Forschungsmonitor Schule
Rezensionen für die Praxis

Johannes Rosendahl, Sonja Hensel

Rezension zu

Nennstiel, R. & Gilgen, S. (2024). Does chubby Can get lower grades than skinny Sophie? Using an intersectional approach to uncover grading bias in German secondary schools. *PLoS ONE 19*(7), 1–23.

Kommentierter Kurzbefund

Nennstiel und Gilgen gehen der Frage nach, inwiefern Zeugnisnoten, die Schülerinnen und Schüler in der Sekundarstufe I bekommen, durch Faktoren beeinflusst werden, die nicht unmittelbar mit ihrer Leistung zu tun haben, etwa durch unangepasstes Verhalten der Schülerinnen und Schüler oder Stereotype in den Köpfen der Lehrkräfte. Hierfür nutzen sie Daten von 14 090 Neuntklässlerinnen und Neuntklässlern der 4. Startkohorte des deutschen Nationalen Bildungspanels (NEPS, 2010/2011) und gehen von der Annahme aus, dass Abweichungen in den Leistungsausprägungen zwischen Fächernoten einerseits und den Ergebnissen in domänenspezifischen standardisierten Tests andererseits einen Benotungsbias, also eine verzerrte Benotung belegen.

Die Forschenden untersuchen, wie sich Geschlecht, sozioökonomischer Status (SES), Migrationshintergrund oder Übergewicht auf die Zeugnisnoten in Deutsch, Mathematik, Biologie, Chemie und Physik auswirken. Darüber hinaus analysieren sie, inwiefern die Effekte kumulieren, ob also beispielsweise bei einem übergewichtigen Jungen mit Migrationshintergrund besonders große Abweichungen der Leistungsausprägungen auf Grundlage von Noten und Testergebnissen zu finden sind.

Im Ergebnis steht die Notenvergabe durch die Lehrkräfte in allen 5 Fächern in Zusammenhang mit den untersuchten Merkmalen, am stärksten in Deutsch. Beispielsweise werden Schülerinnen und Schüler mit Migrationshintergrund oder Übergewicht tendenziell schlechter benotet, selbst wenn sie vergleichbare Testleistungen erzielen wie Schülerinnen und Schüler ohne Migrationshintergrund bzw. ohne Übergewicht. Die Effekte der einzelnen Einflussfaktoren kumulieren zudem, so dass sich beispielsweise bei übergewichtigen türkeistämmigen Schülern starke negative Verzerrungseffekte zeigen, bei Mädchen ohne Migrationshintergrund und Übergewicht lassen sich hingegen positive Verzerrungseffekte nachweisen.

Angesichts dessen schlagen Nennstiel und Gilgen vor, die Rolle von Noten und was sie messen sollten zu diskutieren und beispielsweise in Deutsch strukturiertere Bewertungsschemata zu nutzen, um die Benotung genauer und gerechter zu gestalten. Allerdings wurden an anderer Stelle eher positive Diskriminierungseffekte bei der Benotung nachgewiesen, so dass die Benotungsbias vermutlich weniger auf mangelnde Objektivität der Lehrkraftbewertungen zurückzuführen sind, sondern letztlich dürften andere Gründe ausschlaggebend dafür sein, dass benachteiligte Gruppen ihre Lernpotenziale nicht ausschöpfen und ihre Notenleistungen nicht in höherem Maße den Kompetenzen entsprechen, die sie gemäß ihren Testergebnissen erreichen (könnten).

Hintergrund

Ein Ausgangspunkt der Studie von Nennstiel und Gilgen ist die umfangreiche Forschung zur Beeinflussung der Notengebung durch Faktoren, die jenseits der Leistungen der Schülerinnen und Schüler liegen, dem sogenannten „grading bias“ (Benotungsbias). Hier beziehen sie sich auf Ergebnisse zum Einfluss von Geschlecht, ethnischer Herkunft bzw. Migrationshintergrund, sozioökonomischem Hintergrund und Übergewicht.

Zum Thema Geschlecht geben die Forschenden 22 Studien an, deren Mehrheit eine bessere Benotung von Mädchen bei gleichem Abschneiden in standardisierten Tests feststellt. In den naturwissenschaftlichen Fächern und Mathematik finden sich allerdings umgekehrte Effekte.

