

Schmid, Christine

Asymmetrien zwischen Noten und Testleistungen am Ende der 8. Schulstufe? Eine Analyse auf Grundlage der österreichischen Bildungsstandardüberprüfung Mathematik 2012

formal und inhaltlich überarbeitete Version der Originalveröffentlichung in:

formally and content revised edition of the original source in:

Zeitschrift für Bildungsforschung 8 (2018) 3, S. 249-268



Bitte verwenden Sie in der Quellenangabe folgende URN oder DOI /
Please use the following URN or DOI for reference:

urn:nbn:de:0111-pedocs-169191
10.25656/01:16919

<https://nbn-resolving.org/urn:nbn:de:0111-pedocs-169191>

<https://doi.org/10.25656/01:16919>

Nutzungsbedingungen

Gewährt wird ein nicht exklusives, nicht übertragbares, persönliches und beschränktes Recht auf Nutzung dieses Dokuments. Dieses Dokument ist ausschließlich für den persönlichen, nicht-kommerziellen Gebrauch bestimmt. Die Nutzung stellt keine Übertragung des Eigentumsrechts an diesem Dokument dar und gilt vorbehaltlich der folgenden Einschränkungen: Auf sämtlichen Kopien dieses Dokuments müssen alle Urheberrechtshinweise und sonstigen Hinweise auf gesetzlichen Schutz beibehalten werden. Sie dürfen dieses Dokument nicht in irgendeiner Weise abändern, noch dürfen Sie dieses Dokument für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, aufführen, vertreiben oder anderweitig nutzen.

Mit der Verwendung dieses Dokuments erkennen Sie die Nutzungsbedingungen an.

Terms of use

We grant a non-exclusive, non-transferable, individual and limited right to using this document.

This document is solely intended for your personal, non-commercial use. Use of this document does not include any transfer of property rights and it is conditional to the following limitations: All of the copies of this documents must retain all copyright information and other information regarding legal protection. You are not allowed to alter this document in any way, to copy it for public or commercial purposes, to exhibit the document in public, to perform, distribute or otherwise use the document in public.

By using this particular document, you accept the above-stated conditions of use.

Kontakt / Contact:

peDOCS
DIPF | Leibniz-Institut für Bildungsforschung und Bildungsinformation
Informationszentrum (IZ) Bildung
E-Mail: pedocs@dipf.de
Internet: www.pedocs.de

Mitglied der


Leibniz-Gemeinschaft

Christine Schmid

Asymmetrien zwischen Noten und Testleistungen am Ende der 8. Schulstufe? Eine Analyse auf Grundlage der österreichischen Bildungsstandardüberprüfung Mathematik 2012

Korrespondierende/r Autor/in:

PD Dr. Christine Schmid

DIPF | Leibniz-Institut für Bildungsforschung und Bildungsinformation

Abteilung BIQUA

Rostocker Straße 6

60323 Frankfurt am Main

Tel.: +49 (0)69 24708-162

Email: christine.schmid@dipf.de

Asymmetrien zwischen Noten und Testleistungen am Ende der 8. Schulstufe? Eine Analyse auf Grundlage der österreichischen Bildungsstandardüberprüfung Mathematik 2012

Zusammenfassung

Die österreichische Bildungsstandardüberprüfung für das Fach Mathematik 8. Schulstufe aus dem Jahr 2012 (BIST-Ü M8 2012) erlaubt als Vollerhebung (N = 79.678) einen Vergleich von Noten und Testleistungen nicht nur auf der Populationsebene, sondern auch auf der Ebene von Klassen und Lerngruppen. Regressions- und Mehrebenenanalysen zeigen, dass Schüler/innen der 1. Leistungsgruppe Hauptschule bei gleichen Testleistungen deutlich bessere Noten erhalten als Schüler/innen der Allgemeinbildenden höheren Schulen (AHS). Außerdem erzielen Mädchen, Kinder aus Elternhäusern mit einem höheren Sozialstatus sowie Schüler/innen mit Migrationshintergrund bei gleichen Testleistungen bessere Noten als Jungen, Kinder aus Elternhäusern mit einem niedrigeren Sozialstatus und als Schüler/innen ohne Migrationshintergrund. Die Ergebnisse werden im Hinblick auf mögliche zugrundeliegende Mechanismen diskutiert. Sie spiegeln den Stand der Benotungspraxis vor der flächendeckenden Umwandlung der Hauptschulen in Neue Mittelschulen wieder.

Schlüsselwörter

Noten, Bildungsstandards, leistungsäquivalente Notenvergabe, soziale Bezugsnorm, Mehrebenenanalyse, Sekundarstufe I

Asymmetries between grades and test performances at the end of 8th grade? An analysis based on the Austrian educational standards assessment in Mathematics 2012

Abstract

The Austrian educational standards assessment for 8th grade Mathematics in 2012 (BIST-Ü M8 2012), as a complete survey (N = 79,678), allows to compare grades with test performances not only at the population level, but also at the level of classes or learning groups. Corresponding regression and multi-level analyses show that students in the 1st performance group “Hauptschule” receive significantly better marks for the same test performance than students of the “Allgemeinbildende Höhere Schule” (AHS). In addition, girls, children from households with higher social status, and students with an immigrant background achieve better grades than boys, children from households with lower social status, and students without an immigrant background. The results are discussed with regard to possible underlying mechanisms. They reflect the state of grading practice before the nation-wide conversion of secondary schools into “Neue Mittelschulen” (NMS).

Keywords

Grades, educational standards, performance-equivalent grading, social reference orientation, multi-level analysis, lower secondary education

1. Einleitung

Noten haben in modernen Schulsystemen verschiedene Funktionen. Sie dienen Eltern und Schüler/innen als Feedback über die gezeigten Leistungen, gehen aber auch mit bestimmten

Berechtigungen einher. Vor allem aufgrund der Berechtigungsfunktion besteht in meritokratischen Schulsystemen die Forderung nach einer leistungsäquivalenten Benotung, d.h. gleiche Noten sollen vergleichbare Leistungen ausdrücken (Fend, 2006).

Internationale Schulleistungsstudien (z.B. TIMSS, PIRLS und PISA) ermöglichen es, den Zusammenhang von Noten und Testleistungen über Schulen und Schulzweige hinweg zu untersuchen (Tent & Birkel, 2010). In der Regel werden in diesen Studien jedoch keine ganzen Klassen erhoben, so dass es nicht möglich ist, Aussagen über den Zusammenhang von Noten und Testleistungen auf der Ebene von Klassen zu treffen. Österreichische Bildungsstandardüberprüfungen sind im Unterschied zu internationalen Schulleistungsstudien als Vollerhebungen konzipiert, was die Modellierung der Klassenebene ermöglicht.

Bei der Untersuchung des Zusammenhangs von Noten und Testleistungen rücken zwei Merkmalsgruppen in den Fokus, die systematische Abweichungen zwischen Noten und Testleistungen verursachen können. Zum einen sind dies organisatorische Bedingungen des Lernens, wie etwa die Einteilung der Lernenden in Klassen, Leistungsgruppen oder Schularten. Solche Einteilungen können intendiert, wie im Falle von Schulzweigen oder Leistungsgruppen über unterschiedliche Curricula, oder ungewollt, wie etwa bei einer sozialen Bezugsnormorientierung von Lehrkräften (Köller, 2005), für eigene Referenzrahmen bei der Benotung sorgen. Zur zweiten Gruppe gehören Merkmale wie das Geschlecht, der Migrationshintergrund oder die Bildung der Eltern bzw. deren soziale Stellung, die entweder notenrelevante Verhaltensweisen der Schüler/innen beeinflussen oder unbewusst, beispielsweise in Form von Stereotypen, in die Notengebung mit einfließen können.

1.1 Asymmetrien zwischen Noten und Testleistungen unter messtheoretischen Gesichtspunkten

Unter messtheoretischen Gesichtspunkten wie Objektivität, Reliabilität und Validität haben Noten und Testleistungen unterschiedliche Stärken und Schwächen (Hochweber, 2010). Schulnoten, insbesondere Zeugnisnoten, basieren in der Regel auf mehreren Leistungsbeurteilungen, die über einen längeren Zeitraum (z.B. ein ganzes Halbjahr) ermittelt werden. Noten sind dadurch breiter fundiert und reliabler als einmalig durchgeführte Tests. Zudem nutzen Lehrkräfte bei der Notenfeststellung verschiedene Methoden der Leistungserhebung, insbesondere auch mündliche Formen. Tests im Rahmen von internationalen Schulleistungsstudien oder Bildungsstandardüberprüfungen werden häufig zu einem bestimmten Zeitpunkt, nur schriftlich und überwiegend unter Nutzung von Multiple-Choice-Formaten durchgeführt. Testergebnisse sind daher ebenfalls mit Messfehlern behaftet und d.h. nur eingeschränkt reliabel.

