„Paul ist nicht so gut in Deutsch“

Geschlechtsdifferenzielle Benotung im Fach Deutsch – eine Sekundäranalyse der Daten des IQB-Bildungstrends 2015

Christin Rüdiger[1/2], Dr. Malte Jansen[1/2], Dr. Camilla Rjosk[1]
[1] Institut zur Qualitätsentwicklung im Bildungswesen (IQB) an der Humboldt-Universität zu Berlin
[2] Zentrum für internationale Bildungsvergleichsstudien (ZIB)

Zusammenfassung: Jungen erhalten, insbesondere in den sprachlichen Fächern, im Durchschnitt schlechtere Noten als Mädchen. Eine Ursache dafür sind geringere Kompetenzen der Jungen, aber auch motivationale Faktoren, Lehrkraft- und Elternmerkmale werden als Erklärungsfaktoren diskutiert. Dieser Beitrag untersucht die Frage, ob Notenunterschiede im Fach Deutsch zwischen Jungen und Mädchen nach der Kontrolle von Kompetenzen bestehen bleiben und wie diese erklärt werden können. Dazu wurden die Daten des IQB-Bildungstrends 2015 ($N = 21.432$) genutzt, welche Leistungstests in drei für das Fach Deutsch maßgeblichen Kompetenzbereichen (Lesen, Zuhören und Orthografie) beinhalten. Es zeigte sich erwartungsgemäß, dass die schlechteren Deutschnoten von Jungen der neunten Jahrgangsstufe zum Teil auf ihre niedrigeren Kompetenzen zurückzuführen sind. Darüber hinaus trugen aber auch die geringere Ausprägung motivationaler Merkmale (Selbstkonzept, Interesse, Anstrengungsbereitschaft) und die höhere Ausprägung von Langeweile im Unterricht sowie Lehrkraftüberzeugungen zum Leseverhalten in unterschiedlichem Maße zu den Notenunterschieden bei. Auch nach Kontrolle dieser Faktoren blieb ein Benotungsnachteil für die Jungen erhalten. Mögliche Ursachen und Aufgaben zukünftiger Forschung werden diskutiert.

Schlüsselbegriffe: Benotung, Geschlechterdisparität, Lehrkraftüberzeugung, Deutschunterricht

„Paul is not so good in German“
Gender-differentiated grading in German – a secondary analysis of the IQB Trends in Student Achievement 2015 data

Summary: On average, boys receive lower grades than girls, especially in language subjects. One reason are their lower competencies. However, motivational factors and teacher perceptions also contribute to those grade disparities. This article examines whether gender differences in German grades can be explained by looking beyond the gender gap in competencies. We used a large-scale study of German ninth-grade students ($N = 21.432$), which includes tests in three areas of competence (reading, listening and spelling). As expected, it was shown that the lower grades for boys were partly due to their lower competencies and motivation (self-concept, interest, willingness to make an effort) as well as higher boredom. Controlling for competencies, motivational and cognitive characteristics as well as the teacher's perception of reading behavior, an unexplained grading disadvantage for boys was found. Possible causes and tasks of future research are discussed.

Keywords: Grading, reading, gender inequality, teacher beliefs, subject German

Da Mädchen in Deutschland in allen Jahrgangsstufen durchschnittlich bessere Noten erzielen als Jungen, fragt sich die empirische Bildungsforschung seit mehreren Jahren: *„Sind Jungen die neuen Bildungsverlierer?“* (Diefenbach, 2011; Hannover & Kessels, 2011; Helbig, 2010; Helbig 2012; Kuhl & Hannover, 2012). Benotung ist theoretisch von multiplen Bedingungsfak-

