Migrationsbedingte Disparitäten in der Notenvergabe nach dem Übergang auf das Gymnasium

[Students‘ grading according to migration background]

Meike Bonefeld[1]

Oliver Dickhäuser[1]

Stefan Janke[1]

Anna-Katharina Praetorius[2]

Markus Dresel[3]

[1]Universität Mannheim [2] Deutsches Institut für InternationalePädagogische Forschung (DIPF) [3]Universität Augsburg

Korrespondenz:

Meike Bonefeld, Universität Mannheim, Fakultät für Sozialwissenschaften, Lehrstuhl für Pädagogische Psychologie, A5,6 Gebäudeteil B 68131 Mannheim, meike.bonefeld@staff.uni-mannheim.de

Diese Artikelfassung entspricht nicht vollständig dem in der Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie veröffentlichten Artikel unter https://doi.org/10.1026/0049-8637/a000163. Dies ist nicht die Originalversion des Artikels und kann daher nicht zur Zitierung herangezogen werden. Bitte verbreiten oder zitieren Sie diesen Artikel nicht ohne Zustimmung des Autors.

Zusammenfassung

Disparitäten bei der Leistungsbewertung von Schülern und Schülerinnen mit Migrationshintergrund konnten häufig für den Übergang zwischen Grundschule und weiterführender Schule nachgewiesen werden. In der hier vorgestellten Studie wurde geprüft, ob sich solche Zusammenhänge auch nach dem Übergang auf das Gymnasium zeigen. Untersucht wurde der Effekt des Migrationshintergrundes von Schüler/-n/-innen auf die Leistungsbewertung im Fach Mathematik bei Klassenarbeiten und Zeugnissen. Daten von 1487 Gymnasiastinnen und Gymnasiasten und deren 56 Lehrkräften im Fach Mathematik zu fünf Messzeitpunkten (Beginn der fünften Klasse bis zum Ende der sechsten Klasse) zeigten, dass Schüler/-innen nicht-deutscher Herkunft auch unter Kontrolle von Leistungen in standardisierten Tests signifikant schlechtere Klassenarbeits- und Zeugnisnoten erhielten. Der Effekt blieb im Zeitverlauf stabil. Diese Ergebnisse unterstützen die Hypothese, dass sich Urteilsfehler bei der Benotung von Leistungen von Schülern und Schülerinnen mit nichtdeutscher Herkunft zeigen.

Schlagwörter: Leistungsbewertung, Migrationshintergrund, Mathematik, Noten, Bestätigungsfehler, Erwartungen

Abstract

Disparities in educational outcomes of students with migration background were often detected for the transition from primary school to secondary school track. Here we examined whether such disparities also persist after the transition to secondary school. Thereby, the present study considers the effect of migration background on students' grades in mathematics. Data of 1487 secondary school students and their 56 teachers in mathematics based on five points of measurement (from the start of fifth grade until the end of sixth grade) showed that students with migration background obtained significantly worse grades in exams and final school reports even after controlling for mathematical test performance. The effect was stable over time. These results support the hypothesis that a confirmation bias in giving grades for students with migration background might exist.

Keywords: evaluation, teacher expectation, confirmation bias, performance, grades, migration background

Leistungsbewertung in Abhängigkeit des Migrationshintergrundes der Schüler/-innen

Im Jahr 2013 wiesen in Deutschland 20.5 Prozent der Bevölkerung einen Migrationshintergrund in dem Sinne auf, dass mindestens ein Elternteil nicht in Deutschland geboren war (Statistisches Bundesamt, 2014). Nach wie vor gibt es ungleiche Verteilungen in der Bildungspartizipation und im Bildungserfolg in Abhängigkeit der Migrationsherkunft (Autorengruppe Bildungsberichterstattung, 2016).

In der bisherigen Forschung wurden in Bezug auf diese Bildungsungleichheit im deutschen Schulsystem häufig Bildungsübergänge betrachtet (Schnabel & Schwippert, 2000). Diese Übergänge werden häufig als kritischer Punkt in Bildungslebensläufen, vor allem bei benachteiligten Gruppen, angesehen. Dies ist vor allem deshalb der Fall, da sie eine Vorentscheidung für den folgenden Bildungserfolg darstellen. (Kristen & Dollmann, 2010). Nachteilige Übertrittsmuster als Folge von Disparitäten wurden daher häufig an dem ersten Bildungsübergang (von der Primarstufe zur Sekundarstufe) untersucht. Dabei zeigte sich, dass Schüler/-innen mit Migrationshintergrund deutlich seltener eine Gymnasialempfehlung erhalten als solche ohne Migrationshintergrund (Bos et al., 2004). In Folge gibt es große Unterschiede in der Verteilung von Schülern und Schülerinnen auf die verschiedenen Schulformen. Beispielsweise besuchen lediglich knapp ein Viertel der Jugendlichen mit Migrationshintergrund (24 Prozent) ein Gymnasium, während im Vergleich dazu rund 44 Prozent der Schüler/-innen ohne Migrationshintergrund ein Gymnasium besuchen. (Autorengruppe Bildungsberichterstattung, 2016). Ungleichheiten auch *nach* diesem als kritisch angesehenen ersten Bildungsübergang konnten in bisheriger Forschung - etwa im Rahmen großer Schulleistungsvergleichsstudien - nachgewiesen werden (OECD, 2011; OECD, 2014). Die vorliegenden Daten erstrecken sich typischerweise allerdings über verschiedene Schulformen hinweg. Aufgrund der Tatsache, dass die Zuweisung zu den Schulformen disparat erfolgt und die unterschiedlichen Schulformen differentielle Milieus der Leistungsentwicklung darstellen, könnten migrationsbedingte Leistungsdisparitäten, wie sie etwa

bei PISA für 15jährige nachgewiesen werden unter Umständen auch eine reine Folge der differentiellen Zuweisung sein. Hinzu kommt, dass relevante Kontrollvariablen, wie beispielsweise das objektive Leistungsniveau der Schüler/-innen oft vernachlässigt wurden. So können über mögliche Gründe der Disparitäten keine genaueren Aussagen getroffen werden. Hinsichtlich möglicher Ursachen für Disparitäten ist zu beachten, dass Disparitäten nicht nur auf nicht nur auf tatsächliche Unterschiede in der Leistungsfähigkeit, beispielsweise aufgrund von Sprachschwierigkeiten oder mit der Herkunft verbundenen gegebenenfalls ungünstigen häuslichen Förderungsbedingungen, zurückgeführt werden können. Es muss dabei differenziert werden, ob diese Effekte Disparitäten primärer Natur, das heißt in diesem Fall beispielsweise tatsächliche Unterschiede in den Leistungen darstellen oder vielmehr Disparitäten sekundärer Natur sind, das heißt unabhängig von der objektiven Leistungsfähigkeit und damit verbundene Faktoren, wie beispielsweise den Sprachfertigkeiten, bestehen.

Im Folgenden soll daher zum einen untersucht werden, ob Bildungsungleichheiten in Abhängigkeit des Migrationshintergrundes auch *nach* der Realisierung eines im deutschen Schulsystem gerade für Disparitäten bedeutsamen Bildungsüberganges fortbestehen und zum anderen ob diese Disparitäten auch unter Berücksichtigung relevanter Kontrollvariablen sowie spezifischer Betrachtung innerhalb einer Schulform Bestand haben. Betrachtet wird hierzu die Phase direkt nach dem ersten Bildungsübergang (fünfte und sechste Klasse) auf dem Gymnasium, welche durch den Schulwechsel eine sensible Phase bedeutsamer Anpassungsprozesse für Schüler/-innen und auch Lehrkräfte darstellt. Da Zensuren einen wichtigen Faktor für den Schulerfolg darstellen, aber wie schon vielfach belegt kein fehlerfreies Beurteilungsinstrument darstellen (Ingenkamp, 1995), nimmt diese Arbeit Urteilsfehler in Hinblick auf diese Form der Leistungsbeurteilung in den Blick.

Fehlereinflüsse in der Notenvergabe

Diagnostische Urteile können unter Nutzung verschiedener Bezugsnormen gefällt werden. Das Lehrerurteil bildet dabei häufig nicht nur ab, in welchem Ausmaß das Schülerresultat gesetzten Kriterien entspricht (kriteriale Bezugsnorm). Vielmehr spielt für das Lehrerurteil z.B. die soziale Bezugsnorm ebenfalls eine Rolle. So konnten etwa Trautwein und Baeriswyl (2007) zeigen, dass Lehrerurteile umso positiver ausfallen, je geringer das Leistungsniveau der Klasse ist, innerhalb derer sich der zu beurteilende Schüler/die zu beurteilende Schülerin befindet. Die Befunde illustrieren bereits, dass Lehrerurteile gewissen Fehlereinflüssen unterliegen und das tatsächliche Niveau der Schülerleistung nicht perfekt abbilden. Gründe für Urteilsfehler in der Notengebung durch Lehrkräfte können vielfältiger Natur sein. Gerade im Bereich der Beurteilung von Schülern und Schülerinnen mit Migrationshintergrund könnten Stereotype eine Rolle in diesem Prozess spielen. Analysen zu kulturellen Stereotypen in Deutschland konnten zeigen, dass Deutsche mit Migrationshintergrund von anderen Deutschen weniger stark mit leistungs- und erfolgsbezogenen Attributen beschrieben werden als Deutsche ohne Migrationshintergrund (Kahraman & Knoblich, 2000). Solche Stereotype könnten bedingen, dass Lehrkräfte geringere Erwartungen an das Leistungspotential von Schülern und Schülerinnen mit Migrationshintergrund entwickeln.

