleich zu jener der Testleistungen ab. Hierfür ist es ausreichend, die Kennwerte beider Verteilungen zu dokumentieren. Für die zweite Frage, die auf den Zusammenhang zwischen Testleistungen und Noten gerichtet ist, wird eine einfache logistische Regression mit den Noten als abhängiger und den Testleistungen als unabhängiger Größe berechnet. Zur Beantwortung der dritten Frage wird die einfache logistische Regression um die Faktoren Geschlecht, soziökonomische Herkunft (SES) und Migrationshintergrund erweitert. Von systematischen Abweichungen der Noten von den Testleistungen kann gesprochen werden, wenn die genannten Faktoren bei gegenseitiger Kontrolle und über die Testleistungen hinaus noch einen Effekt auf die Noten haben.
Die Frage, ob die Klasse bei den Testleistungen mehr Varianz als bei den Noten erklärt, wird anhand der Berechnung von Intraklassenkorrelationen beantwortet. Die Intraklassenkorrelation gibt den Anteil an der Gesamtvarianz an, der auf Unterschiede zwischen den Klassen zurückzuführen ist. Ein erster Hinweis auf die Anwendung der sozialen Bezugsnorm durch Lehrkräfte ist gegeben, wenn die Intraklassenkorrelation bei den Testleistungen höher ausfällt als bei den Noten.
Ein weiterer Hinweis auf die Anwendung der sozialen Bezugsnorm wäre gegeben, wenn sich für die leistungsmäßige Zusammensetzung der Klasse ein negativer Effekt auf die Noten zeigt. Um dies zu prüfen, wird eine zweite Mehrebenenanalyse berechnet. Die abhängige Größe bilden die Mathematiknoten, als unabhängige Größen werden die Testleistungen, das Geschlecht, die soziale Schicht und der Migrationshintergrund auf Individualebene sowie die mittleren Testleistungen der Klassen auf Klassenebene berücksichtigt. Technisch entspricht dieses Modell einem *random-intercept model* (type = twolevel mit Prädiktoren auf Individual- und Klassenebene). Erwartet werden ein positiver Effekt der Testleistungen auf Individualebene sowie ein negativer Effekt der mittleren Testleistungen auf Klassenebene.
Zur Beantwortung der letzten Frage, welche auf die Möglichkeit einer weniger leistungsangemessenen Benotung im jeweils unteren und oberen Leistungsquartil der Klassen gerichtet ist, werden zunächst anhand eines eigens erstellten Skripts die klassenspezifischen Leistungsquartile ermittelt. Die Zugehörigkeit zu diesen wird dann mit entsprechenden Interaktionstermen in einer dritten Mehrebenenanalyse berücksichtigt. Mittels der Interaktionseffekte wird überprüft, ob der Zusammenhang zwischen Testleistung und Note bei Zugehörigkeit zum oberen bzw. unteren Leistungsquartil im Vergleich zu den beiden mittleren Leistungsquartilen in den Klassen abgeschwächt oder erhöht ist. Vor dem Hintergrund der Annahmen wären Interaktionseffekte mit negativen Vorzeichen zu erwarten, da diese eine Abschwächung des jeweiligen Zusammenhangs indizieren.

4 Methode

Die Zielpopulation der BildungsstandardüberprüfungMathematik 4 von 2013 bilden alle österreichischen Schüler/innen der 4. Jahrgangsstufe (*N* = 75.797; 3050 Schulen; 4920 Klassen), ausgenommen außerordentliche Schüler/innen sowie jene mit sonderpädagogischem Förderbedarf (Schreiner und Breit 2014). Da einige Schüler/innen am Testtag, beispielsweise auf Grund von Krankheit, abwesend waren, wurden nur 73.655 Personen getestet; dies entspricht einer Ausschöpfungsquote von 97,2%. Um für systematische Ausfälle zu kompensieren, wurden Schüler/innengewichte erzeugt (Trendtel 2015). Die Testung wurde durch zuvor geschulte Lehrkräfte

durchgeführt. Neben dem Test wurden Schüler/innen-, Eltern-, Lehrer/innen- und Schulfragebögen für die Erhebung von Kontextvariablen eingesetzt.