Eine von der Mehrheitsgesellschaft abweichende ethnische Herkunft weist nach den angeführten 17 internationalen Studien auf eine schlechtere Benotung bei gleicher Leistung hin. Wie beim Einfluss des Geschlechts ist auch hier die Studienlage nicht eindeutig, denn einzelne Studien finden positive und keine Verzerrungseffekte.

Zum Einfluss des sozioökonomischen Status fanden die Forschenden 9 Studien. Diese stellen fast durchweg eine Benachteiligung von Kindern aus sozial schwachen Familien besonders in sprachlichen Fächern fest, einhergehend mit geringeren Erwartungen der Lehrkräfte an die Leistungsfähigkeit dieser Kinder.

Übergewicht bzw. Adipositas scheint zu einem negativen Benotungsbias zu führen. Allerdings stützt sich dieser Befund nur auf 5 Studien.

Für diese Verzerrungseffekte bei der Notengebung führen Nennstiel und Gilgen zwei Erklärungsansätze an, die in der Forschung diskutiert werden. Zum einen spiele das Verhalten im Klassenraum eine Rolle, was eine Erklärung für die negativere Beurteilung von Jungen sein könne, die öfter durch herausforderndes Verhalten auffallen. Zum anderen wurden in Studien Stereotype in den Köpfen der Lehrkräfte identifiziert, die beispielweise zu einer schlechteren Einschätzung der Leistungsfähigkeit von Mädchen im Mathematikunterricht oder von Schülerinnen und Schülern mit Migrationshintergrund im Deutschunterricht führen. Die kognitiven Fähigkeiten von Lernenden mit hohem sozioökonomischem Status würden über-, die von sozial benachteiligten Schülerinnen und Schülern hingegen unterschätzt. Mit Bezug auf übergewichtige Kinder und Jugendliche präge die Vorstellung, dass diese sich stärker anstrengen müssten, dennoch oft weniger Leistung erbrächten und mehr Unterstützung benötigten, die Sicht der Lehrkräfte.

Ein Forschungsdesiderat sehen Nennstiel und Gilgen beim Thema der Intersektionalität, also der kumulativen Benachteiligung von Lernenden, die verschiedenen von Diskriminierung betroffenen Gruppen angehören. Dabei seien die Effekte bislang unklar, da Benachteiligungen einerseits kumulieren könnten, andererseits könnte die Zugehörigkeit zu verschiedenen Gruppen aber auch zu gegenläufigen Effekten führen, die sich gegenseitig neutralisieren, wie bei einem Jungen mit Migrationshintergrund aus einem Elternhaus mit hohem sozioökonomischem Status. Möglich wäre außerdem eine mildere Benotung von Lernenden, die mehreren potenziell diskriminierten Gruppen angehören, was den Benotungsbias aufheben würde.

Vor diesem Hintergrund fokussieren Nennstiel und Gilgen drei Forschungsfragen:

- Wie bedeutsam ist der Benotungsbias im Hinblick auf Geschlecht, sozioökonomischen Status, Herkunft und Übergewicht in Deutschland?
- Variieren die Effekte bei unterschiedlichen Fächern?
- Gibt es Gruppen, die besonders stark betroffen sind, weil sie mehrere nachteilige Faktoren aufweisen?

Design

Dem Forschungsdesign zugrunde liegt ein sogenannter „Notengleichungs-Ansatz“ (grade equation approach), bei dem davon ausgegangen wird, dass Unterschiede zwischen den Leistungsausprägungen auf Grundlage von Noten einerseits und den Ergebnissen in standardisierten Tests andererseits einen Benotungsbias belegen, dass also Eigenschaften von Lernenden wie ihr Geschlecht, sozioökonomischer Status, Migrationshintergrund oder Übergewicht einen verzerrenden Einfluss auf die Notengebung der Lehrkräfte haben.

Stichprobe

Die Forschenden nutzten Daten aus dem Nationalen Bildungspanel (National Education Panel Study, NEPS) des Leibniz-Instituts für Bildungsverläufe. In dieser Langzeit-Bildungsstudie werden seit 2009 in bisher 7 Startkohorten die Bildungsverläufe von u. a. Kindern und Jugendlichen in Deutschland untersucht, um Bedingungen, Ergebnisse und Folgen von Bildungsprozessen zu beschreiben und zu erklären. Diese Kohorten umfassen insgesamt mehr als 70 000 Teilnehmende und 50 000 Begleitpersonen wie Eltern oder Lehrkräfte (vgl. die Homepage des Projekts www.neps-data.de).