In bisheriger Forschung konnte gezeigt werden, dass dieselben Leistungen in Abhängigkeit verschiedener anderer Merkmale, die ein Schüler/eine Schülerin aufwies (wie beispielsweise der sozialen Herkunft) unterschiedlich bewertet wurden (Hadley, 1995): Urteiler, die glaubten, dass ein Kind einen höheren/niedrigeren sozioökonomischen Status aufwies, bewerteten die Leistung dieses Kindes bei gleicher sonstiger Information besser/schlechter als diejenigen Urteiler, die keine Informationen über den sozioökonomischen Hintergrund des Kindes hatten (Darley & Gross, 1983).

Entsprechende Urteilsfehler könnten aufgrund der durch die kulturellen negativen Stereotype bedingten spezifischen Lehrererwartungen über das möglicherweise geringere Leistungspotential von Schülern und Schülerinnen mit Migrationshintergrund auch bei der Beurteilung dieser Gruppe von Schülern und Schülerinnen entstehen. Einen erklärenden

Mechanismus für solche Urteilsfehler könnte der sogenannte *expectancy-confirmation-bias* darstellen, welcher beschreibt, dass Personen geneigt sind ihre eigenen Erwartungen bei Bewertungsprozessen zu bestätigen (Nickerson, 1998). Die aus dieser Bestätigungstendenz folgenden Urteilsfehler haben einen nachgewiesenen Einfluss auf den Umgang der Lehrkräfte mit ihren Schülern und Schülerinnen. Beispielsweise konnte in verschiedenen Studien gezeigt werden, dass Lehrkräfte bei Schülerinnen und Schülern, für die sie niedrigere Erwartungen haben, weniger unterstützend agieren, weniger positives Feedback geben, ihnen weniger Aufmerksamkeit schenken und ihnen weniger Gelegenheiten bieten ihre Fähigkeiten zu demonstrieren als bei Schülern/Schülerinnen, für die die Lehrkräfte hohe Leistungserwartungen haben (Brophy, 1983; Jussim,1986; Jussim & Eccles, 1992; Rosenthal, 1974). Im Sinne einer selbsterfüllenden Prophezeiung könnten Schüler/-innen durch die unterschiedliche Förderung so letztendlich auch objektiv unterschiedliche Leistungen erbringen (Brophy, 1983; Jussim, 1986; Jussim & Eccles, 1992; Rosenthal, 1974).

Neben solchen Einflüssen auf die Lehrer-Schüler-Interaktion wird vermutet, dass sich Erwartungen auch direkt in der Notengebung niederschlagen können. So könnten Schüler/-innen mit Migrationshintergrund trotz gleicher objektiver Leistungsfähigkeit schlechtere Noten erhalten. Bei der Analyse der Ursachen von migrationsbedingten Disparitäten im Rahmen der Notengebung ist daher das objektive - unabhängig vom Lehrerurteil erfasste - Leistungsniveau der Schüler/-innen eine bedeutsame Kontrollvariable.

Darüber hinaus werden Unterschiede in Abhängigkeit des Migrationshintergrundes teilweise mit Ungleichheiten im Sprachgebrauch in der Familie sowie mit der sozialen Herkunft assoziiert (Esser, 2006). Ergänzend belegen vergleichende PISA-Analysen aus den Jahren 2000, 2003 und 2006 entsprechend auch, dass der Migrationshintergrund zu Teilen mit dem Sprachgebrauch und der sozialen Herkunft konfundiert ist (Stanat & Christensen, 2006; Walter, 2009). Um den Effekt des Migrationshintergrundes von Effekten solcher möglicher konfundierender Faktoren trennen zu

können, müssen diese Faktoren in den Betrachtungen von Effekten des Migrationshintergrundes berücksichtigt werden.

Für die Entstehung derjenigen Disparitäten, die nach der Kontrolle von Sprachgebrauch, sozialer Herkunft und der objektiven Leistungsfähigkeit verbleiben, spielen möglicherweise die zuvor erwähnten Lehrererwartungen gegenüber Schülern und Schülerinnen mit Migrationshintergrund eine Rolle. Um die Relevanz solcher Effekte auf dem weiteren Bildungsweg der Schüler/-innen einschätzen zu können, ist es wichtig zu prüfen inwieweit diese Effekte über die Zeit stabil verbleiben.

Persistenz von Urteilsfehlern

Häufig wird in diesem Zusammenhang angenommen, dass sich Erwartungseffekte auf Lehrkraftseite vor allem in neuen Situationen zeigen und diese Effekte falscher Erwartungen über die Zeit geringer werden sollten (Fiske & Neuberg, 1990), da Lehrkräfte gerade zu Beginn des Kontaktes mit den Schülern und Schülerinnen noch wenige Informationen über deren Fähigkeiten haben und somit auf alternative Informationsquellen, wie Stereotype zur Einschätzung der Leistungsstärke zurückgreifen müssen (Jussim & Harber, 2005). Das würde bedeuten, dass Urteilsfehler in der Leistungsbewertung eine geringere Rolle spielen, je länger die Schüler/-innen den Lehrkräften bekannt sind (Jussim, 1990; Jussim et al. 1996).

Demgegenüber könnte auf der Basis verschiedener sozialpsychologischer Befunde angenommen werden, dass sich erwartungsabhängige Unterschiede manifestieren und die ersten (möglicherweise fehlerhaften) Erwartungen im Zeitverlauf bestehen bleiben und Einfluss auf die Urteile nehmen. Da die ersten Schlussfolgerungen, die Personen auf Grundlage von Informationen (hier: die Erwartungen über die Leistungen der Schüler/-innen aufgrund ihrer Migrationsherkunft) früh im Prozess gezogen haben, im Laufe der Zeit immer wieder in Folgeurteile integriert werden, fallen sie im Laufe der Zeit wahrscheinlicher stärker ins Gewicht als später erworbene Informationen (hier individuierende Informationen über die Schüler/-innen; Lingle & Ostrom,

1981; Sherman et al., 1983;Smith, Jussim & Eccles, 1999). Im Sinne eines *primacy effects* wäre also denkbar, dass das früh gebildete, erste Urteil über einen Schüler/eine Schülerin die nachfolgenden Lehrkrafturteile beeinflusst, indem nachfolgend primär solche Informationen in das eigene Urteil integriert werden, die dem Anfangsurteil entsprechen (Anderson & Jacobson, 1965; Jones & Goethals, 1972; Nisbett & Ross, 1980). Eine Stabilität von Effekten falscher Erwartungen über die Zeit lässt sich auch vor dem Hintergrund des Konzepts der *belief persistence* ableiten, wonach anfängliche Schlussfolgerungen schwer revidierbar sind und darüber hinaus die Beurteilung neuer Information beeinflussen können (Hayden & Mischel, 1976; Ross & Lepper, 1980). Beispielsweise können Informationen, die für die eigene Position sprechen stärker gewichtet werden oder Informationen, welche mit der eigenen Position in Konflikt stehen in Frage gestellt werden (Ross & Anderson, 1982). Es wäre demnach zu erwarten, dass Lehrkräfte dazu neigen ihre anfänglichen Erwartungen zu bestätigen und beispielsweise Faktoren, welche ihre Erwartungen stützen, stärker in ihre Urteile einzubeziehen (Nickerson, 1998). Daher ist es nicht unplausibel zu vermuten, dass falsche Eingangserwartungen auch zu späteren Zeitpunkten stabile Effekte auf die Beurteilung durch die Lehrkraft haben.

Verschiedene Studien haben diesbezüglich untersucht, inwieweit der Effekt von selbsterfüllenden Prophezeiungen tatsächlich von der Dauer der Interaktion mit der zu beurteilenden Person abhängt (Rosenthal & Jacobson, 1968; West & Anderson 1976; Smith, Jussim & Eccles, 1999). Allen Studien ist gemein, dass sie keine Zunahme der Effekte finden konnten. Ein gemischtes Befundmuster ergab sich allerdings für Aussagen über eine Abnahme oder Stabilität der Befunde. Rosenthal und Jacobson (1968) sowie West und Anderson (1976) fanden Befundmuster, welche eher für eine Abnahme sprachen. Smith und Kollegen (1999) berichteten neben Effekten der Abnahme auch Befunde für stabile Effekte und konnten zeigen, dass die Effekte der selbsterfüllenden Prophezeiung, unabhängig von einer späteren Abnahme oder weiteren Stabilität, in allen Fällen sehr langanhaltend (über mehrere Jahre stabil) nachweisbar waren. Die Stärke des

Zusammenhangs zwischen Lehrererwartung und Noten reduzierten sich anfänglich, blieben dann aber stabil (d=0.15-0.25).