Dagegen kann Testauswertungen eine hohe Objektivität bescheinigt werden, da sie in der Regel hochgradig standardisiert und anonym erfolgen. Demgegenüber berücksichtigen Lehrkräfte bei der Beurteilung von Leistungen stärker den Kontext der Leistungserbringung. In zahlreichen Experimenten konnte zudem nachgewiesen werden, dass die Notenvergabe durch Lehrkräfte nicht frei von deren Erwartungen, Vorurteilen und Stereotypen erfolgt (Tent & Birkel, 2010; Wilhelm & Kunina-Habenicht, 2015).

1 Bildungsstandüberprüfungen beziehen sich in der Regel auf ein konzeptionell gut
2 begründetes, intendiertes Curriculum, weshalb der zu diesem Zweck entwickelte Test als
3 curriculumsvalide gelten kann. Noten dagegen sollen die Erfüllung von unterrichtsbezogenen
4 Anforderungen widerspiegeln und dürften von daher näher am realisierten Curriculum
5 liegen, sie gelten im Allgemeinen als unterrichtsvalider (Tent & Birkel, 2010).
6

7 **1.2 Abweichung der Noten von Testleistungen durch verschiedene** 8 **Bezugsnormorientierungen** 9

10 Mit Noten werden im Unterschied zu Testleistungen häufig auch pädagogische Zielsetzungen
11 verfolgt. Gute Noten können zur Motivierung, schlechte Noten als Disziplinierungsinstrument
12 eingesetzt werden. Mit pädagogischen Zielsetzungen eng verbunden ist die *individuelle* oder
13 *ipsative* Bezugsnorm (Köller, 2005). Unabhängig vom Ausgangsniveau bzw. dem absoluten
14 Niveau der gezeigten Leistungen werden hier gute Noten vergeben, wenn Schüler/innen
15 Leistungsverbesserungen zeigen. Solch eine Benotungspraxis kann pädagogisch sinnvoll sein,
16 ist aber kaum mit der Forderung nach einer leistungsäquivalenten Notenvergabe zu
17 vereinbaren. Dieser Forderung entspricht am ehesten die Orientierung an einer *kriteriellen*
18 Bezugsnorm.
19

20 Von einer *kriteriellen* Bezugsnorm wird gesprochen, wenn die Noten an feste, allgemeingültig
21 definierte Leistungskriterien gebunden sind. Um in der Praxis wirksam werden zu können,
22 müssen die Kriterien allerdings hinreichend konkret beschrieben sein. Zumeist jedoch sind
23 Leistungskriterien für Noten eher abstrakt formuliert, so dass Lehrkräften viel Spielraum für
24 Interpretationen bleibt.
25

26 Im Bestreben dennoch leistungsäquivalent zu benoten, wird häufig auf die *soziale*
27 Bezugsnorm zurückgegriffen (Köller, 2005). Die *soziale* Bezugsnorm hat den Vorteil, dass
28 die Rangreihe der Leistungen in der zu beurteilenden Gruppe (meistens die Klasse)
29 berücksichtigt wird. Häufig jedoch erfolgt eine Anpassung des Notenspektrums an das
30 Leistungsniveau der Lerngruppe. Dies hat zur Folge, dass innerhalb der Klasse bzw.
31 Lerngruppe die Leistungsäquivalenz gewährleistet ist, jedoch nicht über diese hinaus.
32

33 **1.3 Relevanz der Notengebung in der 8. Schulstufe** 34

35 Das österreichische Schulsystem weist im Unterschied zum deutschen nach der 8. Schulstufe
36 einen zweiten Übergang auf. Die Sekundarstufe I wird ab der 5. Schulstufe zweigliedrig
37 geführt. Neben der Unterstufe der „Allgemeinbildenden Höheren Schule“ (AHS, i.d.R. sind
38 dies Gymnasien) gibt es den Zweig der „Allgemeinbildenden Pflichtschule“ (APS), der zum
39 Zeitpunkt der hier zugrunde gelegten Bildungsstandardüberprüfung, d.h. im Schuljahr
40 2011/12, aus der „Hauptschule“ (HS) und der Allgemeinen Sonderschule (ASO) bestand. In
41 HS werden die Fächer Deutsch, Mathematik und Englisch in drei Leistungsgruppen
42 unterrichtet, wobei das curriculare Niveau der 1. Leistungsgruppe demjenigen der AHS
43 entsprechen soll. Schüler/innen der AHS können bei erfolgreichem Abschluss der 8. Klasse
44 ohne weiteres in die Oberstufe der AHS wechseln. Diese nimmt dann weitere vier Jahre in
45 Anspruch und endet nach der 12. Klasse mit der „Matura“.
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

Neben der AHS-Oberstufe gibt es die „Berufsbildenden höheren Schulen“ (BHS), die sowohl einen berufsbildenden Abschluss ermöglichen als auch im Rahmen eines 13. Schuljahres den Erwerb der allgemeinen Hochschulreife. Schüler/innen aus HS können ohne Aufnahmeprüfung in die Oberstufe der AHS oder in die BHS wechseln, sofern sie in allen drei Pflichtgegenständen (Mathematik, Deutsch, Englisch) in der 1. Leistungsgruppe positiv beurteilt wurden oder in der 2. Leistungsgruppe keine schlechtere Note als „gut“ aufweisen. Für die anderen Schüler/innen besteht die Möglichkeit, eine Aufnahmeprüfung zu absolvieren oder in eine „Berufsbildende mittlere Schule“ (BMS) zu wechseln. Für die Aufnahme in eine BMS ist mindestens ein „befriedigend“ in allen drei Pflichtgegenständen in der 2. Leistungsgruppe Voraussetzung. Neben der BMS besteht noch die Möglichkeit, das 9. Schuljahr in einer „Polytechnischen Schule“ zu verbringen. Diese bereitet auf den Übergang in eine Lehre und den Besuch der Berufsschule vor und kann noch ein zweites Jahr besucht werden, wenn keine Lehrstelle gefunden wurde.

Im Zeugnis der 8. Schulstufe werden neben den Noten auch der Schulzweig und die Leistungsgruppe vermerkt, in der die Note erzielt wurde. In welcher Relation die Noten eines Faches über die Leistungsgruppen hinweg stehen sollen, ist im Schulunterrichtsgesetz nicht eindeutig geregelt. Um eine über Leistungsgruppen hinweg vergleichbare Beurteilungsskala zu erzeugen, wird den Noten der 2. Leistungsgruppe üblicherweise der Wert 2, und Noten der 3. Leistungsgruppe der Wert 4 hinzu addiert. Diese Umrechnung entspricht der notengebundenen Umstufungspraxis (Bruneforth, Lassnigg, Vogtenhuber, Schreiner & Breit, 2016).

1.4 Befunde zum Zusammenhang von Noten und Testleistungen

In Deutschland repräsentieren Noten die Testleistungen innerhalb von Schulformen besser als über Schulformen hinweg. Beispielsweise lag in Auswertungen auf Grundlage von PISA 2000 (Baumert, Trautwein & Artelt, 2003) die Korrelation zwischen der Leistung in einem curriculumsnahen Mathematiktest und der Mathematiknote innerhalb der Schulformen bei $r = -.46$, über alle Schulformen hinweg dagegen nur bei $r = -.32$. Zudem erklärte die Schulform 37 % der Varianz des Mathematikleistungstests, aber nur 0,2 % der Varianz der Mathematiknoten. Schulformen bilden demnach jeweils einen eigenen Referenzrahmen für die Notengebung, was dazu führt, dass sich Noten über Schulformen hinweg ähnlicher sind als Testleistungen.

Entsprechendes lässt sich für Österreich zeigen. Beispielsweise ergaben sich auf Grundlage der Pilotuntersuchung zu den Bildungsstandardüberprüfungen der 8. Schulstufe nahezu deckungsgleiche Notenverteilungen für die verschiedenen Leistungsgruppen (AHS sowie 1. bis 3. Leistungsgruppe HS; Eder & Dämon, 2010). Dieselben Noten drückten demnach je nach Leistungsgruppe ein anderes Testleistungsniveau aus (vgl. hierzu auch die Auswertungen im Rahmen des Nationalen Bildungsberichts: Bruneforth et al., 2016, S. 115).

Relativ hoch fiel der Zusammenhang zwischen Mathematiknoten und Testleistungen auf der Grundlage einer repräsentativen Stichprobe von 841 österreichischen Schüler/innen der 7. und

8. Klassen unter Anwendung der o.g. Umrechnungsformel¹ aus ($r = -.69$; Eder, 2003). In der genannten Untersuchung zeigten sich darüber hinaus bedeutsame Abweichungen der Noten von den Testleistungen: Schüler/innen der AHS erhielten ähnliche Noten wie jene der 1. Leistungsgruppe HS, wiesen im Vergleich zu diesen aber im Schnitt deutlich schlechtere Testleistungen auf. Dieser Sachverhalt wird auch durch Auswertungen im Rahmen des Nationalen Bildungsberichts bestätigt (Bruneforth et al., 2016).