Konkret werteten die Forschenden die Daten von 14 090 Schülerinnen und Schülern aus, die zum Befragungszeitpunkt 2011 die 9. Klasse verschiedener Schulformen (keine Förderschulen) besuchten und am Anfang und Ende des Schuljahres mittels eines von ihnen ausgefüllten Fragebogens befragt und mit standardisierten Tests getestet worden waren.

Erhebungsinstrumente

Im Fragebogen gaben die Jugendlichen ihre Zeugnisnoten in Deutsch, Mathematik, Biologie, Chemie und Physik an. Die Leistungen der Schülerinnen und Schüler wurden durch standardisierte Tests zu domänenspezifischen und allgemeinen kognitiven Fähigkeiten gemessen (vgl. zu detaillierteren Informationen die Homepage des Projekts www.neps-data.de).

Der Migrationshintergrund wurde auf zweierlei Weisen operationalisiert: Einerseits wurde anhand der Angaben zu den Geburtsorten der Jugendlichen, ihrer Eltern und Großeltern festgestellt, ob die Jugendlichen einen Migrationshintergrund aufweisen (mindestens 2 Großelternteile im Ausland geboren). Andererseits wurde dieser Migrationshintergrund weiter differenziert in Bezug auf die Herkunftsgebiete der Familien: Türkei, ehemalige Sowjetunion, nordwestliches und südliches Europa, Zentral- und Osteuropa und andere Staaten.

Für die Klassifizierung der Schülerinnen und Schüler als übergewichtig wurde der Body-Mass-Index (BMI) aus den in den Fragebögen erfassten Angaben zu Körpergröße und Gewicht errechnet. Ab einem BMI-Perzentil von 85 galten die Jugendlichen als übergewichtig.

Der sozioökonomische Status (SES) wurde auf Grundlage des International Socio-Economic Index of Occupational Status (ISEI) ermittelt unter Bezug auf die Angaben, die die Befragten zu den Berufen ihrer

Eltern machten, erforderlichenfalls ergänzt durch Angaben aus den Eltern-Interviews. Der ISEI-Index besitzt eine Skala von 10 – 90 (z. B. Erntehelfer: 11; Richter: 89). Für die deskriptive Auswertung wurden die Skalenwerte zum jeweils höchsten ISEI-Wert der Eltern in 10 Perzentile (Dezile) eingeteilt, für die Regressionsanalysen wurden sie z-standardisiert.

Schließlich verwendeten die Forschenden verschiedene Variablen, um das Verhalten im Klassenraum und psychische Merkmale statistisch kontrollieren zu können: das psychologische Modell der „Big Five“, mit dem Grundeigenschaften von Personen beschrieben werden wie *Offenheit*, *soziale Verträglichkeit* oder *Gewissenhaftigkeit* (erhoben durch Selbstauskünfte in dem Fragebogen mit jeweils zwei Items), Skalen aus dem Strengths and Difficulties Questionnaire (SDQ) zur Erhebung von prosozialem Verhalten und Verhaltensproblemen mit Gleichaltrigen, den SCOFF-Fragebogen zu problematischem Essverhalten, die Zufriedenheit mit der eigenen Gesundheit und die Frage, ob die Person eine Klasse wiederholt hatte. Außerdem wurde die besuchte Schulform mitbetrachtet.

Umgang mit fehlenden Werten

In einigen Datensätzen fehlten zu einzelnen Variablen die notwendigen Daten, z. B. zum ISEI in 12.3 %, zum BMI in 15.5 % und zu den Noten in den Naturwissenschaften (zwischen 13.2 % und 18.7 %) – letzteres u. a., weil in einigen Bundesländern oder Schulformen diese Fächer nicht getrennt voneinander unterrichtet werden. Dieser Umstand verringerte in den diesbezüglichen Analysen die Fallzahlen entsprechend. Bei anderen fehlenden Daten wie dem ISEI wurden Daten imputiert, d. h. durch statistische Verfahren auf Grundlage von ähnlichen, vollständig vorliegenden Datensätzen geschätzt und so die Daten vervollständigt.