Die Befunde der bisherigen Längsschnittstudien lassen vermuten, dass Erwartungen über einen langen Zeitraum statistisch bedeutsame Effekte haben können. Sie liefern zu der Frage nach der Persistenz von Erwartungseffekten über die Zeit zusammenfassend jedoch kein klares Ergebnismuster. Darüber hinaus stammen die Studien durchweg aus anderen Schulkontexten (US-amerikanisches Schulsystem). Kontrollvariablen, wie Leistungen in standardisierten Tests zur Operationalisierung des objektiven Leistungsniveaus wurden zum Teil nicht berücksichtigt oder nicht kontinuierlich an die sich verändernden Leistungen der Schüler/-innen angepasst. Daher bleibt in dieser Studie zu untersuchen, ob die Effekte falscher Erwartungen über die Zeit unter Kontrolle von Leistungen aus standardisierten Tests und vor allem auch nach einem im deutschen Schulsystem besonders als kritisch erachteten Bildungsübergang bestehen bleiben oder abnehmen.

Zusammenfassend soll in dieser Arbeit zunächst untersucht werden, ob Notenunterschiede zwischen Schülern und Schülerinnen mit und ohne Migrationshintergrund im Fach Mathematik auf dem Gymnasium, also nach dem ersten Bildungsübergang, fortbestehen und ob diese auch nach Kontrolle der Leistungen in standardisierten Tests der Schüler/-innen und konfundierender Variablen, wie der sozialen Herkunft und dem Sprachgebrauch der Schüler/-innen bedeutsam verbleiben. Dies wäre ein Hinweis auf leistungsunabhängige Disparitäten auch nach einem für Disparitäten als zentral erachteten Bildungsübergang. Zudem soll die Stabilität der Effekte im Zeitverlauf betrachtet werden.

Hypothesen

Vermutet wird, dass es zu einer schlechteren Leistungsbewertung durch die Lehrkraft in Abhängigkeit des Migrationshintergrundes kommt. Erwartet wird, dass dieser Effekt auch unter Kontrolle der Leistungen in standardisierten Tests, des Sprachgebrauchs und der sozialen Herkunft der Schüler/-innen statistisch signifikant verbleibt, was die angenommen Erwartungseffekte

bestätigen würde und diese von Effekten von mit dem Migrationsstatus konfundierenden Variablen trennt. Da in Bezug zu der Persistenz der hier postulierten Effekte sowohl eine Abnahme als auch stabile Einflüsse mit Blick auf die dargestellten theoretischen Annahmen sowie die Empirie plausibel erklärbar sind, soll die Frage der Stabilität der Effekte über die Dauer der Interaktion zwischen Lehrkraft und Schüler/-innen in der vorliegenden Studie exploriert werden.

Methode

Stichprobe

Zur Analyse genutzt wurden Daten einer Längsschnitterhebung in deren Verlauf 1487 Gymnasiastinnen und Gymnasiasten im Laufe der ersten beiden Schuljahre der Sekundarstufe I zu fünf Messzeitpunkten (Beginn der fünften Klasse, Mitte der fünften Klasse und Ende der fünften Klasse sowie Anfang der sechsten Klasse und Ende der sechsten Klasse) befragt wurden. Außerdem lagen Daten zu der Leistungsbewertung durch deren 56 Lehrkräfte im Fach Mathematik vor. Schüler/-innen und Lehrkräfte stammten aus 56 Klassen an 33 Gymnasien in Baden-Württemberg. Die Klassen wurden über den gesamten Zeitraum von der gleichen Lehrkraft unterrichtet. Im Mittel waren die Lehrkräfte (57.1% weiblich) zum ersten Messzeitpunkt 41.1 Jahre alt (*SD*=12.5 Jahre) und 13.3 Jahre (*SD*=12.3 Jahre) im Schuldienst. Die Schüler-/innen (50.4 % weiblich) waren zum ersten Messzeitpunkt im Mittel 10.3 Jahre alt (*SD*=0.57); 24.5 Prozent der befragten Schüler/-innen wiesen einen Migrationshintergrund auf.

Instrumente

Migrationshintergrund. Als Schüler/-innen mit Migrationshintergrund wurden Schüler/-innen definiert, welche mindestens ein Elternteil haben, das in einem anderen Land als Deutschland geboren wurde oder die selbst in einem anderen Land geboren wurden (OECD, 2011; OECD, 2014). Hierzu wurden die Schüler/-innen mithilfe von drei Einzelitems mit halboffenem Antwortformat gefragt in welchem Land ihr Vater, ihre Mutter sowie sie selbst geboren wurden (z.B. „Meine Mutter ist …..“). Mögliche Antwortformate waren dabei „…in Deutschland geboren.“

(1) sowie „…nicht in Deutschland geboren, sondern….“ (2).Von einem Migrationsstatus wurde dann ausgegangen, wenn ein Schüler beziehungsweise eine Schülerin auf mindestens einem dieser Items angab, dass die Geburt nicht in Deutschland war.

Die *soziale Herkunft* wurde über den höchsten erreichten Bildungsabschluss der Eltern erfasst (Jussim et al., 1996). Hierbei wurden die Schüler/-innen gefragt auf welcher Schule ihr Vater beziehungsweise ihre Mutter war (z.B. „Auf welcher Schule war deine Mutter?“). Die Antwort konnte auf fünf Stufen gegeben werden (Keine Schule [keinen Schulabschluss] [1], Volksschule/Hauptschule [Hauptschulabschluss] [2], Realschule [Mittlere Reife] [3], Gymnasium [Abitur] [4], Sonstige Schule [z.B. im Ausland] [5]), wobei im letzteren Fall in einem Freitextfeld genauere Angaben gemacht werden konnten. Hierbei wurden für die Analysen die Angaben aus der Kategorie „Sonstige Schule“ sofern möglich den passenden anderen Kategorien zugeordnet (bspw. ein mit dem Abitur vergleichbarer Schulabschluss aus einem anderen Land wurde der Kategorie 4 zugeordnet). Berücksichtigt wurde in den Analysen jeweils der höchste Bildungsabschluss in der Familie.

Häuslicher Sprachgebrauch. Der häusliche Sprachgebrauch wurde mit der Häufigkeit des Gebrauchs der deutschen Sprache in der Familie operationalisiert. (Item: „Wie oft sprichst du zuhause Deutsch?“). Hierbei konnten die Antwort nach von „nie“(1) bis „immer“(4) auf einer vierstufigen Skala reichen.

Beurteilung durch die Lehrkraft. Zur Operationalisierung der Leistungsbewertung der Lehrkraft wurden von den Schülern und Schülerinnen selbstberichtete Noten der Klassenarbeiten und Zeugnisse im Fach Mathematik (je ein Item, z.B. „Welche Mathe-Note hattest du in der letzten Klassenarbeit?“) mit einem offenen Antwortformat verwendet. Zum ersten Messzeitpunkt wurden die Noten der vierten Klasse retrospektiv erfasst, zu den folgenden Messzeitpunkten jeweils die letzten erreichten Noten (Klassenarbeits- und Zeugnisnote Mitte der fünften Klasse, Klassenarbeits- und Zeugnisnote Ende der fünften Klasse, Zeugnisnote Mitte der sechsten Klasse,

Klassenarbeitsnote Ende der sechsten Klasse). Selbstberichtete Noten weisen eine hohe Validität auf und korrelieren hoch mit den tatsächlichen Noten, weshalb sie als ein akkurates Maß der tatsächlichen Noten gewertet werden können (Dickhäuser & Plenter, 2005).

Leistungen in standardisierte Tests. Zur Erfassung der allgemeinen Mathematiktestleistung der Schüler/-innen wurden curricular-valide allgemeine standardisierte Testverfahren herangezogen. Hierbei wurde zum ersten Messzeitpunkt (Beginn der 5. Klasse) der Hamburger Schulleistungstest für 4. und 5. Klassen (kurz: HST, Mietzel & Willenberg, 2000) eingesetzt. Für die vorliegende Untersuchung wurden alle 30 das Fach Mathematik betreffende Items genutzt (α=0.73). Die Items gliederten sich in drei Subtests („Zahlenverständnis“, „Rechenoperationen“, „Größen“). Zum dritten und fünften Messzeitpunkt (Ende der fünften Klassen beziehungsweise Ende der sechsten Klasse) wurden der Deutsche Mathematiktest für fünfte (DEMAT 5+; Götz, Lingel & Schneider, 2013a; α=0.80)[1] beziehungsweise für sechste Klassen genutzt (DEMAT 6+; Götz, Lingel & Schneider, 2013b; α=0.83). Diese Tests gliedern sich jeweils ebenfalls in drei Subtests zu den Bereichen Arithmetik, Geometrie und Sachrechnen und sind in den Lehrplänen der deutschen Bundesländer fundiert. In den folgenden Analysen wurden bei allen drei Tests jeweils der Summenscore aus den drei Subtests verwendet. Hohe Ausprägungen in den Punktzahlen aller Tests stehen für ein hohes Leistungsniveau.

Datenstruktur und statistische Verfahren

Die folgenden Analysen wurden mit Hilfe von Mplus 7.2 (Muthén & Muthén, 2010) durchgeführt. Da die Daten eine hierarchische Struktur aufwiesen (Schüler/-innen geclustert innerhalb von Klassen), es sich aber um eine Ebene-1-Fragestellung handelt und daher keine mehrebenenanalytischen Berechungen nötig sind, wurde die Analyseoption „type = complex“ verwendet, um Verzerrungen bei der Parameterschätzung zu vermeiden (Cohen et al., 2003). Zum adäquaten Umgang mit fehlenden Werten wurden alle Parameter unter Verwendung der Full information maximum likelihood (FIML)-Prozedur geschätzt (Enders, 2001).