In der Primarstufe, in der nicht leistungsdifferenziert unterrichtet wird, fallen Zusammenhänge zwischen Noten und Testleistungen im Allgemeinen höher aus als in der Sekundarstufe. Beispielsweise fanden sich in der deutschen IGLU-Studie Korrelationen zwischen der Mathematiknote und dem Mathematiktest in Höhe von $r = -.55$ sowie zwischen der Deutschnote und dem Lesetest in Höhe von $r = -.56$ (Bos, Voss, Lankes, Schwippert & Thiel, 2004). Für die österreichische Primarstufe ergab sich auf Grundlage der Bildungsstandarderhebung Mathematik 4. Klasse (BIST-Ü M4) im Jahr 2013 ein Zusammenhang zwischen umgepolter Note und Testleistung von immerhin $\log \beta = .66$ (Schmid, Paasch & Katstaller, 2016).

1.5 Systematische Abweichungen in Abhängigkeit von Geschlecht, sozialem Status und Migrationshintergrund

Systematische Abweichungen von Noten gegenüber den Testleistungen können ihre Ursache sowohl in einer mangelnden Objektivität von Lehrkräften haben, als auch auf der legitimen Berücksichtigung von Leistungsmerkmalen wie etwa der mündlichen Mitarbeit im Unterricht beruhen. Analysen für die Sekundarstufe I zeigen, dass weibliche Schülerinnen sowie Schüler/innen aus sozial begünstigten Elternhäusern im Fach Mathematik bei gleichen Testleistungen häufig besser bewertet werden als männliche Schüler und als Schüler/innen aus sozial schlechter gestellten Elternhäusern (Hochweber, 2010; Schöps, Walter, Zimmer & Prenzel, 2006). In einer repräsentativen österreichischen Studie für die Sekundarstufe I erzielten Jungen die besseren Testleistungen, Mädchen aber erhielten die besseren Noten (Eder, 2003).

In Auswertungen für die Sekundarstufe II (TOSCA; Maaz, Baeriswyl & Trautwein, 2013) ergaben sich für den Migrationshintergrund entweder negative Abweichungen der Noten von den Testleistungen oder keine systematischen Verzerrungen. Hochweber (2010) berichtet auf Grundlage von VERA-GU für die Primarstufe negative, auf Grundlage der MARKUS-Studie für die Sekundarstufe I aber positive Abweichungen in Abhängigkeit von einer nichtdeutschen Herkunftssprache.

Einen differenzierten Befund berichten Schmid et al. (2016) auf Grundlage der österreichischen BIST-Ü M4. Für die Gesamtpopulation aller Primarschüler/innen ergaben sich keine systematischen Abweichungen der Noten von den Testleistungen. Eine Mehrebenenanalyse zeigte hingegen, dass Kinder mit Migrationshintergrund innerhalb von Klassen bei gleichen Testleistungen signifikant schlechtere Noten erhielten als ihre Mitschüler/innen ohne Migrationshintergrund. Der auf den ersten Blick widersprüchlich

¹ Abweichend von der allgemein üblichen Formel wurden hier +2 Noten in der 2. Leistungsgruppe und +3 Noten in der 3. Leistungsgruppe addiert.

erscheinende Befund ließ sich dadurch erklären, dass Kinder mit Migrationshintergrund häufiger in den leistungsschwächeren Klassen zu finden waren. In diesen war es aufgrund der Anwendung der sozialen Bezugsnorm leichter, gute Noten zu erzielen. Der Befund verdeutlicht, dass Mehrebenenanalysen zu anderen Ergebnissen führen können als einfache Regressionsanalysen.

2. Forschungsfragen und Vorgehensweise

Die im Jahr 2012 vom Bundesinstitut für Bildungsforschung, Innovation und Entwicklung des österreichischen Schulwesens (BIFIE) durchgeführte Bildungsstandardüberprüfung Mathematik 8. Schulstufe (Schreiner & Breit, o. J.) ermöglicht als Vollerhebung zum einen den Vergleich von Noten und Testleistungen über die beiden Schulzweige (AHS und APS) und die drei Leistungsgruppen der HS hinweg, sowie zum anderen eine mehrebenenanalytische Betrachtung unter Berücksichtigung von Klassen bzw. Lerngruppen. Im Schuljahr 2011/12 wurde Mathematik in der HS noch in klassenübergreifend gebildeten, leistungsdifferenzierten Lerngruppen unterrichtet.

2.1 Verteilung von Noten und Testleistungen

In einem ersten Schritt sollen deskriptiv die Verteilungen der Noten innerhalb der Schulzweige und Leistungsgruppen dargestellt werden. Vor dem Hintergrund früherer Auswertungen für die Sekundarstufe I (z.B. Eder & Dämon, 2010) ist zu erwarten, dass die Noten innerhalb der Schulzweige jeweils annähernd normal verteilt sind. Dasselbe kann für die Verteilungen der Noten innerhalb der Leistungsgruppen angenommen werden, wobei sich hier die Möglichkeiten zur Auf- und Abstufung bemerkbar machen könnten. Zu erwarten sind weniger „nicht genügend“ in der 1. Leistungsgruppe HS, weniger „sehr gut“ und „nicht genügend“ in der 2. Leistungsgruppe HS sowie weniger „sehr gut“ in der 3. Leistungsgruppe HS.

2.2 Einfache Zusammenhänge zwischen Noten und Testleistungen

Im zweiten Schritt sollen einfache Zusammenhänge zwischen Noten und Testleistungen untersucht werden. Anzunehmen ist, dass der Zusammenhang zwischen Noten und Testleistungen in der Sekundarstufe I, in der aufgrund der gegliederten Schulstruktur unterschiedliche curriculare Referenzrahmen für die Benotung zugrunde liegen, niedriger ausfällt als in der Primarstufe. Darüber hinaus sollte sich der Zusammenhang zwischen Noten und Testleistungen in der Sekundarstufe I erhöhen, wenn statistisch für die Schulform und die Leistungsgruppen, d.h. für die unterschiedlichen Referenzrahmen kontrolliert wird. Dasselbe ist zu erwarten, wenn die Umrechnungsformel (+2 Noten in der 2. Leistungsgruppe, +4 Noten in der 3. Leistungsgruppe) zur Vergleichbarmachung der Noten über Leistungsgruppen hinweg berücksichtigt wird. Außerdem ist zu erwarten, dass der Zusammenhang zwischen Noten und Testleistungen innerhalb des APS-Zweiges, in dem mit den drei Leistungsgruppen der HS auch drei verschiedene Referenzrahmen bei der Benotung zugrunde liegen, schwächer ausfällt als innerhalb des AHS-Zweiges, wo nur ein Bewertungsmaßstab gilt. Innerhalb der drei Leistungsgruppen der HS hingegen sollten die Zusammenhänge jeweils höher liegen als im APS-Zweig insgesamt, da innerhalb der Leistungsgruppen Noten jeweils vergleichbare Leistungen ausdrücken sollten.

2.3 Systematische Abweichungen in Abhängigkeit von Geschlecht, sozialem Status und Migrationshintergrund

Im dritten Schritt steht die Frage nach systematischen Abweichungen der Noten von den Testleistungen in Abhängigkeit vom Geschlecht, dem sozialen Status der Eltern und dem Migrationshintergrund im Mittelpunkt. Hierfür werden multiple logistische Regressionen berechnet, die neben den Testleistungen entsprechende Prädiktoren für diese drei Größen enthalten. Zu erwarten ist, dass männliche Schüler bei gleichen Testleistungen schlechtere Noten erzielen als weibliche Schülerinnen, und dass Kinder aus Elternhäusern mit höherem Sozialstatus bei gleichen Testleistungen bessere Noten erzielen als Kinder aus Elternhäusern mit niedrigerem Sozialstatus. Hinsichtlich der Frage, ob Schüler/innen mit Migrationshintergrund bei gleichen Testleistungen bessere oder schlechtere Noten erhalten, war die Forschungslage uneinheitlich. Hier könnten außerdem – wie für die Primarstufe dargestellt – gewöhnliche Regressionen zu anderen Ergebnissen führen als Mehrebenenanalysen.

2.4 Mehrebenenanalysen

Im letzten Schritt werden Mehrebenenanalysen berechnet, mit denen zwei unterschiedliche Ziele verfolgt werden. Das erste Ziel besteht in der Überprüfung systematischer Abweichungen der Noten von den Testleistungen in Abhängigkeit von sozialen Hintergrundvariablen auf der Ebene von Klassen bzw. Lerngruppen. Von Effekten auf der Klassen- oder Lerngruppenebene kann direkt auf das Handeln von Lehrkräften geschlossen werden, während Abweichungen auf der Ebene von Gesamt- oder Teilpopulationen auch einer unterschiedlichen Zusammensetzung der Schülerschaft in Abhängigkeit vom entsprechenden Merkmal geschuldet sein können.

Das zweite Ziel besteht in der Überprüfung von Referenzgruppeneffekten im Sinne der Anwendung der sozialen Bezugsnorm durch Lehrkräfte. Im AHS-Zweig können soziale Bezugsnormeffekte vor allem durch die Anpassung des Notenspiegels an das mittlere Leistungsniveau der Klasse entstehen. In der HS bilden die klassenübergreifend gebildeten Lerngruppen, die jeweils einer bestimmten Leistungsgruppe zugeordnet sind, den Bezugsrahmen für die Benotung. Da jeder Leistungsgruppe ein eigener Bewertungsmaßstab zugrunde liegt, ist es sinnvoll, die Mehrebenenanalysen getrennt für die Leistungsgruppen zu berechnen.