Auswertung

Die erhobenen Daten wurden deskriptiv und regressionsanalytisch ausgewertet. Mit der deskriptiven Auswertung wurde geprüft, inwiefern überhaupt Unterschiede im Hinblick auf die unabhängigen (vorhersagenden) Variablen (Geschlecht, BMI, sozioökonomischer Status, Migrationshintergrund) in den Fächernoten bestehen. Hierzu wurden stetige Variablen in zwei Kategorien dichotomisiert (z. B. übergewichtig vs. normalgewichtig, 1. SES-Dezil vs. 10. SES-Dezil). Sodann wurden für die beiden Kategorien die Mittelwerte der Fachnoten gebildet und voneinander abgezogen, um den (bivariaten) Benotungsunterschied zu bestimmen. Aufgrund der zuvor durchgeführten z-Standardisierung der Fachnoten, mit der ihr Mittelwert auf 0 und ihre Standardabweichung auf 1 transformiert wird, geben diese Differenzwerte den Benotungsunterschied in Standardabweichungen (*SD*) als Einheit an – beispielsweise bedeutet ein Ergebnis von 0.44 *SD*, dass Mädchen im betroffenen Fach im Schnitt um fast eine halbe Standardabweichung besser benotet werden als Jungen.

Der Schwerpunkt lag jedoch auf den regressionsanalytischen Auswertungen, um die Variation der Fachnoten als abhängige Variablen zu erklären und herauszufinden, inwiefern sich ein Benotungsbias nachweisen lässt und worauf er ggf. zurückgeführt werden kann. Die Daten wurden mit Mehrebenenregressionsanalysen ausgewertet. Dabei handelt es sich um ein statistisches Verfahren für Daten, die eine geschachtelte Struktur aufweisen. Im vorliegenden Fall sind die Schülerinnen und Schüler (Ebene 1) in Klassen (Ebene 2) und diese wiederum in Schulen (Ebene 3) geschachtelt. Mithilfe der Mehrebenenanalyse lässt sich zum einen herausfinden, inwiefern die Benotung bzw. der Benotungsbias zwischen Klassen (z. B. aufgrund der spezifischen Zensurengebung einzelner Lehrkräfte) oder zwischen Schulen variiert. Zum anderen ist die Berücksichtigung der Verschachtelung bei der statistischen Auswertung erforderlich, weil die Effekte der Einflussfaktoren sonst nicht korrekt geschätzt werden.

Für jede der fünf Fachnoten wurden drei Modelle berechnet:

Mit Modell 1 wurde der allgemeine Benotungsbias für das Fach bestimmt. Hierfür wurden das interessierende Merkmal (z. B. Geschlecht oder BMI), der domänenspezifische Kompetenztest (z. B. das Testergebnis in Mathematik), die Ergebnisse des allgemeinen kognitiven Fähigkeitstests und die besuchte Schulform als unabhängige Variablen einbezogen, um die Notenunterschiede zwischen den Schülerinnen und Schülern zu erklären. Für die Fächer Chemie, Biologie und Physik griffen die Forschenden auf einen allgemeinen Kompetenztest für die Naturwissenschaften zurück.

Zur Berechnung von Modell 2 wurden die Merkmale Migrationshintergrund oder Herkunftsgruppe sowie die weiteren Kontrollvariablen (siehe oben) als unabhängige Variablen hinzugenommen, um zu untersuchen, inwieweit psychische Merkmale und Verhalten den Benotungsbias erklären.

In einem dritten Schritt wurden die kumulativen („intersektionalen“) Effekte von sozialer Herkunft, Geschlecht, BMI und Migrationshintergrund bzw. Herkunftsgruppe geprüft. Hierfür wurden zunächst analog zu Modell 1 neben einem Modell mit Haupteffekten für diese Variablen weitere Modelle mit den 2- sowie 4-Wege-Interaktionseffekten berechnet. Mit Interaktionseffekt wird in der Statistik die Beziehung bzw. Wechselwirkung zwischen zwei oder mehr Variablen bezeichnet. Durch das Einbeziehen der Interaktionseffekte wird also untersucht, ob die einzelnen Variablen sich gegenseitig beeinflussen bzw. verstärken. Auf Grundlage der geschätzten Regressionsparameter wurden im Anschluss Vorhersagewerte (predictive margins) ermittelt und daraus Differenzen für Gruppen mit verschiedenen Merkmalskombinationen gebildet. Damit konnten beispielsweise übergewichtige türkeistämmige Schüler mit niedrigem sozialem Status (-1 *SD*) verglichen werden mit normalgewichtigen Mädchen ohne Migrationshintergrund und mit hohem sozialem Status (+1 *SD*).