4.1 Instrumente

Testleistung Mathematik: Für die Testung wurden unterschiedliche Testhefte (insgesamt 30) mit ähnlichem Schwierigkeitsgrad entwickelt. Über alle Testhefte hinweg bestand der Test aus 254 Items (hauptsächlich Multiple-Choice-Aufgaben, aber auch halboffene und offene Aufgaben). Die einzelnen Testhefte enthielten in der Regel 68 Items und waren innerhalb von 80 min zu bearbeiten. Inhaltlich deckte der Test die allgemeinen Kompetenzen Modellieren, Operieren, Kommunizieren und Problemlösen ab sowie die mathematischen Kompetenzen des Arbeitens mit Zahlen, mit Operationen, mit Größen, sowie mit Ebene und Raum (Schreiner und Breit 2014). Der Test enthielt Ankeritems, welche den Vergleich mit einer Baseline-Testung ermöglichen, die im Jahr 2010 auf Grundlage einer repräsentativen Stichprobe von 267 Volksschulen durchgeführt wurde. Die Skalierung der Items erfolgte im Rahmen der Item-Response-Theorie (Yen und Fitzpatrick 2006) auf einer Metrik, die an der Baselinetestung orientiert war und damals mit einem Mittelwert von 500 und einer Standardabweichung von 100 normiert wurde. Die Testwerte liegen in Form von 10 *plausible values* vor (von Davier et al. 2009).

Geschlecht: Das Geschlecht wurde über den Schüler/innenfragen erhoben und weist die Kategorien 0 = weiblich und 1 = männlich auf. Von den getesteten Schüler/inne/n waren 51% männlich.

Sozioökonomischer Hintergrund (SES): Der sozioökonomische Hintergrund wird über den HISEI-Wert abgebildet. Dieser wurde auf Grundlage der Angaben zum Beruf beider Elternteile ermittelt. Kodiert wurden die Berufe nach dem Schema ISCO-08 (Freunberger et al. 2014). Für die Bildung des HISEI wurde der jeweils höhere Wert von Mutter oder Vater bzw. der einzig vorhandene Wert herangezogen. Der HISEI weist einen Mittelwert von 46,72 (SD = 20,67) auf.

Migrationshintergrund: Die Angaben zum Migrationshintergrund stammen ebenfalls aus dem Schüler/innenfragebogen. Ein Migrationshintergrund liegt dann vor, wenn Vater und Mutter im Ausland geboren wurden. Wurde nur ein Elternteil im Ausland geboren oder beide Elternteile in Deutschland, so wurde dies nicht als Migrationshintergrund gewertet (Freunberger et al. 2014). Die Variable hat die Ausprägungen 0 = kein Migrationshintergrund und 1 = Migrationshintergrund. Insgesamt wiesen 19% der Schüler/innen einen Migrationshintergrund auf.

Mathematiknote: Auch die Mathematiknote wurde über den Schüler/innenfragebogen erhoben, konkret wurde nach der Note im letzten Semesterzeugnis (Halbjahresnote) gefragt. Das Notenspektrum in Österreich erstreckt sich auf die Noten 1 (Sehr gut) bis 5 (Nicht genügend). Um die Interpretation der Ergebnisse zu erleichtern, wurde die Note umkodiert, so dass hohe Werte bessere Noten indizieren.

4.2 Umgang mit Missings

Um sicherzustellen, dass es zu keinen Verzerrungen in den Ergebnissen durch fehlende Werte kommt, wurden diese über das Verfahren der *multiple imputation* (Lüdtke et al. 2007) geschätzt. Insgesamt wurden 10 Imputationsdatensätze (je einer pro *plausible value*) erzeugt.

4.3 Analysen

Alle folgenden Analysen mit Ausnahme der Verteilungsdarstellungen wurden mit Mplus Version 7.31 unter Verwendung der Schüler/innengewichte berechnet. Um dem ordinalen Skalenniveau der

Noten Rechnung zu tragen, wurden die Regressionen nicht linear, sondern logistisch berechnet (Eid et al. 2013). Bei den Mehrebenenanalysen wurde für die metrischen Variablen die Voreinstellung von Mplus der Zentrierung am *grand mean* beibehalten.
Mplus gibt für logistische Regressionen unstandardisierte, *y*-standardisierte sowie *yx*-standardisierte Koeffizienten aus. Die unstandardisierten Koeffizienten geben an, um wieviel sich der Logit (= *log odds*) der Mathematiknote bei Erhöhung der unabhängigen Größe um eine Einheit verändert. Für die Berechnung der *y*- sowie der *yx*-standardisierten Koeffizienten wird die Standardabweichung des Logits der Mathematiknoten geschätzt.
Im Folgenden wird für die Interpretation der Ergebnisse auf die *yx*-standardisierten Koeffizienten zurückgegriffen. Diese haben gegenüber anderen Kennwerten (z. B. *odds ratios* oder teilstandardisierten Koeffizienten) den Vorteil, dass sie wie die Koeffizienten gewöhnlicher linearer Regressionen interpretiert werden können (Menard 2011).