Um den statistischen Effekt des Migrationshintergrunds auf die Benotung zu analysieren, wurde für jede der erfassten Zensuren[2] eine lineare Regressionsanalyse gerechnet. Dabei wurden folgende Regressionsmodelle geprüft: In einem ersten Schritt (Modell M1) wurde lediglich die prädiktive Bedeutsamkeit des Migrationshintergrund für die erlangte Note ohne Berücksichtigung von Kontrollvariablen geprüft. Dieses Modell dient dem Nachweis migrationsbedingter Disparitäten in den Zensuren. In einem zweiten Modell (M2) wurde die mathematische Testleistung als Prädiktor mit aufgenommen, um zu testen, ob der Effekt des Migrationshintergrundes auch dann noch verbleibt, wenn die allgemeinen Testleistungen kontrolliert werden. Dieses Modell dient somit dem Nachweis leistungsunabhängiger Disparitäten. Hierbei wurde jeweils die zeitlich aktuellste verfügbare Testleistung als Prädiktor verwendet. In einem dritten Modell (M3) wurden schließlich auch die Häufigkeit des häuslichen Gebrauchs der deutschen Sprache und die soziale Herkunft als Prädiktoren aufgenommen, um zu prüfen, ob ein etwaiger Effekt des Migrationshintergrundes nicht letzten Endes ein Effekt des Sprachgebrauches innerhalb der Familie oder der Bildungsherkunft der Schüler/-innen ist.

Zur Untersuchung der zeitlichen Entwicklung des Effektes des Migrationshintergrundes auf die Noten wurde der statistische Test zur Gleichheit von Regressionskoeffizienten nach Paternoster verwendet, welcher es ermöglicht Regressionskoeffizienten miteinander zu vergleichen (Paternoster et al., 1998). Um statistisch sparsam zu prüfen, ob die Effekte stabil verbleiben, sich verringern oder vergrößern wurden der größte Abfall sowie der größte Anstieg in den Regressionskoeffizienten getestet.

Ergebnisse

Tabelle 1 zeigt die Mittelwerte und Standardabweichungen der allgemeinen Testleistungen und der selbstberichteten Zeugnis- sowie Klassenarbeitsnoten aufgegliedert nach dem Migrationshintergrund der Schüler/-innen über alle fünf Messzeitpunkte hinweg.

Die Tabelle zeigt zunächst deskriptiv in allen betrachteten Variablen schlechtere Leistungen (niedrigere Testleistungen und schlechtere Noten) für Schüler/-innen mit Migrationshintergrund als ohne Migrationshintergrund. Es zeigt sich neben diesen Gruppenunterschieden in den Testleistungen und Zensuren, dass die Zensuren zu den späteren Messzeitpunkten für beide Gruppen schlechter ausgefallen sind als zu den früheren Messzeitpunkten Bei den in Tabelle 1 dargestellten Mittelwertsunterschieden handelt es sich um die noch nicht für Effekte der Kontrollvariablen korrigierte Unterschiede in den Noten und Testleistungen. In drei zusätzlichen linearen Regressionsanalysen (je eine Regression pro standardisiertem Leistungstest) wurde der Effekt des Migrationshintergrundes auf die Leistungen in den standardisierten Tests daher unter Kontrolle der sozialen Herkunft sowie des Sprachgebrauchs in der Familie betrachtet. Bei dieser Betrachtung verbleiben zu keinem Messzeitpunkt signifikante Unterschiede in den Leistungen in den standardisierten Tests nach dem Migrationshintergrund.

Eine Betrachtung der Effektstärken zeigt, dass die Mittelwertsunterschiede bei den Noten zwischen .15 und .34 und bei den Testleistungen zwischen .13 und .28 schwanken. Die Korrelationen zwischen den Noten und den Testleistungen liegen zwischen .3 und .7 (vgl. Elektronisches Supplement 1). Die mäßige Beziehung zwischen Testleistung und Noten weist einerseits darauf hin, dass die Testleistung eine wichtige Variable ist, die es bei der Analyse von migrationsbedingten Disparitäten in den Noten zu kontrollieren gilt. Gleichzeitig ist angesichts der nur mäßigen Interkorrelation nicht zu erwarten, dass sämtliche Mittelwertsunterschiede in den Noten verschwinden, wenn für Mittelwertsunterschiede in den Tests kontrolliert wird.

In Bezug zu der Kontrollvariablen „soziale Herkunft“ ergibt sich für die Gruppe der Schüler/-innen ohne Migrationshintergrund ein Mittelwert von 3.62 (SD=0.59). Der Modus beträgt vier was der Antwort „Gymnasium“ entspricht. Ein ähnliches Muster zeigt sich ebenfalls für die Schüler/-innen mit Migrationshintergrund. Im Mittel ergab sich ein Wert von 3.57 (SD= 0.69) und ein Modus von ebenfalls vier (Gymnasium). Der Bildungshintergrund ist also für beide Gruppen im

Mittel als ähnlich zu bewerten, wobei die Gruppe der Schüler/-innen mit Migrationshintergrund eine heterogenere Verteilung aufweisen. In Bezug zum Sprachgebrauch ergibt sich für die Schüler/-innen ohne Migrationshintergrund ein Mittelwert von 3.90 (SD=0.32) und ein Modalwert von vier, welcher der Kategorie „immer [Deutsch]“ entsprach. Für die Schüler/-innen mit Migrationshintergrund ergab sich im Mittel ein Wert von 3.11 (SD=0.78) und ein Modalwert von 3, welcher der Kategorie „meistens [Deutsch]“ entsprach. Die Schüler/-innen mit Migrationshintergrund sprechen also etwas seltener deutsch in der Familie als die Schüler/-innen ohne Migrationshintergrund und ihre Gruppe ist wie auch in Bezug auf den Bildungshintergrund heterogener als die Gruppe der Schüler/-innen ohne Migrationshintergrund.

Migrationshintergrundes und Benotung

Die Ergebnisse der Regressionsanalysen zur Prüfung des statistischen Effektes des Migrationshintergrundes auf die Benotung sind in Tabelle 2 ersichtlich. Hierbei gliedern sich die Analysen in drei Modelle pro Note.[3]

Modell 1: Zur Rolle des Migrationshintergrundes

In Übereinstimmung mit der deskriptiven Betrachtung der Mittelwertsunterschiede zeigte sich für alle untersuchten Noten, dass Schüler/-innen mit Migrationshintergrund, sowohl bei Klassenarbeiten als auch bei Zeugnisnoten, signifikant schlechtere Zensuren erhalten als Schüler/-innen ohne Migrationshintergrund. Die Unterschiede bewegten sich im Bereich von 0.2 bis 0.3 Notenstufen. Der durch den Migrationshintergrund vorhergesagte Beitrag in der Varianz der Noten betrug 1-2 Prozent und war durchgängig statistisch signifikant.

Modelle 2: Bedeutung der allgemeinen Testleistungen[4]

Bei Einbezug von Leistungen in standardisierten mathematischen Tests als Prädiktor zeigte sich zunächst, dass diese Testleistungen ein bedeutsamer Prädiktor für die Zensuren waren – Schüler/-innen mit höheren Testleistungen erzielten bessere (d.h. numerisch niedrigere) Zensuren. Es zeigte sich aber auch, dass der Effekt des Migrationshintergrunds auf die Zensuren bei Kontrolle der

Leistungen in standardisierten Tests weiterhin statistisch signifikant blieb. Die statistische Stärke dieses Effektes war gegenüber dem Effekt in Modell 1 etwas geringer, jedoch durchgängig signifikant. Diese statistisch signifikanten Effekte bei Kontrolle der Leistungsfähigkeit sind ein Hinweis auf leistungsunabhängige Disparitäten bei der Benotung in Abhängigkeit des Migrationshintergrunds. Die Modelle unter zusätzlichem Einschluss der Testleistung erklärten zwischen 11 und 25 Prozent der Varianz in den Zensuren.

Einfluss der Kontrollvariablen

Die in Modell 3 verwendeten Kontrollvariablen soziale Herkunft und Häufigkeit des häuslichen Gebrauchs von Deutsch waren nicht in allen Fällen statistisch signifikante Prädiktoren. Dort, wo dies der Fall war, erzielten Schüler/-innen, deren Eltern höhere Bildungsabschlüsse hatten und in deren Zuhause häufiger Deutsch gesprochen wurde bessere (numerisch niedrigere) Noten. Für alle betrachteten abhängigen Variablen blieb der Effekt des Migrationshintergrundes ein statistisch signifikanter Prädiktor. Somit zeigte sich, dass die in M1 und M2 festgestellten Disparitäten bei Schüler/-innen mit Migrationshintergrund auch nicht auf zusätzliche gegebenenfalls mit dem Migrationshintergrund verbundenen Variablen, wie beispielsweise ein unterschiedlicher häuslicher Sprachgebrauch oder andere soziale Herkunftskontexte, zurückführen lassen. Die Anteile aufgeklärter Varianz liegen zwischen 13 und 25 Prozent.