Psychologie in Erziehung und Unterricht, 2021, *68*, Open Access
DOI 10.2378/peu2021.art08d

toren abhängig (Hochweber, 2010), die somit auch bei der empirischen Betrachtung dieses Phänomens zur Erklärung der Notenunterschiede zwischen Mädchen und Jungen herangezogen werden können. Zum einen könnten Geschlechterunterschiede in den fachlichen Kompetenzen für Notenunterschiede verantwortlich sein, da Fachkompetenzen gute Prädiktoren für Fachnoten sind und Mädchen insbesondere im sprachlichen Bereich höhere fachliche Kompetenzen zeigen (Fuchs & Brunner, 2017; Hochweber, 2010; Lintorf, 2012). Da konsistente Benotungsunterschiede aber auch noch nach Kontrolle von Kompetenzen auftreten (Hochweber, 2010; Helbig, 2010; Lehmann, Peek & Gänsfuß, 1997; Lehmann et al., 2000; Zinn & Bayer, 2018), kommen fachspezifische motivationale Merkmale wie Selbstkonzept und Interesse sowie fächerübergreifende motivationale Merkmale wie Anstrengungsbereitschaft, Fleiß und Selbstregulationsfähigkeit als weitere erklärende Faktoren in Betracht (Duckworth & Seligman, 2006; Hochweber, 2010; Kuhl & Hannover, 2012; Kessels & Heyder, 2017; Spiel, Wagner & Fellner, 2002). Bei den fächerübergreifenden motivationalen Merkmalen – man könnte auch vom Arbeits- und Sozialverhalten sprechen – zeigen sich durchgehend Vorteile für Mädchen (Han, Elsäßer, Lang & Ditton, 2017; Spiel et al., 2002; Steinmayr & Spinath, 2008; van Ophuysen, 2008; für einen Überblick siehe Kessels & Heyder, 2018). Bei den fachbezogenen motivationalen Merkmalen existieren ebenfalls stereotype Geschlechterunterschiede. Mädchen haben durchschnittlich ein höheres Interesse und Selbstkonzept in den sprachlichen Fächern (Artelt, Naumann & Schneider, 2010; Böhme, Sebald, Weirich & Stanat, 2016; Stanat & Kunter, 2001), Jungen häufiger in den mathematisch-naturwissenschaftlichen Fächern (Jansen, Schneider, Schipolowski & Henschel, 2019). Neben dem Einfluss von Individualmerkmalen der Schülerinnen und Schüler können auch Merkmale der Lehrkräfte, deren Einschätzungen ja letztlich die Benotung bedingen (Hochweber 2010), relevant für die Erklärung von Geschlechterunterschieden werden. Dies könnte zum Beispiel der Fall sein, wenn Lehrkräfte unterschiedliche – potenziell stereotype – Überzeugungen über das schulische Verhalten von Jungen und Mädchen hätten, wofür es bereits erste Evidenz gibt (Holder & Kessels, 2017; Lorenz, Gentrup, Kristen, Stanat & Kogan, 2016; Muntoni & Retelsdorf, 2018). Solche differenziellen Überzeugungen könnten sich zusätzlich zur Beurteilung der Kompetenzen sowie des Arbeits- und Sozialverhaltens der Schülerinnen und Schüler auf die Benotung auswirken. Vor diesem Hintergrund sollen in der vorliegenden Studie Zusammenhänge von diversen Schülermerkmalen und zusätzlich auch mehreren Lehrkräftemerkmalen mit Zeugnisnoten im Fach Deutsch betrachtet werden. Dem Beitrag liegen die Daten des IQB-Bildungstrends 2015, an dem Schülerinnen und Schüler der neunten Jahrgangsstufe teilnahmen, zugrunde. Erstmals werden Notenunterschiede unter Berücksichtigung von Kompetenzausprägungen in drei Bereichen (Lesen, Zuhören, Orthografie) mit an den Bildungsstandards der Kultusministerkonferenz orientierten Kompetenztests untersucht. Es wird der Frage nachgegangen, ob sich bereits bekannte Geschlechterdisparitäten im Fach Deutsch zugunsten von Mädchen erneut zeigen (Replikation von Han et al., 2017; Helbig, 2010; Kuhl & Hannover, 2012; Lehmann et al., 1997; Zinn & Bayer, 2018) und wie sich diese durch die Berücksichtigung von schulischen Kompetenzen sowie kognitiven, motivationalen und soziodemografischen Merkmalen der Schülerinnen und Schüler verändern. Zusätzlich wird die Bedeutung von Lehrkraftmerkmalen für die Benotung untersucht. Dabei liegt ein Fokus auf der möglichen Rolle von geschlechtsspezifischen Lehrkraftüberzeugungen zum Leseverhalten. Die Studie stellt damit bisherige Erkenntnisse und Annahmen zu Notendisparitäten auf eine aktuelle Datenbasis und erweitert den Forschungsstand zudem durch die Berücksichtigung von mehreren zentralen Schülermerkmalen und Lehrkraftfaktoren in einem gemeinsamen Modell.