Zu erwarten ist ein negativer Effekt des mittleren Leistungsniveaus der Klassen bzw. der Lerngruppen auf die Noten. Dieser indiziert, dass es in Klassen bzw. Lerngruppen mit einem höheren durchschnittlichen Leistungsniveau schwieriger ist, eine gute Note zu erzielen als in Klassen bzw. Lerngruppen mit einem niedrigeren durchschnittlichen Leistungsniveau. Außerdem ist zu erwarten, dass sich in Mehrebenenanalysen *innerhalb* von Klassen bzw. Lerngruppen höhere Zusammenhänge zwischen Noten und Testleistungen ergeben als jeweils für die Gesamtheit aller Schüler/innen bzw. in der entsprechenden Leistungsgruppe.

3. Methode

3.1 Daten

Die Zielpopulation der Bildungsstandardüberprüfung Mathematik 8. Klasse von 2012 bilden alle Schüler/innen der 8. Schulstufe in den beiden Schulzweigen APS und AHS (N = 86.868). Von der Testung ausgeschlossen wurden Schüler/innen mit sonderpädagogischem Förderbedarf (SPF) sowie mit außerordentlichem Status (Schreiner & Breit, o. J.). Insgesamt wurden N = 79.678 Schüler/innen getestet, was einer Ausschöpfung von 91,7 % entspricht. Um für systematische Ausfälle zu kompensieren, wurden Schüler/innengewichte erzeugt (George, Itzlinger-Bruneforth & Trendtel, 2015).

3.2 Instrumente

Mathematiktest: Die Testung erfolgte im Mai 2012 im „paper-and-pencil“-Verfahren mittels eines Multi-Matrix-Designs (Schreiner & Breit, o. J.). Die Testzeit betrug 90 Minuten. Inhaltlich waren die Aufgaben entlang dreier Dimensionen konstruiert: 1. Inhaltsbereiche (Zahlen und Maße, Variable, funktionale Abhängigkeiten; geometrische Figuren und Körper; statistische Darstellungen und Kenngrößen), 2. Handlungsbereiche (Darstellen, Modellbilden; Rechnen, Operieren; Interpretieren; Argumentieren, Begründen) und 3. Komplexitätsbereiche (Einsetzen von Grundkenntnissen und -fertigkeiten; Herstellen von Verbindungen; Einsetzen von Reflexionswissen, Reflektieren). Die Items, die überwiegend ein Multiple-Choice-Format aufwiesen, wurden einer IRT-Skalierung (Yen & Fitzpatrick, 2006) unterzogen und auf einer Metrik verankert, die an der Baselinetestung von 2009 orientiert war. Seit dem Jahr 2009 hat sich die Testleistung der Schüler/innen im Mittel verbessert und lag im Jahr 2012 bei 535 Testpunkten mit einer Standardabweichung von 94. Die Testwerte liegen in Form von 10 *plausible values* vor (von Davier, Gonzales & Mislevy, 2009).

Mathematiknote: Die Mathematiknote wurde über einen zusätzlich zum Test eingesetzten 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, wurden die Noten für die Analysen umgepolt, so dass hohe Werte eine bessere Note indizieren.²

Leistungsgruppe: Die Leistungsgruppe wurde ebenfalls über den Schüler/innenfragebogen erhoben, gefragt wurde nach der Leistungsgruppe, die im Halbjahreszeugnis zusammen mit der Note ausgewiesen war. Ein kleiner Teil der Schüler/innen im APS-Zweig besuchte zum Zeitpunkt der Erhebung eine von 67 Modellschulen „Neue Mittelschule“ (NMS), in denen nicht leistungsdifferenziert unterrichtet wurde. Für die Analysen wurden diese Schüler/innen in einer 4. Gruppe NMS zusammengefasst. Die Verteilung der Schüler/innen über die verschiedenen Leistungsgruppen (LG) sah wie folgt aus: 33 % AHS, 30 % 1. LG HS, 23 % 2. LG HS, 12 % 3. LG HS und 3 % 4. Gruppe NMS (vgl. auch Tabelle 1).

Geschlecht: Das Geschlecht der Schüler/innen wurde gleichfalls über den Schüler/innenfragen erhoben und weist die Kategorien 0 = weiblich und 1 = männlich auf. Von den getesteten Schüler/innen sind 51 % männlich.

Sozialer Status: Der soziale Status der Eltern wurde über den HISEI abgebildet, der auf der Grundlage der Angaben zum Beruf beider Elternteile im Schüler/innenfragebogen ermittelt wurde. Kodiert wurden die Berufe nach dem Schema ISCO-08. Für die Bildung des HISEI

² In Abbildung 1 wurden die Originalwerte beibehalten.

wurde der jeweils höhere Wert von Mutter oder Vater bzw. der einzig vorhandene Wert herangezogen. Der HISEI weist einen Mittelwert von 48,91 ($SD = 20,76$) auf.

Migrationshintergrund: Die Angaben zum Migrationshintergrund stammen ebenfalls aus dem Schüler/innenfragebogen. Die Variable hat die Ausprägungen 0 = kein Migrationshintergrund, d.h. kein oder nur ein Elternteil wurde im Ausland geboren, und 1 = Migrationshintergrund, d.h. beide Elternteile wurden im Ausland geboren. Deutschland wurde aufgrund der Ähnlichkeit der Sprache nicht als Ausland gewertet. Ein Anteil von 18 % der Schüler/innen weist einen Migrationshintergrund auf.³

3.3 Umgang mit fehlenden Werten

Um sicherzustellen, dass es in den Analysen zu keinen Verzerrungen durch fehlende Werte kommt, wurde auf Datensätze mit imputierten Werten zurückgegriffen. Diese Datensätze wurden am BIFIE mithilfe des Verfahrens der multiplen Imputation (10 Datensätze) in R erzeugt. Eine genaue Beschreibung der Vorgehensweise findet sich bei Robitzsch, Pham und Yanagida (2016).

3.4 Analysen

Die Analysen wurden mit der Statistiksoftware Mplus unter Anwendung der Schüler/innengewichte durchgeführt. Den Analysen für die Gesamtpopulation liegen alle zehn imputierten Datensätze zugrunde. Für die getrennt nach Leistungsgruppen berechneten Analysen konnte nur jeweils ein imputierter Datensatz berücksichtigt werden,⁴ was mit Einbußen bezüglich der Zuverlässigkeit der geschätzten Parameter einhergeht.

Da es sich bei der Note um eine Messung mit ordinalem Niveau handelt, wurden die Zusammenhangsanalysen in Form von logistischen Regressionen berechnet. In den Analysen ohne die Berücksichtigung der Mehrebenenstruktur wurde die geschachtelte Struktur der Daten jeweils mittels der Option *type = complex* berücksichtigt. Als Clustervariable diente die Angabe zur Klasse bzw. Lerngruppe.

Bei den berechneten Mehrebenenanalysen handelt es sich um sogenannte *random slope* Modelle. Auf der Individualebene wurden die Noten, die Testleistungen, das Geschlecht und die sozialen Hintergrundmerkmale der Schüler/innen berücksichtigt, auf der Ebene der Klassen bzw. Lerngruppen wurden über alle Schüler/innen der Klasse bzw. Lerngruppe aggregierte Mittelwerte der Testleistungen in die Modelle einbezogen.

Mplus gibt für logistische Regressionen unstandardisierte, y-standardisierte sowie yx-standardisierte Koeffizienten aus. Für die Interpretation der Ergebnisse wurde auf die yx-standardisierten Koeffizienten zurückgegriffen. Diese haben gegenüber anderen Kennwerten (z.B. *odds ratios* oder teilstandardisierten Koeffizienten) den Vorteil, dass sie wie standardisierte Koeffizienten gewöhnlicher linearer Regressionen interpretiert werden können

³ Eine genaue Dokumentation aller Variablen sowie der Anteile fehlender Werte findet sich unter https://www.bifie.at/wp-content/uploads/2017/08/M812I_EXTERN_2_0_Almanach.pdf (Zugriff 08.10.2018)

⁴ Die Leistungsgruppe hat fehlende Werte, die bei der multiplen Imputation geschätzt wurden. Der geschätzte Wert variiert über die 10 Datensätze. MPlus kann mit variierenden Werten in Gruppierungsvariablen nicht umgehen.

(Menard, 2011). Die YX-Standardisierung impliziert die Zentrierung aller Variablen am *grand mean*.⁵ Gignac und Szodorai (2016) sprechen sich dafür aus, standardisierte Koeffizienten $< .10$ im Sinne schwacher Zusammenhänge, Koeffizienten zwischen $.10$ und $.30$ als mittlere und Koeffizienten $> .30$ als starke Zusammenhänge zu interpretieren. Da es sich bei Noten und Testleistungen nicht um zwei vollständig verschiedene Konstrukte handelt – in beiden Fällen wird Leistung gemessen – wird für die Beurteilung des Zusammenhangs dieser beiden Größen die Einstufung nach Cohen beibehalten ($r = .10$ niedrig, $r = .30$ mittel, $r = .50$ hoch).