Ergebnisse

Deskriptive Auswertung

Bei der deskriptiven Auswertung zeigen sich Unterschiede in den z-standardisierten Notenmittelwerten zwischen den Gruppen. Zum Beispiel haben Mädchen signifikant bessere Notenmittelwerte in Deutsch (0.44 *SD*) und Biologie (0.19 *SD*) als Jungen, während Jungen in Mathematik und Physik leicht bessere Notenmittelwerte erzielen. Darüber hinaus haben übergewichtige Schülerinnen und Schüler in allen Fächern ungünstigere Notenmittelwerte als nicht übergewichtige – in Mathematik am wenigsten ausgeprägt und in Deutsch und Biologie am stärksten. In allen Schulfächern gibt es einen deutlichen Unterschied in den Notenmittelwerten zugunsten der privilegierteren Schülerinnen und Schüler (1. vs. 10. SES-Dezil). Zudem haben Schülerinnen und Schüler mit Migrationshintergrund – besonders diejenigen mit Wurzeln in der Türkei – im Mittel schlechtere Noten als Schülerinnen und Schüler ohne – erneut in Deutsch am stärksten ausgeprägt.

Benotungsbias bei Kontrolle der Testleistungen

Bei der regressionsanalytischen Auswertung mit Modell 1 – unter Kontrolle der Variablen allgemeine und domänenspezifische Testleistung – zeigen sich ähnliche Ergebnisse. Bei gleichen Testleistungen haben Mädchen weiterhin einen Vorteil in Deutsch (0.37 *SD*) und Biologie (0.21 *SD*), wohingegen sich die zuvor beschriebene schlechtere Benotung in Mathematik als positive Verzerrung erweist (0.06 *SD*) – d. h., Mädchen erhalten durchschnittlich schlechtere Mathematiknoten als Jungen, aber wenn man ihre Testleistungen berücksichtigt, fallen ihre Noten eher noch zu gut aus. In Physik besteht ein Benotungsbias zugunsten der Jungen (-0.10 *SD*). Ein höherer sozioökonomischer Status ist in allen Fächern verbunden mit einer geringfügig besseren Benotung (0.04 *SD* ≤ 0.08 *SD*). Übergewichtige Lernende werden bei gleichen Testleistungen durchgängig tendenziell schlechter beurteilt, besonders in Deutsch (-0,21 *SD*).

Liegt ein Migrationshintergrund vor, werden, außer in Biologie, im Schnitt schlechtere Noten erzielt, am ausgeprägtesten in Deutsch (-0.14 *SD*), was insbesondere auf Schülerinnen und Schüler aus der Türkei (-0.24 *SD*) und aus der ehemaligen Sowjetunion (-0.19 *SD*) zutrifft.

Benotungsbias bei Kontrolle von Persönlichkeitsmerkmalen und Klassenwiederholung
In Modell 2 wurden zusätzlich Persönlichkeitsmerkmale (siehe Abschnitt Design) und Klassenwiederholungen kontrolliert, um herauszufinden, inwiefern diese Merkmale den Benotungsbias ggf. erklären. Dadurch verringern sich die zuvor gefundenen Vorteile von Mädchen, bleiben aber in Deutsch (0.29 *SD*) und Biologie (0.08 *SD*) signifikant. Für Jungen wachsen im Vergleich zu Modell 1 die Vorteile in Physik (-0.2 *SD*) und nun zeigen sich auch Vorteile in Chemie (-0.09 *SD*). Der positive Benotungsbias zugunsten von Jugendlichen mit einem höheren sozioökonomischen Status bleibt bestehen, während der negative Benotungsbias für übergewichtige Personen außer in Deutsch (-0.1 *SD*) verschwindet. Über alle Fächer und Herkunftsgruppen hinweg verkleinert sich die Benachteiligung von Jugendlichen mit Migrationshintergrund. Dies gilt aber kaum für Deutsch, insbesondere für türkeistämmige Schülerinnen und Schüler (-0.22 *SD*), und für Physik.