5 Ergebnisse

Die Testleistung in Mathematik liegt im Mittel bei 533 Punkten (*SD* = 100) und weist einen Interquartilsabstand von 139 Punkten mit einer Untergrenze bei 464 Punkten und einer Obergrenze bei 603 Punkten auf (Schreiner und Breit 2014). Der Mittelwert von 533 zeigt an, dass die Testleistungen gegenüber der Baselinetestung im Schnitt um 33 Punkte gestiegen sind.
Im Unterschied zu den Testleistungen, die (qua Testkonstruktion) normalverteilt sind, sind die Noten wie erwartet rechtsschief verteilt: 43% erhielten im Halbjahreszeugnis die Note 1, 33% die Note 2, 17% die Note 3, 6% die Note 4 und 1% die Note 5.
Die einfache logistische Regression, die zur Ermittlung des Zusammenhangs zwischen den Testleistungen und den Noten in Mathematik berechnet wurde, weist einen standardisierten Koeffizienten in Höhe von $ß_{yx} = 0{,}659$ ($p < 0{,}001$) auf. Die Testleistungen erklären 43% der Varianz der Mathematiknoten.
Tab. 1 zeigt das Ergebnis der multiplen logistischen Regression, die berechnet wurde, um die Frage nach den systematischen Abweichungen der Noten von den Testleistungen beantworten zu können. Das standardisierte Gewicht der Testleistungen auf die Noten sinkt im Vergleich zur einfachen Regression etwas ab ($ß_{yx} = 0{,}612$, $p < 0{,}001$). Das Geschlecht und der Migrationshintergrund zeigen keine signifikanten Effekte, d. h. bezüglich beider Größen ergeben sich keine systematischen Abweichungen der Noten im Vergleich zu den Testleistungen. Anders sieht es für den sozioökonomischen Hintergrund aus ($ß_{yx} = 0{,}137$, $p < 0{,}001$). Schüler/innen aus Elternhäusern mit einem höheren sozioökonomischen Status (SES) erhalten bei gleichen Testleistungen signifikant bessere Noten als jene aus Elternhäusern mit einem niedrigeren SES. Alle vier Größen zusammen erklären 45% der Varianz der Mathematiknote.
Um die Intraklassenkorrelationen von Testleistungen und Noten vergleichen zu können, wurde ein *intercept-only model* (type = twolevel ohne Prädiktoren) berechnet. Die ausgegebenen Intraklassenkorrelationen liegen bei den Testleistungen bei 0,185 und bei den Noten bei 0,096. Die höhere Intraklassenkorrelation bei den Testleistungen indiziert, dass der Anteil an erklärter Varianz durch die Klasse dort wie erwartet höher ausfällt als bei den Noten, was als ein erstes Indiz für die Anwendung der sozialen Bezugsnorm durch die Lehrkräfte interpretiert werden kann.

Tab. 1 Regressionsgewichte einer multiplen logistischen Regression auf die abhängige Größe Halbjahresnote (*N* = 73.655)

	Unstandardisiert			yx-standardisiert		
Prädiktoren	*Est*	*SE*	*p*	*Est*	*SE*	*p*
Testleistung	0,015	0,000	0,000	0,612	0,004	0,000
Geschlecht	0,023	0,016	0,149	0,005	0,003	0,149
Sozioökonomischer Hintergrund (HISEI)	0,016	0,000	0,000	*0,137*	0,004	0,000
Migrationshintergrund	-0,005	0,028	0,853	-0,001	0,005	0,853
R^2	-	-	-	0,452	0,005	0,000

Auf die Dokumentation der thresholds wird verzichtet, diese können auf Anfrage durch die Autoren mitgeteilt werden

Tab. 2 Regressionsgewichte einer logistischen Mehrebenenanalyse (*random intercept model*): abhängige Größe Halbjahresnote (*N* = 73.655)

	Unstandardisiert			yx-standardisiert		
Prädiktoren	*Est*	*SE*	*p*	*Est*	*SE*	*p*
Within-level						
Testleistung	0,020	0,000	0,000	0,715	0,003	0,000
Geschlecht	-0,003	0,017	0,857	-0,001	0,003	0,857
Sozioökonomischer Hintergrund (HISEI)	0,015	0,000	0,000	0,109	0,003	0,000
Migrationshintergrund	-0,182	0,028	0,000	*-0,025*	0,004	0,000
Between-level						
Klassenmittelwert Testleistung	-0,011	0,000	0,000	*-0,530*	0,014	0,000
R^2 *within*	-	-	-	0,588	0,004	0,000
R^2 *between*	-	-	-	0,281	0,015	0,000

Auf die Dokumentation der thresholds wird verzichtet, diese können auf Anfrage durch die Autoren mitgeteilt werden