4. Ergebnisse

4.1 Verteilungen der Noten

Im ersten Schritt werden deskriptiv die Verteilungen der Halbjahresnoten nach Schulzweig und Leistungsgruppen dargestellt. Auf eine Darstellung der Verteilungen der Testleistungen wird verzichtet, diese können dem Bundesergebnisbericht für die Bildungsstandarderhebung (Schreiner & Breit, o. J., S. 23) sowie dem Nationalen Bildungsbericht (Bruneforth et al., 2016, S. 87) entnommen werden.

Etwa hier Abbildung 1 einfügen

Abbildung 1 zeigt die Notenverteilung in der Allgemeinbildenden höheren Schule (AHS) sowie in den drei Leistungsgruppen der Hauptschule (HS). Alle vier Notenverteilungen ähneln wie erwartet einer Normalverteilung, d.h. die Note 3 bildet die am häufigsten vergebene Note, gefolgt von den Noten 2 und 4. Darüber hinaus unterscheiden sich die Verteilungen in charakteristischer Weise voneinander: Die Note 1 wurde in der 2. und 3. Leistungsgruppe HS seltener vergeben als in den anderen beiden Gruppen, auf der anderen Seite wurde die Note 5 in der 1. und 2. Leistungsgruppe HS seltener vergeben als in den anderen beiden Gruppen. Bemerkenswert sind außerdem die Unterschiede in der Notenverteilung von AHS und 1. Leistungsgruppe HS. In Letzterer wurden die Noten 1, 2 und 3 häufiger, die Noten 4 und 5 dagegen seltener vergeben als in Ersterer.

4.2 Einfache logistische Regressionen

Etwa hier Tabelle 1 einfügen

Tabelle 1 gibt einen Überblick über die im zweiten Schritt mittels logistischer Regressionen berechneten einfachen Zusammenhänge zwischen Noten und Testleistungen. Wie erwartet lag der Zusammenhang in der Gesamtpopulation der Sekundarstufe I mit $\log \beta = .34$ ($p < .001$) deutlich niedriger als in der Primarstufe ($\log \beta = .66$; vgl. Schmid et al., 2016). Ebenfalls wie erwartet erhöhte sich der Zusammenhang bei statistischer Kontrolle der Leistungsgruppen auf $\log \beta = .49$ ($p < .001$). Nach Umrechnung der Noten (+2 Noten in der 2. Leistungsgruppe, +4 Noten in der 3. Leistungsgruppe) erhöhte sich der Zusammenhang sogar auf $\log \beta = .66$ ($p < .001$).

⁵ Eine Zentrierung am *grand mean* gewährleistet, dass der L2-Prädiktor (mittlere Testleistung der Klasse bzw. Lerngruppe) für die Effekte auf L1 (insb. individuelle Testleistungen) kontrolliert ist. Somit wird die Signifikanz des partiellen Regressionskoeffizienten geprüft, der die Frage beantwortet, inwieweit die Zusammensetzung der Klasse bzw. Lerngruppe über die individuellen Testleistungen hinaus einen Effekt hat (Enders & Tofighi, 2007; Lüdtke, Robitzsch, Trautwein & Kunter, 2009).

Bei getrennter Betrachtung zeigte sich erwartungsgemäß im AHS-Zweig ein höherer Zusammenhang zwischen Noten und Testleistungen ($\log \beta = .51, p < .001$) als im APS-Zweig ($\log \beta = .36, p < .001$). Überraschend niedrig fielen die Zusammenhänge in den drei Leistungsgruppen der HS aus (1. Leistungsgruppe: $\log \beta = .29, p < .001$; 2. Leistungsgruppe $\log \beta = .21, p < .001$; 3. Leistungsgruppe $\log \beta = .26, p < .001$), diese sind durchweg niedriger als der Zusammenhang im APS-Zweig insgesamt ($\log \beta = .36, p < .001$). Etwas höher fiel der Zusammenhang in der vierten Gruppe aus, die nicht leistungsdifferenziert unterrichtet wurde ($\log \beta = .45, p < .001$).

4.3 Systematische Abweichungen in Abhängigkeit von Geschlecht, sozialem Status und Migrationshintergrund

Im dritten Schritt wurde überprüft, inwieweit sich systematische Abweichungen der Noten von den Testleistungen ergeben. Hierfür wurden logistische Regressionen berechnet, in denen die Noten die abhängige Größe bildeten und Geschlecht, sozialer Status des Elternhauses und Migrationshintergrund neben den Testleistungen als Prädiktoren berücksichtigt wurden. Die Analysen wurden sowohl für die Gesamtpopulation als auch getrennt für die Leistungsgruppen durchgeführt. Die Ergebnisse sind in Tabelle 2 dokumentiert.

Etwas hier Tabelle 2 einfügen

Zunächst kann festgehalten werden, dass sich der Zusammenhang zwischen Testleistungen und Noten in der Gesamtpopulation aller Schüler/innen der Sekundarstufe I (Modell 1) bei Kontrolle von Geschlecht, sozialem Status der Eltern und Migrationshintergrund mit $\log \beta = .35 (p < .001)$ gegenüber dem im vorigen Schritt berechneten Zusammenhang nur geringfügig veränderte. Darüber hinaus ergab sich eine signifikante negative Abweichung in Abhängigkeit vom Geschlecht der Schüler/innen ($\log \beta = -.06, p < .001$) – wie erwartet erhielten männliche Schüler bei gleichen Testleistungen etwas schlechtere Noten als weibliche Schülerinnen. Erwartungswidrig ergab sich keine positive, sondern eine schwache negative Abweichung in Abhängigkeit vom sozialen Status der Eltern ($\log \beta = -.02, p < .001$), was bedeutet, dass Schüler/innen aus Elternhäusern mit höherem sozialen Status bei gleichen Testleistungen etwas schlechtere Noten erhielten als Schüler/innen aus Elternhäusern mit niedrigerem sozialen Status. Für den Migrationshintergrund ergab sich mit $\log \beta = .02 (p < .001)$ eine schwache positive Abweichung, d.h. Schüler/innen mit Migrationshintergrund erhielten bei gleicher Testleistung etwas bessere Noten im Vergleich zu ihren Mitschüler/innen ohne Migrationshintergrund.

Im Modell 2 wurden zusätzlich zu den drei Merkmalen Geschlecht, sozialer Status der Eltern und Migrationshintergrund der Schulzweig bzw. die Leistungsgruppe in Form von Dummyvariablen in das Modell mit aufgenommen – als Referenzkategorie diente die 1. Leistungsgruppe HS. Der Zusammenhang zwischen Testleistung und Note erhöhte sich erwartungsgemäß durch die Kontrolle der Leistungsgruppen ($\log \beta = .51, p < .001$). Die Abweichung in Abhängigkeit vom Geschlecht fiel wiederum negativ aus ($\log \beta = -.08, p < .001$). Die Abweichung in Abhängigkeit vom sozialen Status des Elternhauses weist jetzt in die erwartete Richtung ($\log \beta = .06, p < .001$), was bedeutet, dass Kinder aus Elternhäusern mit höherem sozialen Status bei gleichen Testleistungen signifikant bessere Noten erhielten. Die Abweichung in Abhängigkeit vom Migrationshintergrund weist nach wie vor ein

positives Vorzeichen auf ($\log \beta = .05, p < .001$). Bemerkenswert ist das mittel hohe negative Betagewicht der AHS-Schülerinnen ($\log \beta = -.27, p < .001$), welches anzeigt, dass Schüler/innen der AHS bei gleichen Testleistungen (sowie unter Kontrolle von Geschlecht, sozialem Status und Migrationshintergrund) deutlich schlechtere Noten erhielten als die Schülerinnen der Referenzgruppe, d.h. der 1. Leistungsgruppe HS. Schüler/innen der 2. und 3. Leistungsgruppe HS dagegen erhielten im Schnitt etwas bessere Noten im Vergleich zur Referenzgruppe.

Die Modelle 3 und 4 weisen die Ergebnisse getrennter Analysen in den beiden Schulzweigen aus. Sie zeigen, dass auch bei Kontrolle von Geschlecht, sozialem Status des Elternhauses sowie dem Migrationshintergrund der Zusammenhang von Testleistungen und Noten im AHS-Zweig höher ausfiel als im APS-Zweig ($\log \beta = .52, p < .001$ vs. $\log \beta = .37, p < .001$). Darüber hinaus ergaben sich wiederum in beiden Schulzweigen negative Abweichungen der Noten von den Testleistungen für männliche Schüler (AHS: $\log \beta = -.11, p < .001$; APS: $\log \beta = -.07, p < .001$), sowie positive Abweichungen für Schüler/innen aus Elternhäusern mit höherem sozialen Status (AHS: $\log \beta = .07, p < .001$; APS: $\log \beta = .04, p < .001$). Für Schüler/innen mit Migrationshintergrund ergaben sich ebenfalls in beiden Schulzweigen positive Abweichungen (AHS: $\log \beta = .03, p < .001$; APS: $\log \beta = .05, p < .001$).