Kumulative Effekte (Intersektionalität)
Zur Untersuchung der Intersektionalität wurde anhand der regressionsanalytisch ermittelten Vorhersagewerte beispielsweise der Benotungsbias bei einem übergewichtigen türkeistämmigen Jungen mit geringem sozioökonomischem Status verglichen mit dem eines türkeistämmigen Jungen mit hohem sozioökonomischem Status und ohne Übergewicht. Dabei finden die Forschenden Hinweise darauf, dass die negativen Effekte der einzelnen Variablen kumulieren, dass es also intersektionale Benachteiligung gibt.

Die Ergebnisse der Modelle mit und ohne Interaktionseffekten sind sich recht ähnlich, so dass die Forschenden keine Hinweise auf kompensatorische oder zusätzliche negative Effekte über die additiven hinaus sehen, zumal wenige der zweifachen oder vierfachen Interaktionen statistisch signifikant sind. Die kumulativen Effekte sorgen jedoch für bedeutsame Unterschiede, was den Grad der Notenverzerrung angeht. Beispielsweise erhält ein türkeistämmiger Junge mit geringem sozioökonomischem Status und Übergewicht im Schnitt eine um 0.8 *SD* schlechtere Deutschnote als ein Mädchen ohne Migrationshintergrund und Übergewicht mit einem hohen sozioökonomischen Status (bei vergleichbaren Testleistungen). Zwischen den Schulen gibt es deutliche Unterschiede, wie stark der Benotungsbias in Bezug auf die Variablen Geschlecht und BMI ausgeprägt ist, in geringerem Maße variiert er für Migrationshintergrund, der Einfluss des sozioökonomischen Status besteht klassenübergreifend in ähnlichem Maße.

Insgesamt finden die Forschenden in allen fünf untersuchten Fächern Auswirkungen der untersuchten Merkmale auf die Notenvergabe von Lehrkräften, am stärksten im Fach Deutsch. Die Effekte der einzelnen Variablen kumulieren zudem, so dass beispielsweise bei übergewichtigen türkeistämmigen Schülern starke negative Verzerrungseffekte, bei Mädchen ohne Migrationshintergrund und Übergewicht positive Verzerrungseffekte resultieren.

Diskussion und Einschätzung

Zum Hintergrund
Nennstiel und Gilgen können sich für ihre Studie auf zahlreiche Forschungsergebnisse beziehen, die in einer Mehrzahl darauf hindeuten, dass es „grading bias“ (Benotungsbias) gibt, dass also Schülerinnen und

Schüler trotz vergleichbarer Testleistungen bei der Notenvergabe von Lehrkräften ungleich behandelt werden. Die referierte Forschungslage belegt insbesondere Benotungsbias zugunsten von Mädchen, bei Mathematik und naturwissenschaftlichen Fächern teilweise zugunsten von Jungen sowie zum Nachteil von Schülerinnen und Schülern, die einen niedrigen sozioökonomischen Status aufweisen oder nicht der ethnischen Mehrheitsgesellschaft angehören oder übergewichtig sind.

Dadurch ist die Studie inhaltlich und methodisch nachvollziehbar in einen Forschungskontext eingebettet und stellt einen bedeutsamen Erkenntnisgewinn in Aussicht, denn sie zielt darauf ab, das Ausmaß dieser Benotungsbias anhand einer für Deutschland repräsentativen Stichprobe für die fünf Fächer Deutsch, Mathematik, Chemie, Biologie und Physik zu ermitteln und herauszufinden, inwiefern die Effekte bei Schülerinnen und Schülern, die mehreren der benachteiligten Kategorien angehören (z. B. Übergewichtige mit Migrationshintergrund), kumulieren („Intersektionalität“).

Zum Design

Die Anlage der Untersuchung ist stimmig und methodisch gut gestaltet. Vorteilhaft sind insbesondere die große repräsentative Stichprobe (N = 14 090) aus dem deutschen nationalen Bildungspanel (NEPS) sowie die Zahl der in den Blick genommenen Fächer. Die Forschenden selbst weisen darauf hin, dass die Aussagekraft der Studie dadurch eingeschränkt wird, dass trotz der großen Grundgesamtheit in einigen Gruppen, z. B. übergewichtige Mädchen, nur sehr wenige Individuen zu finden waren, was den Nachweis signifikanter Ergebnisse erschwert.