Um geschlechtsspezifische Unterschiede ausschließen zu können, wurde in weiterführenden Regressionsanalysen für das Geschlecht kontrolliert. Die Analysen entsprechen denjenigen welche unter Modell 3 beschrieben wurden unter zusätzlicher Hinzunahme des Geschlechts als Kovariate. Die Ergebnisse verändern sich durch die Kontrolle des Geschlechtes nicht. Darüber hinaus hat das Geschlecht ähnlich wie die Kontrollvariablen „Sprachgebrauch“ und „soziale Herkunft“ keinen durchgängig signifikanten Einfluss auf die Noten der Schüler/-innen.

Effekte im Zeitverlauf

Bei deskriptiver Betrachtung der Entwicklung der Notenunterschiede im Zeitverlauf (vgl. Tab.1) ist anzumerken, dass sich die Noten der Schüler/-innen mit und ohne Migrationshintergrund nicht anglichen. Die Schüler/-innen starteten mit kaum unterschiedlichen Noten aus der Grundschule in die weiterführende Schule. Zu Beginn der fünften Klassen erhielten die Schüler/-innen mit Migrationshintergrund schlechtere Noten bei gleichen Leistungen in standardisierten Tests. Diese Unterschiede verblieben bis zum Ende der sechsten Klasse auch unter Kontrolle der sozialen Herkunft, des Sprachgebrauchs und der Testleistungen stabil. Die Unterschiede zwischen den Zeugnis- und Klassenarbeitsnoten der Schüler/-innen mit und ohne Migrationshintergrund waren zu jedem Zeitpunkt über den Zeitverlauf hinweg auf dem 1%- Niveau signifikant.

Zur näheren Untersuchung des Effektes des Migrationshintergrundes auf die Noten über den Zeitverlauf wurde der der statistische Test zur Gleichheit von Regressionskoeffizienten nach Paternoster verwendet, welcher es ermöglicht Regressionskoeffizienten miteinander zu vergleichen (Paternoster et al., 1998). Um die statistisch sparsamste Methode zu verwenden wurde dieser Test für den größten Abfall in den Regressionsgewichten (Klassenarbeit Mitte der fünften Klasse [b=0.20] vs. Zeugnisnote Beginn der sechsten Klasse [b =0.12]) sowie den größten Anstieg (Zeugnisnote Beginn der sechsten Klasse [b=0.12] vs. Klassenarbeitsnote Ende der sechsten Klasse [b=0.29]) in den Regressionsgewichten berechnet, um so zu prüfen, ob die gefundenen Unterschiede in den Noten kleiner oder größer werden oder aber stabil verbleiben. Hierbei ergaben sich weder für den größten Abfall noch für den größten Anstieg statistisch signifikante Unterschiede, was dafür spricht, dass die Unterschiede in den Noten zwischen Schüler/-innen mit und ohne Migrationshintergrund im Zeitverlauf stabil verblieben und weder bedeutsam abfallen noch bedeutsam anstiegen.

Diskussion

Bedeutung des Migrationshintergrundes

Das Ziel der vorliegenden Untersuchung bestand darin, die Rolle des Migrationshintergrundes eines Schülers beziehungsweise einer Schülerin für die Notengebung durch die Lehrkraft zu untersuchen. Dabei sollten Ursachen für die Disparitäten nach dem Migrationshintergrund untersucht werden. Die Ergebnisse zeigen, dass der Migrationshintergrund der Schüler/-innen auch unter Kontrolle der Leistungen in standardisierten Tests einen Effekt auf die Benotung durch die Lehrkräfte hat und dass dieser Effekt auch über den Effekt der Bildung der Eltern und des Sprachgebrauchs in der Familie hinausgeht: Die Benotung fiel für Schüler/-innen mit Migrationshintergrund durchgängig statistisch signifikant schlechter aus als für Schüler/-innen ohne Migrationshintergrund. Dies stellt einen Indikator für sekundäre migrationsbedingte Disparitäten in der Notengebung durch die Lehrkräfte dar.

Die Entwicklung migrationsbedingter Disparitäten nach dem ersten, im deutschen Schulkontext besonderen, Übergang wurde in bisheriger Forschung zwar betrachtet und Ungleichheiten konnten nachgewiesen werden (OECD, 2011; OECD, 2014), Kontrollvariablen, wie Leistungen in standardisierten Tests zur Operationalisierung des objektiven Leistungsniveaus und Veränderungen *innerhalb* von Schulformen wurden dabei allerdings unzureichend berücksichtigt. Forschung, welche Schüler/-innen höherer Klassenstufen betrachtet findet sich außerdem vor allem im US-amerikanischen Raum (Smith et al., 1999; West & Anderson, 1976). Das deutsche Schulsystem mit seiner frühen Segregation und einer anderen Migrationsgeschichte nimmt hier allerdings eine Sonderstellung ein. Bisher wurde insbesondere die sensible Phase *direkt* nach dem Übergang in die Sekundarstufe wenig betrachtet.

Die vorliegenden Befunde zeigen in diesem Zusammenhang, dass auch direkt nach der Realisierung eines bedeutsamen Bildungsüberganges und *innerhalb* einer Schulform - in diesem Fall dem Übergang von der Grundschule zum Gymnasium - Disparitäten zwischen Leistungen in standardisierten Tests der Schüler/-innen und der Notengebung durch Lehrkräfte existieren und keinesfalls an Relevanz verlieren.

Bedeutung im Zeitverlauf

Die vorliegenden Analysen weisen auf eine zeitliche Stabilität der Notenunterschiede vom Anfang der fünften Klasse bis zum Ende der sechsten Klasse hin. Die vergleichenden Analysen der Regressionskoeffizienten ergaben keinen statistisch bedeutsamen Anstieg oder Abfall in dem Einfluss des Migrationshintergrundes auf die Notengebung durch die Lehrkraft. Diese Ergebnisse lassen darauf schließen, dass der Migrationshintergrund nicht nur zu Beginn der fünften Klasse einen Einfluss auf die Note hat, sondern bis zum Ende der sechsten Klasse beständige Effekte dieser Herkunft verbleiben. Daraus ist abzuleiten, dass die Bedeutung der migrationsbedingten (wahrscheinlich erwartungsbasierten) Disparitäten nicht wie von einigen Forscherinnen und Forschern nahegelegt (Rosenthal, 1974; West & Anderson, 1976) mit einem längeren Kontakt zwischen den Lehrkräften und Schülern bzw. Schülerinnen abnimmt sondern weiterhin Bestand hat.

Gerade aufgrund der nachgewiesenen zeitlichen Stabilität sind die Effekte der Migrationsherkunft nicht zu vernachlässigen und auch nicht allein durch geringe Vorinformationen über die Schüler/-innen beim Kontakt mit einer neuen Schülergruppe zu erklären. Es ist zu vermuten, dass es auch nach längerem Kontakt mit einem Schüler/einer Schülerin immer wieder Situationen gibt, in denen die zu beurteilende Leistung ambigue ist und in denen sich die Information über die Migrationsherkunft niederschlagen kann.

Die in der vorliegenden Studie gefundenen Effekte bewegen sich in einem für den Forschungsgegenstand der Erwartungseffekte bisher in einschlägigen Studien ebenso berichteten Bereich (Jussim & Harber, 2005). Bei der Bewertung der für den Migrationseffekt insgesamt als gering zu bezeichnenden Effektstärken ist zu bedenken, dass auch kleine numerische Unterschiede im Rahmen der Notengebung große Konsequenzen haben können, vor allem, wenn diese zeitlich stabil verbleiben. Unterschiede in der Größenordnung wie sie in der vorliegenden Untersuchung gefunden werden (im Bereich von 0.2 bis 0.3 Notenstufen), können vor allem bei der Integration von Teilleistungen in ein umfassendes Urteil (z. B. Entscheidung über Fragen der Versetzung oder

des Wechsels der Schulform) eine Rolle spielen. Gerade wenn eine beurteilende Lehrkraft stark zwischen zwei Notenalternativen schwankt, kann schon ein kleiner Urteilsfehler in Bezug auf eine Teilleistung den Ausschlag für ein anderes Urteil geben. Sollten solche Unterschiede bis zum Abschluss der weiterführenden Schule (in diesem Fall der allgemeinen Hochschulreife) stabil verbleiben, so könnten sie einen weiteren Einfluss auf die Bildungslaufbahnen und Bildungschancen nehmen, in dem Bildungszugänge möglicherweise verweigert werden (in diesem Fall der Zugang zu bestimmten Studiengängen) oder ungünstige Bildungsentscheidungen, wie beispielsweise die Entscheidung zu einer Abschulung getroffen werden. Daher ist es wichtig diese Effekte nicht zu vernachlässigen.

Gerade hier ist die Wahl des Gymnasiums als Untersuchungsgegenstand von besonderem Interesse. Das Gymnasium stellt einerseits wie auch die anderen weiterführenden Schulen im Vergleich zu der Primarstufe eine Schulform dar, nach welcher noch weitere potentielle Übergänge folgen (beispielsweise der Hochschulzugang), welche Einfluss auf den Bildungserfolg der Schüler/-innen nehmen und auf die Urteilsfehler innerhalb des Gymnasiums in verschiedener Weise Einfluss nehmen können. Andererseits ist das Bild des Gymnasiums häufig (fälschlicherweise) mit der Annahme verknüpft eine vergleichsweise homogene Schülerschaft in Bezug auf Leistungsstufen zu beherbergen.