4.4 Mehrebenenanalysen

Im vierten Schritt wurden Mehrebenenanalysen jeweils getrennt nach Schulzweig und Leistungsgruppen durchgeführt (vgl. Tabelle 3). Im AHS-Zweig dienten die Klassen, im APS-Zweig die klassenübergreifend gebildeten Lerngruppen als Clustervariablen zur Definition der zweiten Ebene.⁶

Etwa hier Tabelle 3 einfügen

Die Ergebnisse für die *within* Ebene zeigen zunächst, dass der Zusammenhang zwischen Testleistungen und Noten im AHS-Zweig innerhalb von Klassen mit $\log \beta = .66 (p < .001)$ höher ausfiel als im APS-Zweig innerhalb der jeweiligen Lerngruppen. Allerdings war er hier wiederum in der 1. Leistungsgruppe mit $\log \beta = .56 (p < .001)$ höher als in den anderen beiden Leistungsgruppen ($\log \beta = .33, p < .001$ bzw. $\log \beta = .30, p < .001$). Wiederum zeigte sich in allen vier Leistungsgruppen eine negative Abweichung der Noten von den Testleistungen in Abhängigkeit vom Geschlecht. Die Abweichung in Abhängigkeit vom sozialen Status des Elternhauses war ebenfalls in allen vier Leistungsgruppen signifikant, aber nur schwach ausgeprägt. Dasselbe gilt für die Abweichungen in Abhängigkeit vom Migrationshintergrund, mit Ausnahme der 3. Leistungsgruppe, für die sich kein signifikanter Effekt ergab.

Auf der *between* Ebene zeigte der Klassen- bzw. Lerngruppenmittelwert in allen Untergruppen, mit Ausnahme der 3. Leistungsgruppe HS, einen signifikant negativen Effekt auf die Mathematiknote. Das bedeutet, dass es in Klassen bzw. Lerngruppen mit höheren durchschnittlichen Testleistungen schwieriger war, eine gute Note zu erzielen. Relativ hoch liegt der Anteil an Varianz, der durch den Klassen- bzw. Leistungsgruppenmittelwert auf der

⁶ Die Zahl der Cluster variiert je nach verwendetem imputierten Datensatz. Für den in Tabelle 3 verwendeten Datensatz lag sie in der AHS bei 1190 Klassen, in der 1. LG HS bei 2600, in der 2. LG HS bei 2562 und in der 3. LG HS bei 1951 Lerngruppen.

between Ebene erklärt wird, in der AHS (50 %) sowie in der 1. Leistungsgruppe HS (70 %). In der 2. Leistungsgruppe HS beträgt er dagegen nur niedrige 17 %.

5. Diskussion

Der vorliegende Beitrag verfolgte zum einen das Ziel, schulformübergreifend den Zusammenhang zwischen Noten und Testleistungen auf der Grundlage der 2012 zum ersten Mal bundesweit durchgeführten Bildungsstandardüberprüfungen für das Fach Mathematik in der österreichischen Sekundarstufe I zu untersuchen. Zum anderen wurden Abweichungen zwischen Noten und Testleistungen in Abhängigkeit von zwei verschiedenen Merkmalsgruppen analysiert. Die erste Merkmalsgruppe umfasste institutionelle Bedingungen des Lernens, d.h. die Gruppierung von Lernenden in Form von Schulzweigen, Leistungsgruppen und Klassen bzw. Lerngruppen. Zur zweiten Merkmalsgruppe gehörten das Geschlecht, der soziale Status des Elternhauses und der Migrationshintergrund der Schüler/innen.

5.1 Abweichungen in Abhängigkeit von der Organisation des Lernens

Zunächst einmal kann festgehalten werden, dass die Noten in der Sekundarstufe I, wie erwartet, eher normal verteilt waren. Hier zeichnet sich ein Unterschied zur Notenverteilung in der Primarstufe ab, wo die Verteilung der Mathematiknoten deutlich schief ausfiel und durch die Dominanz der Note 1 gekennzeichnet war (Schmid et al., 2016). Darüber hinaus ergaben sich charakteristische Abweichungen von der Normalverteilung innerhalb der einzelnen Leistungsgruppen. Diese lassen sich jeweils durch die Umstufungsregelungen erklären und können somit als durch die Vorgaben des Bildungssystems intendiert betrachtet werden.

Als problematisch einzustufen sind jedoch die Abweichungen, die sich im Vergleich der AHS mit der 1. Leistungsgruppe HS abzeichneten. In der 1. Leistungsgruppe HS wurden die Noten 1, 2 und 3 häufiger, die Noten 4 und 5 dagegen seltener vergeben. Die dadurch im Schnitt besseren Noten der 1. Leistungsgruppe HS entsprechen keineswegs besseren Testleistungen, was in den multiplen Regressionen deutlich zum Ausdruck kam: Für die Schüler/innen der AHS ergab sich mit $\log \beta = -.27$ ($p < .001$, vgl. Tabelle 2) ein mittel hohes negatives standardisiertes Gewicht, welches anzeigt, dass die Schüler/innen der AHS bei gleichen Testleistungen (sowie unter Kontrolle von sozialen Hintergrundmerkmalen) deutlich schlechtere Noten erhielten als die Schüler/innen der 1. Leistungsgruppe HS. Dieses Ergebnis weist auf eine erhebliche Gerechtigkeitslücke hin, denn für beide Schüler/innengruppen bestehen dieselben Berechtigungen hinsichtlich des Übergangs in weiterführende Schulen in der Sekundarstufe II (vgl. hierzu auch die Diskussion im NBB, Bruneforth et al. 2016). Über mögliche Gründe für die wohlwollendere Benotung in der 1. Leistungsgruppe HS im Vergleich zur AHS kann nur spekuliert werden. Unterschiedliche curriculare Schwerpunkte bei gleicher Ausschöpfung des Notenspektrums sowie unterschiedliche pädagogische Kulturen in den beiden Schulzweigen könnten mögliche Erklärungen bieten.

Die Betrachtung der einfachen Zusammenhänge (vgl. Tabelle 1) zeigte, dass der Zusammenhang zwischen Noten und Testleistungen leistungsgruppenübergreifend nur mittel hoch ausfiel ($\log \beta = .34$, $p < .001$), sich bei statistischer Kontrolle der Leistungsgruppen ($\log \beta = .49$, $p < .001$) sowie unter Anwendung der gängigen Umrechnungspraxis von Noten (\log

$\beta = .66, p < .001$) aber erwartungsgemäß erhöhte. Letzteres entspricht zum einen einer recht guten Wiedergabe der Rangfolge der Leistungen durch Noten im Zusammenhang mit der Leistungsgruppe, und zum anderen in etwa der Höhe des Zusammenhangs in der Primarstufe.

Für den AHS-Zweig ergab sich immerhin ein eher hoher Regressionskoeffizient von $\log \beta = .51$ ($p < .001$), so dass hier über Schulen und Klassen hinweg von einer recht guten Übereinstimmung zwischen Noten und Testleistungen gesprochen werden kann. Im APS-Zweig bzw. in den Leistungsgruppen der Hauptschule fielen die Zusammenhänge dagegen substantiell niedriger aus ($\log \beta < .30$). Offenbar fällt es Lehrkräften an den Hauptschulen schwer, in den jeweiligen Leistungsgruppen über Schulen hinweg bei der Benotung vergleichbare Leistungsstandards zugrunde zu legen.

Die mehrbenenanalytische Betrachtung des Zusammenhangs zwischen Testleistungen und Noten (vgl. Tabelle 3) offenbarte darüber hinaus, dass Lehrkräfte beider Schulzweige dazu tendieren, den Notenspiegel an das Leistungsniveau ihrer jeweiligen Klasse bzw. Lerngruppe anzupassen. In der AHS wie in der 1. und 2. Leistungsgruppe HS ergab sich ein negativer Effekt des Mittelwerts der Testleistungen der Klasse bzw. Lerngruppe auf die Noten. Dass dieser negative Effekt in der 1. Leistungsgruppe HS stärker ausfiel als in der AHS, zeigt, dass die soziale Bezugsnormorientierung in der 1. Leistungsgruppe HS eine größere Rolle spielte, was noch einmal die bereits getroffene Feststellung unterstreicht, dass es Lehrkräften an Hauptschulen offenbar schwerer fällt, die Noten an feste Leistungskriterien zu binden als Lehrkräften in der AHS. Ein Grund hierfür könnte sein, dass die Leistungszusammensetzung von Lerngruppen in den jeweiligen Leistungsgruppen der HS über Schulen hinweg stärker variiert als die entsprechende Zusammensetzung der Klassen in der AHS.