Deutlich gravierender ist, dass die zentrale Annahme fraglich erscheint, nach der Abweichungen der Leistungsbewertungen auf Grundlage von einerseits Noten und andererseits Ergebnissen in standardisierten Tests Benotungsbias darstellen, die auf nicht valide Benotung zurückzuführen sind. So erfasst der eingesetzte Deutschtest lediglich Leseverstehen und der „domänenspezifische“ Naturwissenschaftstest wurde sowohl für die Fächernoten in Biologie und Chemie als auch in Physik als Kontrollvariable genutzt. Neben dem Einwand, dass mit diesen Tests die fachbezogenen Leistungen wohl kaum umfänglich und valide abgebildet werden, lässt sich argumentieren, dass Zeugnisnoten nicht nur auf schriftlichen, sondern auch auf mündlichen bzw. sonstigen Leistungen im Unterricht beruhen und dass sich in der Benotung von fachlichen Leistungen legitimierweise auch unzulängliche sprachliche Leistungen niederschlagen können. Insofern sind die ermittelten Benotungsbias möglicherweise (teilweise) auch inhaltlich begründet und nicht ausschließlich verzerrt durch Persönlichkeitsmerkmale, Klassenraumverhalten und Stereotype in den Köpfen der Lehrkräfte.

Zu den Ergebnissen

Für Deutsch, Mathematik und Physik lassen sich Benotungsbias hinsichtlich aller untersuchten Merkmale (Geschlecht, sozioökonomischer Status, Übergewicht, Migrationshintergrund) nachweisen, in Chemie besteht nur für das Geschlecht kein Benotungsbias und in Biologie nur für Migrationshintergrund nicht. Die am stärksten ausgeprägten Benotungsbias finden sich in Deutsch zugunsten von Mädchen und zum Nachteil türkeistämmiger Jugendlicher. Schade ist, dass die Streuung der Noten nicht angegeben wird, was Voraussetzung wäre, um den in Standardeinheiten angegebenen Differenzen einen Realitätsbezug zu geben und das Ausmaß der Benotungsbias in Notenstufen bestimmen zu können.

Da Lernende gleichzeitig von Benotungsbias aufgrund verschiedener Gruppenzugehörigkeiten betroffen sein können, vergleichen die Forschenden die Gruppendifferenzen für Schülerinnen und Schüler mit spezifischen Kombinationen von Merkmalen. Sie stellen fest, dass die Effekte der einzelnen Merkmale kumulieren, d. h., dass beispielsweise bei übergewichtigen türkeistämmigen Schülern mit niedrigem sozioökonomischem Status (SES) starke negative Verzerrungseffekte, bei Mädchen ohne

Migrationshintergrund und Übergewicht und mit hohem SES positive Verzerrungseffekte bestehen, so dass der Unterschied in der Benotung zwischen diesen Gruppen in Deutsch trotz vergleichbarer Testleistungen recht hoch ausfällt, in Biologie ist er etwa halb so groß und in den anderen Fächern tendenziell geringer. Insgesamt deuten die Ergebnisse darauf hin, dass Benotungsbias weit verbreitet sind und dass Lernende von intersektionalen Ungleichheiten betroffen sein können.

Nennstiel und Gilgen weisen darauf hin, dass die Gründe für die Benotungsbias in zukünftigen Studien untersucht werden müssten und alternative Erklärungen, wie tatsächliche Leistungsunterschiede, nicht ausgeschlossen werden könnten. Dennoch halten sie ihre Forschungsergebnisse für geeignet, die Entwicklung von Maßnahmen zu fördern, mit denen gerechtere schulische Bedingungen geschaffen werden. Eine allgemeine Diskussion über die Rolle von Noten und was sie messen sollten, könnte das System der Notenvergabe genauer und gerechter machen. Konkret schlagen die Forschenden vor: Da es den größten Benotungsbias im Fach Deutsch gebe, wo viel Bewertungsspielraum existiere, könnte die Implementation von strukturierteren Benotungsschemata zu gerechteren Noten beitragen.

Im Hinblick auf schulische Praxis geben diese Ausführungen Impulse für Reflexionen der Bewertungspraxen sowohl auf individueller wie auch auf Schulebene. Gemeinsame Reflexionen könnten dazu führen, kooperativ Bewertungen zu diskutieren und durch gemeinsame Bewertungsschemata zu mehr Objektivität zu kommen.