Benotungsunterschiede zwischen Mädchen und Jungen

Ergebnisse aus einer Metaanalyse (Voyer & Voyer, 2014) mit 369 Primärstudien – allerdings keine aus Deutschland – fanden fächer-, alters- und länderübergreifende Evidenz für einen generellen Notenvorsprung von Mädchen bzw. Frauen. Die Effektstärke des Geschlechtsunterschiedes über alle Fächer und Altersgruppen betrug $d = 0.23$. Der größte Notenunterschied zwischen den Geschlechtern zeigte sich im sprachlichen Bereich (Primarstufe: $d = 0.20$, Sekundarstufe I: $d = 0.45$, Sekundarstufe II: $d = 0.47$), welcher auch in der vorliegenden Studie untersucht wird. Ein Benotungsvorteil für Mädchen konnte auch für Deutschland in mehreren Einzelstudien für verschiedene Fächer und Klassenstufen nachgewiesen werden. Eine mit Voyer und Voyer (2014) vergleichbare Metaanalyse gibt es bislang nicht. Im Folgenden wird der Forschungsstand zu Notenunterschieden zwischen Mädchen und Jungen für unterschiedliche Fächer berichtet, obwohl die vorliegende Studie nur das Fach Deutsch in den Blick nimmt. Das Problem der geschlechtsdifferenziellen Benotung kann so zunächst überblicksartig dargestellt werden. Die vergleichende Betrachtung erscheint uns zudem aufgrund der Abgrenzung potenziell unterschiedlicher geschlechtsspezifischer (stereotyper) Wahrnehmungen der Domänen Mathematik bzw. Deutsch einerseits und möglicher fächerübergreifender Faktoren andererseits gewinnbringend. So werden bereits in der Primarstufe stereotype Unterschiede zwischen der Beurteilung in den Fächern Mathematik und Deutsch erkennbar: In der 4. Jahrgangsstufe sind die geschlechtsspezifischen Benotungsunterschiede im Fach Mathematik (zugunsten der Jungen) und im Fach Sachkunde (zugunsten der Mädchen) vorhanden, aber eher gering (Lintorf, 2012; Wendt, Steinmayr & Kasper, 2016). Im Fach Deutsch (zugunsten der Mädchen) sind die Unterschiede hingegen bereits stärker ausgeprägt (Kuhl & Hannover, 2012). In der 6. Jahrgangsstufe zeigen sich weiterhin kleine Unterschiede in Mathematik (zugunsten der Jungen), größere wiederum im Fach Deutsch (zugunsten der Mädchen) (Helbig, 2010). Wie von Voyer und Voyer (2014) beschrieben, scheint sich dann auch in Deutschland die Notendifferenz zwischen Jungen und Mädchen, insbesondere in den sprachlichen Fächern, während der Schullaufbahn zu verstärken. Dresel, Stöger und Ziegler (2006) zeigten dies – mit schulartspezifisch differenziellen Befunden – für die Jahrgangsstufen 5 bis 10.