Um einen weiteren Hinweis auf die Anwendung der sozialen Bezugsnorm zu gewinnen, wurde im nächsten Schritt überprüft, ob die leistungsmäßige Zusammensetzung der Schulklasse einen Effekt auf die individuelle Note hat. Hierfür wurde ein *random-intercept model* (type = twolevel) berechnet (vgl. Tab. 2). Der Zusammenhang zwischen individuellen Testleistungen und Noten steigt bei Kontrolle der Leistungszusammensetzung der Klassen auf $ß_{yx} = 0{,}715$ ($p < 0{,}001$) an. Innerhalb von Klassen fällt der Zusammenhang zwischen Testleistungen und Noten demnach höher aus als über alle Klassen hinweg. Das Geschlecht zeigt keinen signifikanten Effekt, der HISEI hingegen wiederum einen positiven ($ß_{yx} = 0{,}109$, $p < 0{,}001$). Im Unterschied zum einfachen Regressionsmodell ergibt sich in der Mehrebenenanalyse ein signifikant negativer Effekt für den Migrationshintergrund ($ß_{y} = –0{,}025$, $p < 0{,}001$). Das bedeutet, dass innerhalb von Klassen Kinder mit Migrationshintergrund bei gleichen Testleistungen, gleichem sozioökonomischen Status und gleichem Geschlecht etwas schlechtere Noten erhalten als Kinder ohne Migrationshintergrund. Für die Leistungszusammensetzung der Klasse ergibt sich, wie erwartet, ein negativer Effekt auf die individuellen Noten ($ß_{yx} = –0{,}530$, $p < 0{,}001$). Kinder in Klassen mit einem hohen mittleren Leistungsniveau erhalten demnach signifikant

schlechtere Noten als Kinder in Klassen mit einem niedrigeren mittleren Leistungsniveau. Insgesamt erklären die Größen auf der Individualebene 59% der Varianz und auf Klassenebene 28%.

Tab. 3 Regressionsgewichte einer logistischen Mehrebenenanalyse (*random intercept model*) mit Interaktionseffekten: abhängige Größe Halbjahresnote (*N* = 73.655)

	Unstandardisiert			yx-standardisiert		
Prädiktoren	*Est*	*SE*	*p*	*Est*	*SE*	*p*
Within-level						
Testleistung	0,021	0,000	0,000	*0,754*	0,011	*0,000*
Zugehörigkeit unterstes Leistungsquartil in der Klasse	0,665	0,202	0,001	*0,105*	0,032	*0,001*
Zugehörigkeit oberstes Leistungsquartil in der Klasse	0,791	0,339	0,020	*0,118*	0,051	*0,020*
Testleistung x Zugehörigkeit unterstes Leistungsquartil	-0,001	0,000	0,002	*-0,090*	0,028	*0,002*
Testleistung x Zugehörigkeit oberstes Leistungsquartil	-0,002	0,001	0,006	*-0,147*	0,054	*0,006*
Geschlecht	-0,003	0,017	0,867	-0,001	0,003	0,867
Sozioökonomischer Hintergrund (HISEI)	0,015	0,000	0,000	*0,108*	0,003	*0,000*
Migrationshintergrund	-0,181	0,028	0,000	*-0,025*	0,004	*0,000*
Between-level						
Klassenmittelwert Testleistung	-0,011	0,000	0,000	*-0,545*	0,016	*0,000*
R^2 within	-	-	-	0,589	0,004	0,000
R^2 between	-	-	-	0,297	0,017	0,000

Auf die Dokumentation der thresholds wird verzichtet, diese können auf Anfrage durch die Autoren mitgeteilt werden

Die letzte Analyse galt der Frage, ob besonders leistungsstarke und besonders leistungsschwache Schüler/innen weniger leistungsadäquat benotet werden als jene mit einem mittleren Leistungsniveau. Hierfür wurden auf Grundlage der Testleistungen klassenspezifisch Quartile gebildet und für die Zugehörigkeit zum unteren wie zum oberen Leistungsquartil Interaktionsterme gebildet. Die Referenzkategorie für die Überprüfung der Interaktionsterme bilden Schüler/innen, die den beiden mittleren Leistungsquartilen angehören. Das Ergebnis (vgl. Tab. 3) wurde ebenfalls mittels einer Mehrebenenanalyse ermittelt, bei dem es sich um das *random-intercept model* aus Tab. 2 handelt, das lediglich um die Interaktionsterme erweitert wurde.
Zunächst kann festgehalten werden, dass sich die Effekte für das Geschlecht, den HISEI, den Migrationshintergrund und die mittlere Testleistung der Klassen gegenüber dem in Tab. 2 dargestellten Modell nur unwesentlich verändern. Für die individuellen Testleistungen ergibt sich jetzt ein signifikant positiver Zusammenhang zwischen Leistungstestwerten und Noten in Höhe von $\beta_{yx} = 0{,}754$ ($p < 0{,}001$). Dieser indiziert den Zusammenhang zwischen Testleistungen und Noten in den beiden mittleren Leistungsquartilen und liegt erwartungsgemäß über dem in Tab. 2 dokumentierten Zusammenhang für die Gesamtgruppe aller Schüler/innen. Die Regressionsgewichte für die Zugehörigkeit zum oberen und unteren Leistungsquartil indizieren, dass Schülerinnen und Schüler, die dem unteren Leistungsquartil angehören, bei Konstanthaltung aller anderen Größen (inklusive der Testleistung) bessere Noten erhalten als Schülerinnen und Schüler, die den mittleren