Die Mehrebenenanalysen zeigen darüber hinaus, dass es Lehrkräften in der AHS auch innerhalb von Klassen bzw. Lerngruppen besser gelang, die Rangfolge der Leistungen mit Noten abzubilden als denjenigen in den Hauptschulen. Der Zusammenhang zwischen Testleistungen und Noten *innerhalb* von Klassen bzw. Lerngruppen lag in der AHS bei $\log \beta = .66$ ($p < .001$), in der 1. Leistungsgruppe HS immerhin bei $\log \beta = .56$ ($p < .001$). Nur mittel hoch fiel er in der 2. und 3. Leistungsgruppe HS aus ($\log \beta = .33, p < .001$ bzw. $\log \beta = .30; p < .001$). Eine mögliche Erklärung hierfür wäre, dass der Anteil pädagogisch motivierter Noten – im disziplinierenden wie im motivierenden Sinne – in den niedrigeren Leistungsgruppen zunimmt. Außerdem ist nicht auszuschließen, dass eine niedrigere Varianz der Testleistungen innerhalb der leistungsdifferenzierten Lerngruppen der HS den Zusammenhängen entgegen gewirkt haben könnte.

5.2 Abweichungen in Abhängigkeit von Geschlecht, sozialem Status und Migrationshintergrund

Die gewöhnlichen Regressionsanalysen (vgl. Tabelle 2) dokumentieren, dass männliche Schüler wie erwartet sowohl insgesamt als auch bei getrennter Betrachtung beider Schulzweige bei gleichen Testleistungen schlechtere Noten erzielten als weibliche Schülerinnen. Die präsentierten Mehrebenenanalysen (vgl. Tabelle 3) zeigen darüber hinaus, dass dieses Ergebnis kein Artefakt einer unterschiedlichen Zusammensetzung der Geschlechter in den jeweiligen Schulzweigen bzw. Leistungsgruppen darstellt, sondern sich

negative Abweichungen der Noten von den Testleistungen auch auf Ebene der einzelnen Klassen bzw. Lerngruppen ergaben.

Der Befund schlechterer Noten von männlichen Schülern bei gleichen Testleistungen stimmt mit den berichteten Ergebnissen anderer Studien überein. Er könnte hier wie dort auf ein schulunzuträgliches Verhalten von männlichen Schülern (z.B. weniger Anstrengungsbereitschaft und Gewissenhaftigkeit; Maaz et al., 2013) zurückzuführen sein. Möglicherweise spielt dabei das pubertäre Alter der getesteten Achtklässler eine zusätzliche Rolle, denn in einer Mehrebenenanalyse auf Grundlage der österreichischen BIST-Ü M4 wurde kein entsprechender Effekt gefunden (Schmid et al., 2016).

Im Allgemeinen erzielten Kinder aus Elternhäusern mit höherem Sozialstatus bei gleichen Testleistungen bessere Noten. Vor diesem Hintergrund überraschte das negative Vorzeichen für den Sozialstatus des Elternhauses in der multiplen Regressionsanalyse für die Gesamtpopulation (Tabelle 2, Modell 1). Der Grund hierfür dürfte in der unterschiedlichen Zusammensetzung der Schülerschaft in den beiden Schulzweigen liegen: Schüler/innen aus Elternhäusern mit einem höheren Sozialstatus finden sich zu höheren Anteilen in der AHS wieder, wo es schwieriger ist, bei gleichen Testleistungen gute Noten zu erzielen. Erst bei getrennter Betrachtung der Schulzweige (vgl. Tabelle 2, Modelle 3 und 4) zeigte sich das erwartete Muster, demnach Kinder aus Elternhäusern mit einem höheren Sozialstatus bei gleicher Testleistung bessere Noten erhielten. Auch in den Mehrebenenanalysen fielen die Effekte entsprechend aus (vgl. Tabelle 3). In allen Leistungsgruppen benoteten Lehrkräfte Kinder aus Elternhäusern mit einem höheren sozialen Status besser als Kinder aus Elternhäusern mit einem niedrigeren sozialen Status. Ein möglicher Grund hierfür könnte in einem schulunzuträglicheren Verhalten (z.B. höhere Motivation, mehr Gewissenhaftigkeit, mehr mündliche Mitarbeit) von Kindern aus besser gestellten Elternhäusern liegen (Maaz et al., 2013; vgl. auch Ditton, 2006).

Hinsichtlich der Abweichungen der Noten von den Testleistungen in Abhängigkeit vom Migrationshintergrund war die berichtete Forschungslage uneindeutig. Die vorliegenden Ergebnisse zeigen jedoch, dass sich in allen Analysen, d.h. auch auf der Klassen- bzw. Lerngruppenebene – mit Ausnahme der 3. Leistungsgruppe HS – für den Migrationshintergrund positive Abweichungen ergaben. Dieses Ergebnis, demnach Kinder mit Migrationshintergrund bei gleichen Testleistungen bessere Noten erzielten als Kinder ohne Migrationshintergrund, steht im Gegensatz zu entsprechenden Befunden für die österreichische Primarstufe, wo sich auf der Ebene von Klassen negative Effekte zeigten (Schmid et al., 2016). Allerdings decken sich die Ergebnisse insgesamt mit den Befunden bei Hochweber (2010), der für die Primarstufe ebenfalls negative und für die Sekundarstufe I positive Abweichungen fand. Hochweber erklärte die negativen Abweichungen in der Primarstufe damit, dass auf dieser Stufe den sprachlichen Fähigkeiten der Kinder insgesamt ein höheres Gewicht beigemessen wird, was sich im Sinne eines Halo-Effektes auch auf die Mathematiknoten ausgewirkt haben könnte. Für die positiven Abweichungen in der Sekundarstufe dagegen könnte eine höhere Motivation von Schüler/innen mit Migrationshintergrund im Sinne eines „immigrant optimism“ ausschlaggebend gewesen sein (Walter, 2014).

5.3 Limitationen und Ausblick

Zugegebenermaßen bleiben die angebotenen Erklärungen recht spekulativ. Für eine Überprüfung bedürfte es entsprechender Operationalisierungen im Rahmen weiterer Studien. Offen blieb insbesondere, inwieweit die gefundenen Abweichungen der Noten in Abhängigkeit von Geschlecht, sozialem Status oder Migrationshintergrund auf leistungsferne oder leistungsgedeckte Merkmale zurückzuführen sind. Zu den möglichen leistungsfernen Merkmalen gehören Vorurteile und Stereotype von Lehrkräften, aber auch bestimmte Verhaltensweisen von Schüler/innen wie etwa Disziplinschwierigkeiten oder sprachliche Probleme, die nicht direkt Bestandteil von Fachleistung sind, aber dennoch in Noten einfließen können. Zu den leistungsgedeckten Merkmalen, die legitimerweise in die Noten einfließen, nicht aber in Testleistungen Berücksichtigung finden, gehört vor allem die mündlich erbrachte Leistung.

Es soll noch einmal erwähnt werden, dass die Analysen zum Teil auf nur einem, statt auf allen 10 imputierten Datensätzen basieren. Höhe und Signifikanz von Koeffizienten können dadurch sowohl leicht unter- als auch überschätzt sein. Die im Wesentlichen diskutierten Ergebnisse sollten davon jedoch unbenommen sein, da sich bestimmte Asymmetrien über die Leistungsgruppen hinweg zwar als zum Teil nur sehr schwach ausgeprägt ($\log \beta < .05$),⁷ aber in der Richtung als stabil erwiesen.

Insgesamt kann der vorliegende Beitrag als eine Art Baseline-Studie zu Asymmetrien zwischen Noten und Testleistungen vor der flächendeckenden Umwandlung der HS in NMS betrachtet werden. Mit der Umwandlung gingen die Abschaffung der Leistungsgruppen zugunsten eines binnendifferenzierten Unterrichts innerhalb von Klassen sowie eine Änderung des Benotungssystems einher. Inwieweit die neuen Regelungen zu einer höheren oder niedrigeren Übereinstimmung zwischen Noten und Testleistungen beitragen ist eine Frage, die dann auf Grundlage der BISTÜ-M8 2017 untersucht werden kann.

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 und Budrich.

Bos, W., Voss, A., Lankes, E.-M., Schwippert, K., & Thiel, O. (2004). Schullaufbahneempfehlungen von Lehrkräften für Kinder am Ende der vierten Jahrgangsstufe. In W. 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.

Bruneforth, M., Lassnigg, L., Vogtenhuber, S., Schreiner, C., & Breit, S. (Hrsg.). (2016). Nationaler Bildungsbericht Österreich 2015, Band 1. Das Schulsystem im Spiegel von Daten und Indikatoren. Graz: Leykam.

⁷ Die praktische Relevanz von Effekten mit einem $\log \beta < .05$ dürfte nicht sehr hoch sein.

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

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

7
8
9 Eder, F., & Dämon, K. (2010). Leistungsvergleiche zwischen Hauptschule und AHS-
10 Unterstufe. In F. Eder & G. Hörl (Hrsg.), *Schule auf dem Prüfstand* (S. 13–56). Wien: Lit
11 Verlag.

12
13
14 Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multi-
15 level models: a new look to an old issue. *Psychological Methods*, 12(2), 121–138.

16
17 Fend, H. (2006). *Neue Theorie der Schule. Einführung in das Verstehen von*
18 *Bildungssystemen*. Wiesbaden: VS-Verlag für Sozialwissenschaften.