Bedeutung von Erwartungseffekten

Da ein Teil der durch den Migrationshintergrund vorgesagten Ungleichheit in den Bildungsergebnisse weder auf Leistungen in standardisierten Tests, den Sprachgebrauch noch auf die soziale Herkunft der Schüler/-innen zurückzuführen ist, deuten die vorliegenden Befunde darauf hin, dass es im Prozess der Notengebung zu systematischen Verzerrungen kommt. Einen möglichen Einflussfaktor bei der Notengebung und Erklärungsfaktor für die fehlerbehaftete Bewertung könnten falsche Erwartungen der Lehrkraft über das Leistungspotential von Schülern bzw. Schülerinnen mit Migrationshintergrund darstellen.

Auch wenn auf Grundlage der vorliegenden Daten keine Aussagen darüber getroffen werden können, ob tatsächlich die Erwartungen der Lehrkräfte die unterschiedliche Notengebung bedingen, können die Ergebnisse als Bestärkung gesehen werden, die Forschung dahingehend fortzusetzen diese verbleibenden mit dem Migrationshintergrund zusammenhängenden Disparitäten in der Leistungsbewertung von Schüler/-innen mit Migrationshintergrund weiter aufzuklären.

In diesem Zusammenhang wird die Betrachtung solcher Effekte auf dem Gymnasium als konservative Schätzung angesehen, da Lehrkräfte hier mit Schüler/-innen im guten Leistungsbereich konfrontiert sind und vermutet wird, dass in diesem Kontext weniger Spielraum für die Aktivierung negativer Leistungserwartungen geboten sein sollte (Neal et al., 2003). Dass die Effekte auch in diesem Kontext gefunden werden, spricht für die Relevanz der Effekte.

Limitationen

Obwohl es sich bei den verwendeten standardisierten Tests um curricular valide Tests handelt, bilden sie nicht das komplette Spektrum der Leistung eines Schülers/einer Schülerin ab und sind damit auch keine perfekte Operationalisierung des tatsächlichen Leistungsniveaus. So ist beispielsweise das Ausmaß mündlicher Mitarbeit ein wichtiger (für die Zeugnisnote in dem entsprechen-den Schuljahr auch relevanter) Leistungsaspekt, der aber vermutlich auch andere Leistungsfacetten wider spiegelt als das mittels standardisiertem Test gemessene Leistungsniveau. Insofern kann aus der Tatsache, dass selbst nach Kontrolle von Testleistungen ein Effekt des Migrationshintergrundes auf die Noten gefunden werden kann, nicht zwingend auf einen Bias im Lehrerurteil geschlossen werden, weil die Noten unter Umständen auch andere Leistungsaspekte als die Tests valide abbilden und sich Schüler/-innen mit und ohne Migrationshintergrund in eben diese Facetten unterscheiden. Auffallend ist allerdings, dass in der vorliegenden Studie ein Effekt des Migrationshintergrundes durchgängig und sowohl für Klassenarbeits- wie für Zeugnisnoten gefunden werden kann. Zumindest bei ersteren spielt die durch die Leistungstests nicht abgebildete mündliche Mitarbeit keine Rolle, weshalb wir einen Bias als Grund für die Mittelwertsunterschiede

in Abhängigkeit des Migrationshintergrunds für wahrscheinlich halten. In weiterführender Forschung ist es aber sinnvoll, auch andere noch stärker auf die lokalen Curricula hin zugeschnittene standardisierte Leistungsindikatoren zu verwenden, um Urteilsfehler nachweisen zu können. Alternativ – wenn auch weniger extern valide – bieten sich experimentelle Studie an, bei denen die zu beurteilende Schülerleistung konstant gehalten und lediglich der Migrationshintergrund variiert wird (Glock, Krolak-Schwerdt, Klapproth & Böhmer, 2013).

Durch die Entscheidung Erwartungseffekte *innerhalb* einer Schulform (in diesem Fall dem Gymnasium) zu untersuchen können auf Grundlage des vorliegenden Datensatzes nur Aussagen über eine Gymnasialstichprobe getroffen werden. Um allgemeingültige Aussagen über den Verlauf der Notengebung in anderen Schulformen nach dem ersten Bildungsübergang treffen zu können, wäre es wünschenswert auch Effekte in anderen Schultypen genauer zu untersuchen. Möglicherweise werden beispielsweise in Hauptschulen negative leistungsbezogene Stereotype noch stärker aktiviert als in anderen Schultypen, da schon bei der Zuweisung zu diesem Schultyp ein eher geringes Leistungsniveau betont wird, welches in Verbindung mit den Informationen über den Migrationsstatus zu stärkeren Effekten führen könnte. Auf Gymnasien könnte die Notwendigkeit eines besonders hohen Leistungsniveaus mit den Informationen über den Migrationsstatus in Konflikt treten und Effekte falscher Erwartungen möglicherweise abschwächen. Diese Überlegungen legen nahe, dass die Effekte falscher Erwartungen je nach Schulform variieren könnten.

Darüber hinaus wurden in dieser Arbeit nur Effekte in Bezug auf Noten und Testleistungen im Fach Mathematik betrachtet. So können nur Aussagen über Effekte im Fach Mathematik getroffen werden. Mathematik stellt ein vergleichsweise klar zu beurteilendes Fach dar (Ziegenspeck, 1999), weshalb zu vermuten ist, dass Urteilsfehler aufgrund des Migrationshintergrundes in anderen Fächern größer sein könnte. Es wäre also wünschenswert die Effekte auch in anderen Schulfächern zu untersuchen.

Auch eine Differenzierung der Gruppe der Schüler/-innen mit Migrationshintergrund nach Herkunftsgruppen verschiedenen Herkunftsländern wäre in zukünftigen Studien interessant und wichtig, um näher auf die Bedeutung stereotypabhängiger Erwartungen in Bezug zu verschiedenen Herkunftsgruppen und mit diesen Gruppen verbundenen spezifischen Stereotypen einzugehen.

Allgemeines Fazit

Zusammenfassend ist zu sagen, dass migrationsbedingte Disparitäten im Schulkontext in der vorliegenden Studie auch für die Sekundarstufe I und damit die Zeit nach dem ersten Bildungsübergang nachgewiesen werden konnten. Effekte des Migrationsstatus sind also nicht nur punktuell (z.B. im Zuge des ersten Bildungsüberganges) oder in den Anfängen der Bildungslaufbahn (in der Primarstufe) relevant sondern stellen wahrscheinlich ein über einen längeren Zeitraum stabiles Phänomen dar. In diesem Zusammenhang verdeutlichen die vorliegenden Befunde die Relevanz von migrationsbedingten Disparitäten im deutschen Schulsystem. Der Nachweis migrationsbedingter Disparitäten, auch nach einem viel betrachteten Bildungsübergang und auf dem Gymnasium zeigt, dass auch gute Schüler/-innen noch mit Benachteiligungen konfrontiert sind und diese also auch in diesem Kontext nicht an Relevanz verlieren.

Dass sich diese Bildungsdisparitäten auf Basis des Migrationshintergrundes auch unabhängig von Leistungen in standardisierten Tests und dem Sprachgebrauch von Schülern und Schülerinnen zeigen, verdeutlicht nicht nur, dass sie auch nach dem Übergang auf eine weiterführende Schule weiterhin relevant sind, sondern dass sie zumindest anteilig systematische Benachteiligungsprozesse reflektieren.

Gerade auch mit Blick auf aktuelle Diskurse in Bezug zur Flüchtlingsthematik und der damit verbundenen zusätzlich vermehrten Zuwanderung von Menschen aus anderen Kulturkreisen, ist Forschung zu stereotypbasierter Verzerrung bei der Leistungsbeurteilungen von hoher gesellschaftlicher Relevanz. Um zukünftige Bildungsprozesse zu gestalten ist es wichtig diese

Mechanismen zu verstehen und zu durchbrechen. Dies stellt eine wichtige Herausforderung für zukünftige Forschung dar.