Leistungsquartilen angehören ($ß_{yx}$ = 0,105, p = 0,001); dasselbe gilt für Schülerinnen und Schüler des oberen Leistungsquartils im Vergleich zu jenen der mittleren Leistungsquartile ($ß_{yx}$ = 0,118, p = 0,020). Die Regressionsgewichte für die Interaktionsterme verdeutlichen das erwartete Ergebnis: Die Zugehörigkeit zum unteren Leistungsquartil geht mit einer signifikanten Verminderung des Zusammenhangs zwischen Testleistungen und Noten einher ($ß_{yx}$ = –0,090, p = 0,002), dasselbe gilt für die Zugehörigkeit zum oberen Leistungsquartil ($ß_{yx}$ = –0,147, p = 0,006).

6 Diskussion

Im vorliegenden Beitrag wurde der Zusammenhang zwischen der Mathematiktestleistung und der Mathematiknote auf Grundlage der Bildungsstandardüberprüfung 2013 in der 4. Klasse Volksschule analysiert. Dabei wurden Forschungsfragen zur leistungsangemessenen Notenvergabe durch Lehrkräfte fokussiert. Zum einen wurde untersucht, inwieweit sich systematische Abweichungen zwischen Noten und Testleistungen in Abhängigkeit vom Geschlecht, dem sozioökonomischen Status und dem Migrationshintergrund ergeben; zum anderen wurde analysiert, ob die Leistungszusammensetzung der Klassen bei der Benotung eine Rolle spielt. Schließlich wurde überprüft, ob Schüler/innen, die jeweils dem unteren oder oberen Leistungsquartil innerhalb von Klassen angehörten, weniger leistungsangemessen benotet wurden als jene der beiden mittleren Leistungsquartile.

Zunächst kann festgehalten werden, dass die Mathematikhalbjahresnote erwartungskonform eine rechtsschiefe Verteilung aufwies und somit einen Deckeneffekt bei der Benotung ausdrückte. Vor diesem Hintergrund lag die Vermutung nahe, dass an der Leistungsspitze die Noten die Testleistungen nicht hinreichend differenziert wiedergeben.

Insgesamt zeigte sich zwischen Note und Testleistung mit einem standardisierten logistischen Regressionsgewicht von über 0,60 ein etwas stärkerer Zusammenhang als für die deutsche IGLU-Studie berichtet wurde (Bos et al. 2004). Der Zusammenhang liegt jedoch im Rahmen dessen, was international in Untersuchungen gefunden wurde (Tent und Birkel 2010). Obwohl Lehrkräfte in der Regel nur den Bezugsrahmen ihrer eigenen Klasse sowie eventuell Parallelklassen oder ihre Erfahrungen mit vergangenen Klassen vor Augen haben, gelingt es ihnen recht gut, die Testleistungsrangfolge über diese Referenzgruppen hinweg abzubilden. Verbindliche Lehrpläne könnten hierfür eine wichtige Grundlage sein. Bei der Einschätzung der Stärke des Zusammenhangs ist zu berücksichtigen, dass Tests eine unterschiedliche curriculare Validität aufweisen, und diese in Bildungsstandardüberprüfungen höher sein dürfte als in international vergleichenden Schulleistungsstudien.

Mindernd auf den Zusammenhang zwischen Note und Testleistung haben sich die gefundenen Referenzgruppeneffekte ausgewirkt. Für deren Vorliegen ergaben sich zwei Hinweise: Erstens lag die Intraklassenkorrelation bei den Testleistungen höher als bei den Noten, was dahingehend interpretiert werden kann, dass die Leistungsunterschiede zwischen den Klassen durch das Anlegen einer sozialen Bezugsnorm bei der Benotung durch die Lehrkräfte minimiert wurden. Zweitens ergab sich ein signifikant negativer Effekt der Leistungszusammensetzung der Klassen auf die Noten. Dieser Kompositionseffekt kann ebenfalls auf die Anwendung einer sozialen Bezugsnorm durch Lehrkräfte zurückgeführt werden. Schüler/innen in leistungsstarken Klassen erhielten tendenziell schlechtere Noten im Vergleich zu jenen in leistungsstarken Klassen. Gleiche Noten spiegeln demnach nicht gleiche Leistungen wider.

Eine solche Benotungspraxis ist insofern kritisch zu betrachten, als die Mathematiknote des Halbjahreszeugnisses für den Übergang in die Sekundarstufe 1 ein wichtiges Kriterium darstellt. Für die Anmeldung an einer AHS werden in den drei Fächern Deutsch, Lesen und Mathematik die Semesternoten „Sehr gut“ oder „Gut“ benötigt. Ein „Befriedigend“ in einem der drei Fächer steht der Anmeldung zwar nicht prinzipiell entgegen, es bedarf in diesem Fall aber einer positiven Einschätzung der Leistungsfähigkeit durch die Schulkonferenz der Volksschule (Bergmüller und Steger 2009).