Da Variation in Schulnoten zunächst Variation in den zugrunde liegenden Kompetenzen abbildet, sind Studien, die Kompetenzen kontrollieren, aussagekräftiger als der bloße Vergleich der Noten von Mädchen und Jungen. Passend zum Befundmuster der Notenunterschiede fallen geschlechtsspezifische Kompetenzunterschiede in den sprachlichen Fächern zumeist zugunsten der Mädchen aus, während in Mathematik häufig keine Kompetenzunterschiede oder kleinere Vorteile für Jungen auftreten (Deutsch: Böhme et al., 2016; Lehmann et al., 1997; Mathematik: Reiss, Weis, Klieme & Köller, 2019; Schipolowski, Wittig, Mahler & Stanat, 2019). Unter Kontrolle dieser Unterschiede durch das Berücksichtigen von Kompetenztests blieben in der Sekundarstufe Benotungsvorteile für Mädchen im Fach Deutsch bestehen (Zinn & Bayer, 2018) und zeigten sich in schwächerer Form auch im Fach Mathematik (Hochweber, 2010; Lehmann et al., 2000). Ein ähnliches Bild zeigte sich für beide Fächer in der 6. (Helbig, 2010) sowie in der 4. Jahrgangsstufe im Fach Deutsch (Hannover & Kessels, 2011; Kuhl & Hannover, 2010), allerdings nicht im Fach Mathematik (Lintorf, 2012; Wendt et al., 2016). Die dargestellten Befunde zeigen für die Hauptfächer der allgemeinbildenden Schularten, aber insbesondere im Fach Deutsch, einen Vorteil der Mädchen gegenüber den Jungen bei gleichen fachlichen Kompetenzen. Allerdings basieren diese Studien auf teilweise über 20 Jahre alten Daten und sind hauptsächlich auf einzelne Bundesländer und selektierte Stichproben (aufgrund freiwilliger Teilnahme) beschränkt. Darüber hinaus variieren Breite und Umfang der genutzten Kompetenztests sowie der Kontrollvariablen. Weil die Notenunterschiede auch nach Kontrolle der fachlichen Kompetenzen bestanden, wur-

den in den referierten Studien weitere für Notendisparitäten relevante Faktoren untersucht (Han et al., 2017; Helbig, 2010; Hochweber, 2010; Kuhl & Hannover, 2012; Lehmann et al.,1997; Lehmann et al., 2000; Zinn & Bayer, 2018). Dies sind mehrheitlich motivationale Merkmale der Schülerinnen und Schüler, ihr Arbeits- und Sozialverhalten sowie Lehrkraftfaktoren. Einige dieser Studien nutzten zudem auch noch Merkmale der Eltern oder Elterneinschätzungen der Kinder bzw. Jugendlichen, die aber in unserer Studie nicht untersucht und daher im Folgenden nicht detaillierter beschrieben werden.

Gründe für Benotungsunterschiede zwischen Mädchen und Jungen bei vergleichbaren Kompetenzen