Tabelle 1.
Mittelwerte und Standardabweichungen (in Klammern)sowie Effektstärken (d) der Mittelwertsdifferenz für Zeugnis- und Klassenarbeitsnoten sowie das objektive Leistungsniveau in Mathematik nach dem Migrationshintergrund

		Schüler/-innen		
		ohne Migrationshintergrund	*mit Migrationshintergrund*	
		M (SD)	*M (SD)*	*d*
Beginn 5.Klasse[1]	*Zeugnis*	1.76 (0.53)	1.84 (0.54)	-0.15
Mitte 5. Klasse	*Klassenarbeit*	1.93 (0.80)	2.12 (0.86)	-0.23
Ende 5.Klasse	*Zeugnis*	2.15 (0.66)	2.32 (0.70)	-0.25
	Klassenarbeit	2.14 (0.85)	2.33 (0.93)	-0.21
Beginn 6.Klasse	*Zeugnis*	2.12 (0.72)	2.33 (0.85)	-0.27
Ende 6.Klasse	*Zeugnis*	2.28 (0.77)	2.55 (0.83)	-0.34
	Klassenarbeit	2.42 (0.91)	2.74 (0.99)	-0.34
		Allgemeine Testleistungen		
HST4/5[2]		24.61 (3.65)	24.12 (3.70)	0.13
DEMAT5+[3]		26.27 (5.46)	25.06 (5.51)	0.29
DEMAT6+[4]		24.03 (5.81)	22.44 (5.62)	0.28

[1] *Endnote aus der 4. Klasse*
[2] *erfasst zu Beginn der 5.Klasse*
[3] *erfasst zum Ende der 5. Klasse*
[4] *efasst zum Ende der 6.Klasse*

Tabelle 2.
Einfluss des Migrationshintergrundes, des objektiven Leistungsniveaus[1] sowie verschiedener Kontrollvariablen auf die Noten getrennt nach den Messzeitpunkten[2,3]
***p<0.01 *p<0.05*

	Abhängige Variable	Noten					
	Messzeitpunkt	Mitte 5. Klasse	Ende 5.Klasse		Beginn 6. Klasse	Ende 6. Klasse	
		Klassenarbeit	Zeugnis	Klassenarbeit	Zeugnis	Zeugnis	Klassenarbeit
	Koeffizienten	b (SE)	b (SE)	b (SE)	b (SE)	b (SE)	b (SE)
Modell	Prädiktoren	ß	ß	ß	ß	ß	ß
M1	Migrationshintergrund	0.19(0.04)**	0.17(0.04)**	0.18(0.04)**	0.21(0.06)**	0.27(0.06)**	0.31(0.08)**
		0.10**	0.11**	0.09**	0.12**	0.14**	0.14**
	R^2	0.01	0.01	0.01	0.02	0.02	0.02
M2	HST 4/5	-0.07(0.01)**	-	-	-	-	-
		-0.32**	-	-	-	-	-
	DEMAT5+	-	-0.05(0.00)**	-0.05(0.00)**	-0.06(0.00)**	-	-
		-	-0.36**	-0.33**	-0.40**	-	-
	DEMAT6+	-	-	-	-	-0.07(0.00)**	-0.08(0.00)**
		-	-	-	-	-0.48**	-0.46**
	Migrationshintergrund	0.15(0.04)**	0.12(0.03)**	0.11(0.04)**	0.14(0.05)**	0.16(0.06)**	0.19(0.07)**
		0.08**	0.07**	0.06**	0.08**	0.08**	0.08**
	R^2	0.11	0.14	0.11	0.18	0.25	0.23
M3	Soziale Herkunft	-0.09 (0.03)**	-0.09(0.02)**	-0.09(0.02)**	-0.10(0.00)*	-0.03(0.04)	-0.07(0.04)
		-0.08**	-0.10**	-0.10(0.02)**	-0.10*	-0.03	-0.06
	Sprachgebrauch	-0.05 (0.04)**	-0.04(0.03)	-0.05(0.04)	-0.10(0.05)*	-0.07(0.06)	-0.05(0.05)
		-0.04**	-0.04	-0.03	-0.08*	-0.05	-0.03
	HST 4/5	-0.07 (0.01)**	-	-	-	-	-
		-0.32**	-	-	-	-	-
	DEMAT5+	-	-0.05(0.00)**	-0.05(0.00)**	-0.05(0.00)**	-	-
		-	-0.37**	-0.34**	-0.40**	-	-
	DEMAT6+	-	-	-	-	-0.06(0.01)**	-0.07(0.01)**
		-	-	-	-	-0.45**	-0.42**
	Migrationshintergrund	0.20(0.06)**	0.16(0.04)**	0.12(0.05)*	0.12(0.07)*	0.17(0.08)*	0.29(0.08)**
		0.10**	0.10**	0.06*	0.07*	0.09*	0.13**
	R^2	0.13	0.17	0.13	0.19	0.25	0.24

[1]Objektives Leistungsniveau: HST 4/5 (gemessen zu Beginn der 5. Klasse), DEMAT5+ (gemessen zum Ende der 5. Klasse), DEMAT6+ (gemessen zum Ende der 6. Klasse)
[2]Aus Gründen der direkten Interpretierbarkeit in Bezug auf die Änderungen in den Noten werden in dieser Tabelle neben den standardisierten Koeffizienten(ß-Werte) auch unstandardisierte Koeffizienten (b-Werte) berichtet. Bei der vergleichenden Betrachtung des Einflusses der Prädiktoren ist die unterschiedliche Metrik dieser zu beachten.
[3]Alle Koeffizienten unter Berücksichtigung der Mehrebenenstruktur (gerechnet in Mplus 7.2)

ESM 1.: Korrelationskoeffizienten nach Pearson der Leistungen in den standardisierten Tests und Klassenarbeits- sowie Zeugnisnoten

		HST[1]	DEMAT5+[2]	DEMAT6+[3]	Mitte 5. Klasse	Ende 5. Klasse	Ende 5. Klasse	Beginn 6. Klasse	Ende 6. Klasse	Ende 6. Klasse
					Klassenarbeit	Zeugnis	Klassenarbeit	Zeugnis	Zeugnis	Klassenarbeit
HST		-	-	-	-	-	-	-	-	-
DEMAT5+		0.448**	-	-	-				-	-
DEMAT6+		0.406**	0.480**	-	-	-	-	-	-	-
Mitte 5. Klasse	Klassen-arbeit	-0.331**	-0.372**	-0.368**	-	-	-	-	-	-
Ende 5. Klasse	Zeugnis	-0.351**	-0.373**	-0.456**	0.649**	-	-	-	-	-
Ende der 5. Klasse	Klassen-arbeit	-0.345**	-0.332**	-0.296**	0.435**	0.677**	-	-	-	-
Beginn 6. Klasse	Zeugnis	-0.362**	-0.408**	-0.392**	0.541**	0.621**	0.545**	-	-	-
Ende 6. Klasse	Zeugnis	-0.341**	-0.369**	-0.478**	0.477**	0.554**	0.498**	0.609**	-	-
Ende 6. Klasse	Klassen-arbeit	-0.368**	-0.409**	-0.475**	0.413**	0.441**	0.477**	0.511**	0.707**	-

[1] erfasst zu Beginn der 5. Klasse [2] erfasst zum Ende der 5. Klasse [3] erfasst zum Ende der 6. Klasse

Literaturverzeichnis

Anderson, N. H. & Jacobson, A. (1965). Effect of stimulus inconsistency and discounting instructions in personality impression formation. *Journal of Personality and Social Psychology*, *2*, 531-539.

Autorengruppe Bildungsberichterstattung (2016). *Bildung in Deutschland 2016: Ein indikatorengestützter Bericht mit einer Analyse zur Bildung und Migration*. Bielefeld: W. Bertelsmann Verlag.

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

Brophy, J. (1983). Research on the self-fulfilling prophecy and teacher expectations. *Journal of Educational Psychology, 75,* 631–661.

Cohen, J., Cohen, P., West, S. G. & Aiken, L. S. (2003). *Applied multiple Regression/correlation analysis for the behavioral sciences.* Mahwah: Lawrence Erlbaum.

Darley, J. M. & Gross, P. H. (1983). A hypothesis-conforming bias in labeling effects. *Journal of Personality and Social Psychology*, *44*, 590-598.

Dickhäuser, O. & Plenter, I. (2005). "Letztes Halbjahr stand ich zwei". Zur Akkuratheit selbst berichteter Noten. *Zeitschrift für Pädagogische Psychologie*, *19*, 219-224.

Enders, C. K. (2001). A primer on maximum likelihood algorithms available for use with missing data. *Structural Equation Modeling, 8,* 128–141.

Esser, H. (2006). *Sprache und Integration. Die sozialen Bedingungen und Folgen es Spracherwerbs von Migranten*. Frankfurt a.M.: Campus.

Fiske, S. T. & Neuberg, S. L. (1990). A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. *Advances in experimental social psychology*, *23*, 1-74.

Glock, S., Krolak-Schwerdt, S., Klapproth, F. & Böhmer, M. (2013). Beyond judgment bias: How students' ethnicity and academic profile consistency influence teachers' tracking judgments. *Social Psychology of Education, 16*, 555-573.

Götz, L, Lingel, K. & Schneider, W. (2013a). *DEMAT 5+. Deutscher Mathematiktest für fünfte Klassen.* Göttingen: Hogrefe.

Götz, L., Lingel, K. & Schneider, W. (2013b). *DEMAT 6+. Deutscher Mathematiktest für sechste Klassen.* Göttingen: Hogrefe.

Hadley, S. T. (1995). Feststellungen und Vorurteile in der Zensierung. In K. Ingenkamp (Hrsg.) *Die Fragwürdigkeit der Zensurengebung.* (S.159-166), Weinheim/Basel: Beltz.

Hayden, T. & Mischel, W. (1976). Maintaining trait consistency in the resolution of behavioral inconsistency: The wolf in sheep's clothing? *Journal of Personality*, *44*, 109-132.