Vor dem Hintergrund des beschriebenen Deckeneffekts bei der Benotung lag die Vermutung nahe, dass die Testleistungsunterschiede im oberen Leistungsbereich nicht hinreichend differenziert durch Noten abgebildet werden. Mit Blick auf den unteren Leistungsbereich wurde angenommen, dass Lehrkräfte dazu tendieren, leistungsschwachen Schüler/inne/n ein negatives Feedback zu ersparen bzw. diese mit guten Noten zur Mitarbeit zu motivieren versuchen. Beide Phänomene konnten bestätigt werden: Im oberen wie im unteren Leistungsbereich ergaben sich schwächere Zusammenhänge zwischen Testleistungen und Noten im Vergleich zu den beiden mittleren Leistungsbereichen.

Ebenfalls mindernd auf den Zusammenhang zwischen Noten und Testleistungen wirken sich systematische Abweichungen der Noten in Abhängigkeit von bestimmten sozialen Merkmalen aus. Keine solchen Abweichungen fanden sich in Abhängigkeit vom Geschlecht. Zwar erzielten Buben eine um 14 Punkte bessere Testleistung als Mädchen (Schreiner und Breit 2014), aber bei gleicher Testleistung erhielten Buben wie Mädchen dieselben Noten.

Anders verhielt es sich mit dem sozioökonomischen Hintergrund. Ein höherer Sozialstatus wirkte sich nicht nur positiv auf die Testleistungen aus (Schreiner und Breit 2014), sondern führt bei gleichen Testleistungen zusätzlich auch zu besseren Noten. Inwieweit diese Abweichungen der Noten von den Testleistungen als problematisch einzustufen ist, kann auf Grundlage der durchgeführten Analysen nicht beurteilt werden. Die Abweichungen können sowohl leistungsbasiert (z. B. bessere mündliche Leistungen, engagiertere Mitarbeit von Schüler/inne/n aus höherer sozialer Schicht) als auch nicht leistungsbasiert (z. B. aufgrund von Interventionen durch Eltern, Urteilsverzerrungen der Lehrkräfte oder unter Einbezug der Antizipation von Konsequenzen, vgl. Lintorf 2012) zustande gekommen sein.

Der Migrationshintergrund wirkte sich nicht nur negativ auf die Testleistungen aus, vielmehr erhielten Kinder mit Migrationshintergrund innerhalb von Klassen bei gleichen Testleistungen auch schlechtere Noten. Auch hier kann auf Grundlage der durchgeführten Analysen nicht beurteilt werden, inwieweit die Abweichungen leistungsbasiert sind. Möglich wäre, dass Kinder mit Migrationshintergrund aufgrund mangelnder sprachlicher Fähigkeiten in der Unterrichtssprache schlechtere mündliche Leistungen zeigen als Kinder ohne Migrationshintergrund. Auch könnte es sein, dass kulturell bedingt die schulbezogenen Verhaltensweisen der Kinder mit Migrationshintergrund weniger gut an die Erwartungen der Lehrkräfte angepasst sind und sich dies negativ in den Noten niederschlägt (Hochweber et al. 2014).

Interessant ist, dass der Migrationshintergrund nur innerhalb von Klassen, d. h. erst unter Berücksichtigung der Klassenebene, einen negativen Effekt auf die Noten aufwies. Die Erklärung hierfür ist, dass Kinder mit Migrationshintergrund häufiger aus leistungsschwächeren Klassen stammten und Kinder aus diesen Klassen im Vergleich zu Kindern aus leistungsstärkeren Klassen bessere Noten erhielten. Der negative Effekt auf Klassenebene hebt sich dadurch auf der Ebene der Gesamtschülerschaft wieder auf.

Eine Limitation dieser Studie ist darin zu sehen, dass es sich bei den analysierten Mathematiknoten um Selbstberichte handelt. Fehlerhafte Angaben könnten insbesondere im unteren Leistungsbereich