19
20
21 George, A. C., Itzlinger-Brune forth, U., & Trendtel, M. (2015). *Stichprobenziehung in*
22 *Bildungsstandardüberprüfungen und Pilotierungen. Technische Dokumentation – BIST-Ü*.
23 Salzburg: BIFIE.

24
25
26 Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences
27 researchers. *Personality and Individual Differences*, 102, 74–78.

28
29
30 Hochweber, J. (2010). *Was erfassen Mathematiknoten?* Münster, New York, München,
31 Berlin: Waxmann.

32
33 Köller, O. (2005). Bezugsnormorientierung von Lehrkräften: Konzeptuelle Grundlagen,
34 empirische Befunde und Ratschläge für praktisches Handeln. In R. Vollmeyer & J. C.
35 Brunstein (Hrsg.), *Motivationspsychologie und ihre Anwendungen* (S. 189–202). Stuttgart:
36 Kohlhammer.

37
38
39 Lüdtke, O., Robitzsch, A., Trautwein, U., & Kunter, M. (2009). Assessing the impact of
40 learning environments: How to use student ratings of classroom or school characteristics in
41 multilevel modeling. *Contemporary Educational Psychology*, 34, 120–131.

42
43
44 Robitzsch, A., Pham, G., & Yanagida, T. (2016). Fehlende Daten und Plausible Values. In S.
45 Breit & C. Schreiner (Hrsg.), *Large-Scale Assessment mit R: Methodische Grundlagen der*
46 *österreichischen Bildungsstandard-Überprüfung* (259–249). Wien: facultas.

47
48
49 Maaz, K., Baeriswyl, F., & Trautwein, U. (2013). Herkunft zensiert? Leistungsdiagnostik und
50 soziale Ungleichheiten in der Schule. In D. Deißner (Hrsg.), *Chancen bilden. Wege zu einer*
51 *gerechteren Bildung - ein internationaler Erfahrungsaustausch* (S. 184–341). Wiesbaden:
52 Springer.

53
54
55 Menard, S. (2011). Standards for standardized logistic regression coefficients. *Social Forces*,
56 89(4), 1409–1428.

- Schmid, C., Paasch, D., & Katstaller, M. (2016). Kompositionseffekte bei der Notenvergabe in Mathematik auf der 4. Schulstufe der österreichischen Volksschule. *Zeitschrift für Bildungsforschung*, 6, 265–283. <https://doi.org/10.1007/s35834-016-0170-3>
- Schöps, K., Walter, O., Zimmer, K., & Prenzel, M. (2006). Disparitäten zwischen Jungen und Mädchen in der mathematischen Kompetenz. In PISA-Konsortium Deutschland (Hrsg.), *PISA 2003. Untersuchungen zur Kompetenzentwicklung im Verlauf eines Schuljahres* (S. 209–224). Münster: Waxmann.
- Schreiner, C., & Breit, S. (Hrsg.). (o. J.). *Standardüberprüfung 2012 Mathematik, 8. Schulstufe: Bundesergebnisbericht*. Abgerufen von https://www.bifie.at/system/files/dl/01_BiSt-UE_M8_2012_Bundesergebnisbericht.pdf
- Tent, L., & Birkel, P. (2010). Zensuren. In D. H. Rost (Hrsg.), *Handwörterbuch Pädagogische Psychologie* (S. 949–958). Weinheim, Basel: Beltz.
- 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, NJ: IEA-ETS Research Institute.
- Walter, O. (2014). Immigrant optimism in Deutschland? Die Überprüfung einer Hypothese anhand der mathematikbezogenen Motivation und realistischen Bildungsaspiration von Schülerinnen und Schülern mit Migrationshintergrund. *Empirische Pädagogik*, 28(1), 36–53.
- 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. Brennan (Hrsg.), *Educational measurement* (S. 111–154). Westport: Praeger Publisher.

Abbildung 1: Notenverteilung in der Allgemeinbildenden höheren Schule (AHS) und in den drei Leistungsgruppen (LG) der Hauptschule

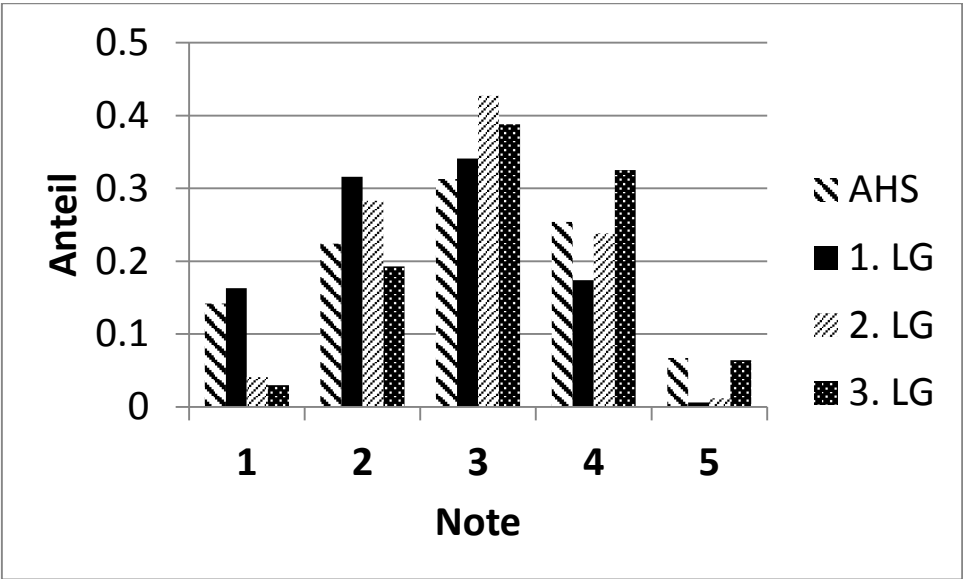


Tabelle 1: Einfache logistische Regressionen der Noten auf die Testleistungen in der Gesamtpopulation sowie in verschiedenen Untergruppen (BIST-Ü M8 2012, N = 79 588)

	Est. ¹	% von N ²
Gesamtpopulation	.34***	100,0
Statistische Kontrolle Leistungsgruppen (LG) ³	.49***	100,0
Notenumrechnung (2. LG: +2 Noten, 3. LG: +4 Noten) ⁴	.66***	100,0
AHS (Allgemeinbildende höhere Schule)	.51***	33,1
APS (Allgemeinbildender Pflichtschulzweig)	.36***	67,0
1. Leistungsgruppe HS	.29***	29,7
2. Leistungsgruppe HS	.21***	22,7
3. Leistungsgruppe HS	.26***	11,6
4. Gruppe NMS	.45***	3,1

¹ YX-standardisierte Gewichte logistischer Regressionen (Mplus), *** $p < .001$

² Die Prozentanteile von AHS und 1. bis 3. Leistungsgruppe HS sowie der 4. Gruppe NMS summieren sich aufgrund von Rundungsfehlern auf über 100%.

³ Multiple logistische Regression mit Dummyvariablen für AHS, 2. und 3. Leistungsgruppe HS sowie 4. Gruppe NMS, Referenzkategorie: 1. Leistungsgruppe HS

⁴ Die Umrechnung der Noten erfolgte vor der Umpolung für die Analysen.

Tabelle 2: Multiple logistische Regressionen der Noten auf die Testleistungen¹ in der Gesamtpopulation sowie in verschiedenen Untergruppen (BIST-Ü M8 2012, N = 79 588)

	Modell 1 Gesamt	Modell 2 Gesamt	Modell 3 AHS	Modell 4 APS
Testleistung	.35***	.51***	.52***	.37***
Geschlecht (männlich)	-.06***	-.08***	-.11***	-.07***
SES Eltern (HISEI)	-.02***	.06***	.07***	.04***
Migrationshintergrund	.02***	.05***	.03***	.05***
Leistungsgruppen ²				
AHS		-.27***		
2. Leistungsgruppe HS		.05***		
3. Leistungsgruppe HS		.05***		
4. Gruppe NMS		.01		
R ²	.12***	.18***	.28***	.14***

¹ YX-standardisierte Gewichte logistischer Regressionen (Mplus), *** $p < .001$

² Referenzkategorie 1. Leistungsgruppe HS

Tabelle 3: Logistische Mehrebenenanalysen¹ in verschiedenen Untergruppen, abhängige Größe Noten (BIST-Ü M8 2012, N = 79 588)

	AHS	1. LG	2. LG	3. LG
<i>within (Klasse/Lerngruppe)</i>				
Testleistung	.66***	.56***	.33***	.30***
Geschlecht (männlich)	-.11***	-.10***	-.08***	-.05***
SES (HISEI)	.06***	.03***	.03***	.05***
Migrationshintergrund	.02**	.04***	.04***	-.01
<i>between (Klasse/Lerngruppe)</i>				
Testleistung (Klassen-/Lerngruppenmittelwert)	-.71***	-.84***	-.41***	.01
R ² <i>within</i>	.44***	.32***	.11***	.09***
R ² <i>between</i>	.50***	.70***	.17***	.00

¹ YX-standardisierte Gewichte logistischer Regressionen (Mplus), *** $p < .001$, ** $p < .01$