In verschiedenen theoretischen Modellen (Helbig, 2012; Hochweber, 2010; Tent, 2001) wird angenommen, dass über die Schulleistung hinaus weitere Merkmale der Schülerinnen und Schüler zu Benotungsunterschieden beitragen. Dazu zählen etwa motivationale (z. B. Fachinteresse), soziodemografische (z. B. sozioökonomischer Status) und familiale Faktoren (z. B. elterliche Unterstützung). Des Weiteren werden strukturelle (z. B. Schulart) und kulturell geteilte Merkmale (z. B. gruppenbezogene Stereotype) sowie Lehrkraftfaktoren (z. B. Geschlecht, Überzeugungen) als Faktoren diskutiert. Zu motivationalen Schülermerkmalen liegen bereits einige Studien vor: Han et al. (2017) konnten für Grundschulkinder zeigen, dass Unterschiede im Arbeitsverhalten (bei Jungen weniger positiv ausgeprägt als bei Mädchen) zusätzlich zu den Kompetenzen Geschlechterdisparitäten in den Fachnoten erklären. Lintorf (2012) konnte, ebenfalls für Schülerinnen und Schüler an Grundschulen, einen Zusammenhang von Gewissenhaftigkeit (bei Mädchen stärker ausgeprägt) und Benotung unter Kontrolle der Kompetenzen in den Fächern Mathematik und Sachkunde feststellen. Kuhl und Hannover (2012) untersuchten den Einfluss der Lehrkrafteinschätzung des selbstgesteuerten Lernens (bei Mädchen stärker ausgeprägt) auf die Benotung in der Grundschule. Sie konnten zeigen, dass diese Einschätzung hochgradig notenrelevant und ein stärkerer Prädiktor für die Deutschnote war als die Lesekompetenz. In Bezug auf die Rolle fachlicher Motivation fand Hochweber (2010) für Sekundarstufenschülerinnen und -schüler, dass das Fachinteresse (bei Jungen stärker ausgeprägt) und die Anstrengung im Fach Mathematik deutliche positive Zusammenhänge mit der Fachnote zeigten. Lauermann, Meißner und Steinmayr (2020) zeigten für die gleiche Altersgruppe starke Zusammenhänge von Note und Selbstkonzept in den Fächern Mathematik (bei Jungen stärker ausgeprägt) und Deutsch (bei Jungen geringer ausgeprägt) auf. Arens (2019) untersuchte bei Grundschulkindern Zusammenhänge von Noten und dem intrinsischen Wert sowie der Wichtigkeit des schulischen Lernens (im Fach Deutsch höher für Mädchen und in den Fächern Mathematik und Sachkunde höher für Jungen). Sie fand signifikante positive Zusammenhänge mit den Fachnoten. Als Gründe für diese ungleiche geschlechts- und fachspezifische Verteilung von Selbstkonzept, Motivation, Interesse sowie Anstrengung, Gewissenhaftigkeit und Selbststeuerung werden verschiedene Gründe diskutiert: Für die fachspezifischen Unterschiede können laut Helbig (2012) unter anderem stereotype Kompetenzzuschreibungen durch Sozialisationsagenten verantwortlich sein. Für die Unterschiede im Arbeits- und Sozialverhalten wird als ein erklärender Aspekt eine für Jungen hinderliche – weil identitätsinkongruente – Wahrnehmung von Schule und Lernen als feminin diskutiert (Heyder & Kessels, 2013; Kessels & Heyder, 2017; Heyder, van Hek & van Houtte, 2020).

Neben Merkmalen der Schülerinnen und Schüler könnten aber auch Merkmale und Überzeugungen von Lehrkräften Geschlechterunterschiede in der Benotung miterklären. Dass Notendisparitäten mit geschlechterstereotypen Überzeugungen aufseiten der Lehrkräfte zusammenhängen könnten, lässt sich ebenfalls aus Überlegungen von Helbig (2012) sowie Heyder und Kessels (2013) und Kessels und Heyder (2017) ableiten. Auch empirische Befunde sprechen für die Annahme, dass Lehrkräfte teilweise

Schülerfähigkeiten verzerrt wahrnehmen (Gentrup, Rjosk, Stanat & Lorenz, 2018; Lorenz et al., 2016; Muntoni & Retelsdorf, 2018). Stereotype Überzeugungen von Lehrkräften können sich dabei zum Beispiel auf eine größere Sprachbegabung von Mädchen im Gegensatz zu einer stärker ausgeprägten mathematisch-naturwissenschaftlichen Begabung von Jungen beziehen (Hannover, Wolter & Zander, 2017). Oder auf die Generalisierung, Mädchen seien grundsätzlich in der Schule anstrengungs- und leistungsbereiter (Baudson & Preckel, 2013; Jones & Myhill, 2004).

Das Geschlecht der Lehrkraft wurde in den letzten Jahren insbesondere in Bezug auf die Primarstufe als ein weiterer Faktor der Notengebung diskutiert. Bislang konnte keine konsistente Evidenz für einen Haupt- oder Interaktionseffekt des Lehrkraftgeschlechts auf die Benotung in der Grundschule gefunden werden (Helbig, 2010; Neugebauer, Helbig & Landmann, 2010). Für die Sekundarstufe liegen dazu keine Studien vor.

Die Berufserfahrung von Lehrkräften könnte bei der Beurteilung von Schülerleistungen ebenfalls eine Rolle spielen, da es Hinweise darauf gibt, dass erfahrenere Lehrkräfte bei der Schülerbeurteilung weniger stark urteilsverzerrenden Einflüssen unterliegen als Novizen (Krolak-Schwerdt, Böhmer & Gräsel, 2009; van Ophuysen, 2006).