eine Rolle spielen (Hochweber et al. 2014), sodass das Ergebnis des verringerten Zusammenhangs zwischen Testleistung und Note im unteren Leistungsquartil mit Vorsicht zu betrachten ist.
Offen bleibt, inwieweit der gefundene Deckeneffekt bei der Benotung kriterial gerechtfertigt ist. Vor dem Hintergrund, dass in der Bildungsstandardüberprüfung nur 12% der Schüler/innen die Kompetenzstufe 3 erreichten (Schreiner und Breit 2014), überrascht der hohe Anteil an sehr guten Noten (43%). Ein Vergleich der gesetzlichen Bestimmungen zur Notenvergabe mit der Beschreibung der Kompetenzstufen (vgl. Abschn. 1.3 und 1.4) lässt die Note „Sehr gut" für das Erreichen der Kompetenzstufe 2 (diese wurde von 65% der Schüler/innen erreicht) als eher nicht angemessen erscheinen.[6] Inwieweit die mündlichen Leistungen und die Mitarbeit im Unterricht die Abweichungen der Noten nach oben erklären können, bleibt fraglich und wäre aufzuklären.
Zu berücksichtigen ist jedoch, dass eine Verschiebung der Notenskala in Richtung schlechterer Noten gravierende Konsequenzen hätte. Beispielsweise wäre mit deutlich niedrigeren Übergangsquoten in die AHS oder mit einem erheblichen Mehraufwand für die Schulkonferenzen zu rechnen, um Schüler/inne/n mit einer Note 3 den Übergang zu ermöglichen. Solch niedrigere Übergangsquoten dürften sowohl den Interessen der Eltern als auch möglicherweise denen der aufnehmenden Schulen entgegenstehen. Inwieweit sie bildungspolitisch erwünscht wären, sei ebenfalls dahingestellt.
Wünschenswert wäre die Aufklärung der systematischen Abweichungen der Noten von den Testleistungen in Abhängigkeit von sozialer Schicht und dem Migrationshintergrund. Für weitergehende Analysen kämen als Indikatoren die Einschätzungen der mündlichen Leistungen sowie der Mitarbeit im Unterricht durch die Lehrkräfte in Frage. Auch die Qualität der Erledigung von Hausaufgaben, arbeitsbezogenes Verhalten sowie Interesse und Motivation für das Fach wären von Interesse. Bei Ditton und Krüsken (2006) erwiesen sich insbesondere die durch die Lehrkräfte wahrgenommenen sozial- und arbeitsstilbezogenen Kompetenzen der Schüler/innen als mediierende Faktoren für entsprechende Abweichungen.

Literatur

Baumert, J., Trautwein, U., & Artelt, C. (2003). Schulumwelten – institutionelle Bedingungen des Lehrens und Lernens. In Deutsches *PISA Konsortium (Hrsg.), PISA 2000 – Ein differenzierter Blick auf die Länder der Bundesrepublik Deutschland* (S. 261–331). Opladen: Leske + Budrich.

Bergmüller, S., & Steger, E. (2009). Gerechte Beurteilung? Leistungs- und Eignungsdiagnostik durch Lehrkräfte. In B. Suchan, C. Wallner-Paschon & C. Schreiner (Hrsg.), *PIRLS 2006. Die Lesekompetenz am Ende der Volksschule. Österreichischer Expertenbericht* (S. 202–218). Graz: Leykam.

Bos, W., Voss, A., Lankes, E.-M., Schwippert, K., & Thiel, O. (2004). Schullaufbahnempfehlungen von Lehrkräften für Kinder am Ende der vierten Jahrgangsstufe. InW. Bos, E.-M. Lankes, M. Prenzel, K. Schwippert, R. Valtin & G. Walther (Hrsg.), *IGLU. Einige Länder der Bundesrepublik Deutschland im nationalen und internationalen Vergleich* (S. 191–228). Münster: Waxmann.

Bundesministerium für Unterricht, Kunst und Kultur (Hrsg.). (2007). *Informationsblätter zum Schulrecht. Teil 3: Leistungsfeststellung und Leistungsbeurteilung*. Wien: Jugend & Volk.

von Davier, M., Gonzales, E., & Mislevy, R. J. (2009). What are plausible values and why are they useful? In IEA-ETS Research Institute (Hrsg.), *IERI Monograph Series Issues* and Methodologies in Large- Scale Assessments, (Bd. 2, S. 9–36). Hamburg, Princeton: IEA-ETS Research Institute.

Ditton, H., & Krüsken, J. (2006). Der Übergang von der Grundschule in die Sekundarstufe I. *Zeitschrift für Erziehungswissenschaft*, 9(3), 348–372.

[6] Von den Schüler/inne/n, welche die Kompetenzstufe 2 erreichten, erhielten 46% die Note 1, von denjenigen, welche die Stufe 1 erreichten, waren es 13 %.

Eder, F. (2003). Tests und Lehrerurteil. Wie gut stimmen externe Leistungstests mit Lehrereinstufungen überein? In E. J. Brunner, P. Noack, G. Scholz & I. Scholl (Hrsg.), *Diagnose und Intervention in schulischen Handlungsfeldern* (S. 125–140). Münster: Waxmann.

Eder, F. (2007). *Das Befinden von Kindern und Jugendlichen in der österreichischen Schule. Befragung 2005*. Innsbruck: Studien Verlag.