Für die Beseitigung von Bildungsungleichheiten, die damit verbunden sind, dass Schülerinnen und Schüler ihre Lernpotenziale nicht voll ausschöpfen und Notenleistungen nicht entsprechend ihrer Kompetenzen aus standardisierten Tests erbringen, kommt der Objektivität der Benotung durch die Lehrkräfte allerdings vermutlich eher eine untergeordnete Rolle zu. Im Gegenteil könnten Maßnahmen zur objektiveren Benotung schließlich dazu führen, dass Notenunterschiede, beispielsweise zwischen Schülerinnen und Schülern mit und ohne Migrationshintergrund, größer werden und Bildungschancen sinken, da Lehrkräfte an deutschen Schulen im Mittel dazu tendieren, Schülerinnen und Schüler aus ethnischen Minderheiten und aus ungünstigen sozialen Verhältnissen vergleichsweise besser zu benoten, insbesondere die Leistungsstärkeren unter ihnen (Bredtmann, Otten & Vonnahme, 2024).

Reflexionsfragen für die Praxis

Nachfolgende Reflexionsfragen sind ein Angebot, die Befunde der rezensierten Studie auf das eigene Handeln als Lehrkraft oder Schulleitungsmitglied zu beziehen und zu überlegen, inwiefern sich Anregungen für die eigene Handlungspraxis ergeben. Die Befunde der rezensierten Studien sind nicht immer generalisierbar, was z. B. in einer begrenzten Stichprobe begründet ist. Aber auch in diesen Fällen können die Ergebnisse interessante Hinweise liefern, um über die eigene pädagogische und schulentwicklerische Praxis zu reflektieren.

Reflexionsfragen für Lehrkräfte

- Bei welchen meiner Schülerinnen und Schüler entsprechen die Notenleistungen nicht den Ergebnissen in standardisierten Tests (z. B. VERA)?
- Welche Zusammenhänge bestehen ggf. mit Geschlecht, sozioökonomischem Hintergrund, Migrationshintergrund und Übergewicht?
- Was kann ich in meinem Unterricht tun, um Lern- und Leistungspotenziale bei potenziell benachteiligten Schülerinnen und Schülern zu entdecken?

- Inwiefern nutze ich Möglichkeiten, um Leistungen der Lernenden möglichst objektiv und ohne Einfluss von bestimmten Merkmalen zu beurteilen?
- Inwieweit tausche ich mich mit anderen Lehrkräften über dieses Thema aus, damit unsere Bewertungen nicht durch Stereotype und Vorurteile beeinflusst werden?

Reflexionsfragen für Schulleitungen

- Welche Rolle spielt ein möglicher Benotungsbias in meiner Reflexion von Unterrichtspraxis?
- Inwiefern gibt es Austausch unter den Lehrkräften über die individuelle Bewertungspraxis?
- Wo kann ich ansetzen, um die Ressourcen der Kolleginnen und Kollegen zu Verbesserungen in diesem Bereich zu stärken?

Literatur

Bredtmann, J., Otten, S. & Vonnahme, C. (2024). *Discrimination in Grading? Evidence on Teachers' Evaluation Bias Towards Minority Students* (Ruhr Economic Papers, #1122). Essen: RWI – Leibniz-Institut für Wirtschaftsforschung.

Rezensent/-in

Dr. Johannes Rosendahl, Sonja Hensel Dr. Johannes Rosendahl ist Referent an der Qualitäts- und UnterstützungsAgentur - Landesinstitut für Schule (QUA-LiS NRW). Dr. Sonja Hensel ist Lehrerin am Berufskolleg in Siegburg sowie Lehrbeauftragte an der Universität Siegen. Arbeitsschwerpunkte: Rechtschreib-, Schreib- und Lesedidaktik, selbstreguliertes und kooperatives Lernen.

Zitiervorschlag

Johannes Rosendahl, Sonja Hensel (2025). Rezension zu Nennstiel, R. & Gilgen, S. (2024). Does chubby Can get lower grades than skinny Sophie? Using an intersectional approach to uncover grading bias in German secondary schools. PLoS ONE 19(7), 1–23. *Forschungsmonitor Schule, 194*. Abgerufen von https://www.forschungsmonitor-schule.de/print.php?id=187

Urheberrecht