Fragestellung

Zusammengefasst liefern die referierten Studien zu Benotungsunterschieden zwischen Mädchen und Jungen in Deutschland erste Hinweise zur Genese dieser Unterschiede (z. B. Han et al., 2017; Helbig, 2010; Kuhl & Hannover, 2012; Hochweber, 2010; Lehmann et al., 1997; Lehmann et al., 2000; Lintorf, 2012; Zinn & Bayer, 2018). Es gibt bislang jedoch keine Untersuchung für Notendisparitäten in der 9. Jahrgangsstufe im Fach Deutsch, die Kompetenztests, motivationale, kognitive und soziodemografische Schülermerkmale sowie Lehrkraftmerkmale und Lehrkraftüberzeugungen berücksichtigt. Diese Forschungslücke möchte der vorliegende Beitrag schließen.

Es wird untersucht, ob auf Basis des IQB-Bildungstrends 2015 (I) die bekannten Geschlechtsunterschiede in den Zeugnisnoten im Fach Deutsch repliziert werden können und ob diese Benotungsunterschiede (II) unter Kontrolle von curricular validen Kompetenztests in den drei Bereichen Lesen, Zuhören und Orthografie bestehen bleiben. Weiterhin wird untersucht, ob Notendifferenzen (III) auch unter zusätzlicher Kontrolle von motivationalen Schülermerkmalen (Anstrengungsbereitschaft, Selbstkonzept, Interesse, Leseverhalten und Langeweile) sowie der kognitiven Fähigkeiten, des sozioökonomischen Status und der zu Hause gesprochenen Sprache bestehen bleiben. Tritt in der Stichprobe der erwartete Geschlechtseffekt auf und bleibt er nach Kontrolle der Schülermerkmale bestehen, wird untersucht, ob (IV) Lehrkraftmerkmale (Geschlecht, Berufserfahrung) sowie geschlechtsspezifische Lehrkraftüberzeugungen und die Schulart mit der Benotung von Mädchen und Jungen zusammenhängen.

Methodisches Vorgehen

Stichprobe

Die verwendeten Daten des IQB-Bildungstrends 2015 sind am Forschungsdatenzentrum des Instituts zur Qualitätsentwicklung im Bildungswesen (FDZ am IQB) verfügbar (Stanat et al., 2018). Die Studie wurde im Frühjahr 2015 u. a. im Fach Deutsch in der neunten Jahrgangsstufe zur Überprüfung des Erreichens der in den Bildungsstandards der Kultusministerkonferenz beschriebenen Kompetenzen in allen deutschen Bundesländern durchgeführt. Es wurden die Kompetenzen der Jugendlichen sowie zahlreiche Hintergrundmerkmale der Schülerinnen und Schüler und ihrer Lehrkräfte erfasst (Skalenhandbuch: Schipolowski, Haag, Milles, Pietz & Stanat, 2018). Die Stichprobe des IQB-Bildungstrends 2015, aus welcher die Analysestichprobe gebildet wurde, umfasst $N = 36.542$ Schülerinnen und Schüler aus $N = 1513$ Schulen und deren $N = 1575$ Deutschlehrkräfte. Auf Schülerebene wurde die Stichprobe repräsentativ gezogen. In jeder Schule wurde nur eine Klasse gezogen (Details zur Lehrkräftestichprobe: Hoffmann & Richter, 2016; zur Gesamtstichprobe: Schipolowski, Haag, Böhme & Sachse, 2016).