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

Jones, E. E. & Goethals, G. (1972). Order effects in impression formation: Attribution context and the nature of the entity. In E. E. Jones, D.E. Kanouse, H.H.Kelley, R.E. Nisbett, S. Valins & B. Weiner (Eds.), *Attribution: Perceiving the causes of behavior (p.27-46).*Morristown, NJ: General Learning Press.

Jussim, L. (1986). Self-fulfilling prophecies: A theoretical and integrative review. *Psychological Review, 93,* 429–445.

Jussim, L. (1990). Social reality and social problems: The role of expectancies. *Journal of Social Issues*, *46*, 9-34.

Jussim, L. & Eccles, J. (1992). Teacher expectations II: Construction and reflection of student achievement. *Journal of Personality and Social Psychology*, *63*, 947-961.

Jussim, L., Eccles,J. & Madon,S. (1996). Social perception, social stereotypes, and teacher expectations: Accuracy and the quest for the powerful self-fulfilling Prophecy. In Zanna, M.P. (Eds.), *Advances on experimental social psychology*. *28*, 281-388.

Jussim, L. & Harber, K. (2005). Teacher expectations and self-fulfilling prophecies: Known and unknowns, resolved and unresolved controversies. *Personality and Social Psychology Review*, 9, 131-155.

Kahraman, B. & Knoblich, G. (2000). Stechen statt Sprechen. Valenz und Aktivierbarkeit von Stereotypen über Türken. *Zeitschrift für Sozialpsychologie. 31*, 31-43.

Kristen, C. & Dollmann, J. (2010). Sekundäre Effekte der ethnischen Herkunft: Kinder aus türkischen Familien am ersten Bildungsübergang. In B. Becker & D,. Reimer, *Vom Kindergarten bis zur Hochschule* (S.117-144). VS Verlag für Sozialwissenschaften.

Lingle, J. H. & Ostrom, T. M. (1981). Principles of memory and cognition in attitude formation. In R. E. Petty, T. M. Ostrom, & T. C. Brock (Eds.), *Cognitive responses in persuasive communications: A text in attitude change* (p. 399-420). Hillsdale, NJ: Erlbaum.

Mietzel, G. & Willenberg, H. (2000). *HST 4/5. Hamburger Schulleistungstest für 4. und 5. Klassen.* Göttingen: Hogrefe.

Muthén, L. K. & Muthén, B. O. (2010). *Mplus User's Guide,* Los Angeles: Muthen & Muthen.

Neal, L. I., McCray, A. D., Webb-Johnson, G. & Bridgest, S. T. (2003). The effects of African American movement styles on teachers' perceptions and reactions. *The Journal of Special Education*, *37*, 49–57.

Nickerson, R. S. (1998). Confirmation Bias. *Review of General Psychology, 2,* 175-220.

Nisbett, R. E. & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgement.* Englewood Cliffs, NJ: Prentice-Hall

OECD (2011). *PISA 2009: Potenziale nutzen und Chancengerechtigkeit sichern*, OECD Publishing, Paris.

OECD (2014). *PISA 2012 Ergebnisse: Exzellenz durch Chancengerechtigkeit (Band II): Allen Schülerinnen und Schülern die Voraussetzungen zum Erfolg sichern*, W. Bertelsmann Verlag, Bielefeld.

Paternoster, R., Brame, R., Mazerolle, P. & Piquero, A. (1998). Using the correct statistical test for the equality of regression coefficients. *Criminology,* 36, 859-866.

Rosenthal, R. (1974). *On the social psychology of the self-fulfilling prophecy: Further evidence for Pygmalion effects and their mediating mechanisms.* New York: MSS Modular.

Rosenthal, R. & Jacobson, L. (1968). *Pygmalion in the classroom: teacher expectations and students intellectual development.* New York: Holt.

Ross, L. & Anderson, C. (1982). Shortcomings in the attribution process: On the origins and maintenance of erroneous social assessments. In A. Tversky, D. Kahneman, & P. Slovic (Eds.), *Judgement under uncertainty: Heursitics and biases (p.129-152),* Cambridge, England: Cambridge University Press.

Ross, L. & Lepper, M. R. (1980). The perseverance of beliefs: Empirical and normative considerations. In R. Shweder & D. Fiske (Eds.), *New directions for methodology of social and behavioral science: Fallibel judgement in behavioral research* (Vol. 4, pp.17-36). San Francisco: Jossey-Bass.

Schnabel, K. & Schwippert, K. (2000). Schichtenspezifische Einflüsse am Übergang auf die Sekundarstufe II. In J. Baumert, W. Bos & R. Lehmann (Hrsg.), *TIMSS/ III. Dritte Internationale Mathematik- und Naturwissenschaftsstudie. Mathematisch-naturwissenschaftliche Bildung am Ende der Schullaufbahn. Band I* (S. 261-281). Opladen: Leske & Budrich.

Sherman, S. J., Zehner, K. S., Johnson, J. & Hirt, E. R. (1983). Social explanation: The role of timing, set, and recall on subjective likelihood estimates. *Journal of Personality and Social Psychology*, *44*, 1127-1143.

Smith, A. E.; Jussim, L. & Eccles, J. (1999). Do self-fulfilling prophecies accumulate, dissipate, or remain stable over time? *Journal of Personality and Social Psychology*, *77*, 548-565.

Statistisches Bundesamt (2014). *Bevölkerung und Erwerbstätigkeit. Bevölkerung mit Migrationshintergrund. Ergebnisse des Mikrozensus 2013 (*Fachserie 1. Reihe 2.2.).Wiesbaden: Statistisches Bundesamt.

Stanat, P. & Christensen, G. S. (2006).*Where immigrant students succeed: A comparative review of performances and engagement in PISA 2003.* Paris: OECD.

Trautwein, U. & Baeriswyl, F. (2007). Wenn leistungsstarke Klassenkameraden ein Nachteil sind: Referenzgruppeneffekte bei Übertrittsentscheidungen. *Zeitschrift für Pädagogische Psychologie*, 21, 119-133.

Walter, O. (2009). Herkunftsassoziierte Disparitäten im Lesen, der Mathematik und den Naturwissenschaften: ein Vergleich zwischen PISA 2000, PISA 2003 und PISA 2006. In M. Prenzel & J. Baumert (Hrsg.), *Vertiefende Analysen zu PISA 2006* (S.149-168). VS Verlag für Sozialwissenschaften.

West, C. & Anderson, T. (1976). The question of preponderant causation in teacher expectancy research. *Review of Educational Research, 46*, 613-630.

Ziegenspeck, J. W. (1999). *Handbuch Zensur und Zeugnis in der Schule.* Bad Heilbrunn: Klinkhardt.

Fußnoten

[1]Der DEMAT 5+ wurde in einer Vorversion eingesetzt, der uns vor der Veröffentlichung zur Verfügung gestellt wurde und in wenigen Details von der endgültigen Fassung abweicht.

[2]Für die zum ersten Messzeitpunkt erfasste Note (Note in Mathematik im Abschlusszeugnis der Grundschule) wurden dieses Analysen nicht berechnet, da für diese Zensur die ehemalige Art der Zusammensetzung der Grundschulklassen – eine Gruppenzugehörigkeit, für die in den Analysen zu kontrollieren wäre – nicht bekannt ist.

[3]Zusätzlich zu den angeführten Analysen, führten wir auch weitere Regressionsanalysen (eine für jede Note) durch, in welchen der Migrationshintergrund dreistufig operationalisiert wurde (kein Elternteil, ein Elternteil, beide Elternteile außerhalb Deutschlands geboren) und als Dummy-Variablen mit der Referenzkategorie „kein Elternteil“ eingeschlossen wurde. Im Rahmen der Analysen wurden analog zu Modell 3 die soziale Herkunft, der häusliche Sprachgebrauch und Leistungen in standardisierten Tests kontrolliert. Hierbei ergab sich, ein statistisch höherer signifikanter Einfluss des Migrationshintergrundes auf die Note für die Schüler/-innen deren beide Elternteilen nicht in Deutschland geboren waren ($b=0.26$-0.40; $p<0.001$), während sich bei Schülern/Schülerinnen mit nur einem Elternteil aus dem Ausland ein ebenfalls statistisch signifikanter aber geringer zu bewertender Einfluss des Migrationshintergrundes ($b=0.10$-0.20; $p<0.05$) fand.

[4]In weiterführenden Regressionsanalysen (entsprechend Modell 3 unter Kontrolle der sozialen Herkunft und des häuslichen Sprachgebrauchs) zeigte sich, dass Schüler/-innen mit Migrationshintergrund bei gleichen Noten höhere standardisierte Testleistungen erreichen. Nach Migrationshintergrund getrennte Regressionsanalysen zeigten auf, dass Schüler/-innen mit Migrationshintergrund mehr Punkte in den standardisierten Testleistungen im Vergleich zu Schülern/Schülerinnen ohne Migrationshintergrund erreichen mussten, um die gleiche Note zu erhalten (Vergleich des Achsenabschnitts). Die Punktzahlen, die die Schüler/-innen mit beziehungsweise ohne Migrationshintergrund jeweils für einen Notensprung erreichen

mussten, waren dabei signifikant voneinander verschieden ($p<0.05$). Diese Befunde waren zu jedem Messzeitpunkt statistisch signifikant und unterstützen die hier angeführten Ergebnisse.