Eder, F., & Dämon, K. (2010). Leistungsvergleiche zwischen Hauptschule und AHS-Unterstufe. In F. Eder & G. Hörl (Hrsg.), *Schule auf dem Prüfstand* (S. 13–56). Wien: LIT.

Eid, M., Gollwitzer, M., & Schmitt, M. (2013). *Statistik und Forschungsmethoden*. Weinheim, Basel: Beltz.

Freunberger, R., Robitzsch, A., & Pham, G. (2014). *Hintergrundvariablen und spezielle Analysen*. Technische Dokumentation – BIST-Ü Mathematik, 4. Schulstufe, 2013. Salzburg: BIFIE.

Hochweber, J., Hosenfeld, I., & Klieme, E. (2014). Classroom composition, classroom management, and the relationship between student attributes and grades. *Journal of Educational Psychology*, 106(1), 289–300.

Ingenkamp, K.-H. (1995). *Die Fragwürdigkeit der Zensurengebung*. Weinheim: Beltz.

Köller, O. (2005). Bezugsnormorientierung von Lehrkräften: Konzeptuelle Grundlagen, empirische Befunde und Ratschläge für praktisches Handeln. In R. Vollmeyer & J.C. Brunstein (Hrsg.), *Motivationspsychologie und ihre Anwendungen* (S. 189–202). Stuttgart: Kohlhammer.

Lintorf, K. (2012). *Wie vorhersagbar sind Grundschulnoten? Prädiktionskraft individueller und kontext-spezifischer Merkmale*. Wiesbaden: VS.

Lüdtke, O., Robitzsch, A., Trautwein, U., & Köller, O. (2007). Umgang mit fehlenden Werten in der psychologischen Forschung: Probleme und Lösungen. *Psychologische Rundschau*, 58(2), 103–117.

Maaz, K., Baeriswyl, F., & Trautwein, U. (2013). Herkunft zensiert? Leistungsdiagnostik und soziale Ungleichheiten in der Schule. In D. Deißner (Hrsg.), *Chancen bilden. Wege zu einer gerechteren Bildung – ein internationaler Erfahrungsaustausch* (S. 184–341). Wiesbaden: Springer.

Menard, S. (2011). Standards for standardized logistic regression coefficients. *Social Forces*, 89(4), 1409–1428.

Rheinberg, F., & Fries, S. (2010). Bezugsnormorientierung. In D. H. Rost (Hrsg.), *Handwörterbuch Pädagogische Psychologie* (4. Aufl. S. 61–68). Weinheim, Basel: Beltz.

Rheinberg, F., & Vollmeyer, R. (2008). Motivationsförderung. In W. Schneider & M. Hasselhorn (Hrsg.), *Handbuch der Pädagogischen Psychologie* (S. 391–403). Göttingen: Hogrefe.

Sacher, W. (2009). *Leistungen entwickeln, überprüfen und beurteilen* (5. Aufl.). Bad Heilbrunn: Julius Klinkhardt.

Schreiner, C., & Breit, S. (Hrsg.). (2014). *Standardüberprüfung 2013 Mathematik, 4. Schulstufe. Bundesergebnisbericht*. Salzburg: BIFIE.

Streckeisen, U., Hänzi, D., Hungerbühler, A., & Tritten, S. (2006). Fördern und Auslesen: Deutungsmuster von Lehrpersonen zu einem beruflichen Handlungsproblem. In K.-S. Rehberg (Hrsg.), *Soziale Ungleichheit, kulturelle Unterschiede: Verhandlungen des 32. Kongresses der Deutschen Gesellschaft für Soziologie in München* (S. 4363–4372). Frankfurt a.M.: Campus.

Tent, L., & Birkel, P. (2010). Zensuren. In D. H. Rost (Hrsg.), *Handwörterbuch Pädagogische Psychologie* (S. 949–958). Weinheim, Basel: Beltz.

Trendtel, M. (2015). *Skalierung der Leistungsdaten und Linking zur Baseline-Erhebung*. Technische Dokumentation – BIST-Ü Mathematik, 4. Schulstufe, 2013. Salzburg: BIFIE.

Wallner-Paschon, C. (2009). Notengerechtigkeit bei der Risiko- und Spitzengruppe. In B. Suchan, C.Wallner-Paschon & C. Schreiner (Hrsg.), *PIRLS 2006. Die Lesekompetenz am Ende der Volksschule. Österreichischer Expertenbericht* (S. 45–51). Graz: Leykam.

Wilhelm, O., & Kunina-Habenicht, O. (2015). Pädagogisch-psychologische Diagnostik. In E. Wild & J. Möller (Hrsg.), *Pädagogische Psychologie* (S. 305–327). Berlin, Heidelberg: Springer.

Yen,W.M., & Fitzpatrick, A. R. (2006). Item response theory. In R.L. Brennon (Hrsg.), *Educational measurement* (S. 111–154). Westport: Praeger.