Für die vorliegenden Analysen wurden einige Schülerinnen und Schüler ausgeschlossen und zwar (a) Jugendliche mit sonderpädagogischem Förderbedarf sowie zieldifferent unterrichtete Jugendliche ($N = 4881$, davon $N = 1560$ an Förderschulen), (b) Jugendliche, die aufgrund diverser Ursachen nicht an der verpflichtenden Testung teilnehmen konnten ($N = 1029$) sowie (c) Schülerinnen und Schüler, die keine (Ziffern)Noten erhalten hatten oder deren Noten nicht nachvollziehbar waren ($N = 2317$). Zudem wurden nur Schülerinnen und Schüler berücksichtigt, deren Lehrkräfte an der Befragung teilgenommen hatten (die Teilnahmepflicht für Lehrkräfte variierte bundeslandspezifisch). Der Umfang der Analysestichprobe reduzierte sich daher auf $N = 21.432$ Jugendliche in $N = 980$ Klassen mit einer durchschnittlichen Anzahl von $N = 21.87$ Schülerinnen und Schülern. Die Anzahl der dazugehörigen Deutschlehrkräfte betrug ebenfalls $N = 980$ Personen. Zu Beginn des Testzeitraums waren die Jugendlichen (50,9 % weiblich) im Mittel 15.44 Jahre ($SD = 0.58$) alt. Die Schülerinnen und Schüler besuchten Hauptschulen und Realschulen (5,8 % und 14,1 %), Gesamtschulen sowie Schulen mit mehreren Bildungsgängen (11 % und 17,8 %) oder Gymnasien (51,3 %).

Messinstrumente

Note im Fach Deutsch

Die Deutschnoten aus den Halbjahreszeugnissen der neunten Jahrgangsstufe im Schuljahr 2014/2015 wurden durch eine designierte Lehrkraft (Schulkoordination für die Durchführung der Erhebung) für jede Schülerin und jeden Schüler auf Basis der Schulakten angegeben.

Kompetenzen im Fach Deutsch

Die Testaufgaben für das Fach Deutsch in den Kompetenzbereichen Lesen, Zuhören und Orthografie wurden am IQB durch erfahrene Lehrkräfte in Zusammenarbeit mit fachdidaktischen Kooperationspartnern erstellt (für Details siehe Becker-Mrotzek et al., 2016). Die Kompetenzschätzungen werden auf einer metrischen Skala mit einem Mittelwert von 500 Punkten und einer Standardabweichung von 100 Punkten angegeben (Sachse, Haag & Weirich, 2016). Für die nachfolgenden Analysen wurden 15 Plausible Values für die Kompetenzschätzungen verwendet (Lüdtke & Robitzsch, 2017).

Motivationale Merkmale der Schülerinnen und Schüler

Aus den im Schülerfragebogen erfassten Konstrukten wurden Anstrengungsbereitschaft, Deutschinteresse, Selbstkonzept im Fach Deutsch und Langeweile im Deutschunterricht sowie das Leseverhalten als Einflussfaktoren für Notendisparitäten über die Kompetenzen hinaus ausgewählt. Die *Anstrengungsbereitschaft* wurde mit drei Items *(„Ich bin fleißig", „Ich arbeite hart", „Was ich anfange, das beende ich auch")* einer am IQB entwickelten fünfstufigen Skala (1 = *„Trifft überhaupt nicht zu"* bis 5 = *„Trifft voll und ganz zu"*) in Anlehnung an Litman (2008) erfasst ($\alpha = 0.73$).

Das *Interesse am Fach Deutsch* wurde mit vier Items auf einer vierstufigen Skala („1 = *„Trifft gar nicht zu"* bis 4 = *„Trifft völlig zu"*) aus PISA 2003 (Ramm et al., 2006) erfasst (z. B. *„Das Fach Deutsch ist mir persönlich wichtig.";* $\alpha = 0.87$).

Das *Selbstkonzept im Fach Deutsch* wurde mit sieben Items einer vierstufigen Skala (1 = *„Stimmt ganz genau"* bis 4 = *„Stimmt überhaupt nicht"*) aus der DESI-Studie (Wagner, Helmke & Rösner, 2009) erfasst (z. B. *„Für das Fach Deutsch habe ich einfach keine Begabung.";* $\alpha = 0.88$). Die positiv gepolten Items der Skala wurden rekodiert, sodass höhere Werte einem höheren Selbstkonzept entsprechen.

Die *Langeweile im Deutschunterricht* wurde mit drei Items einer am IQB entwickelten vierstufigen Skala (1 = *„Stimmt überhaupt nicht"* bis 4 = *„Stimmt ganz genau"*) in Anlehnung an Preckel, Götz und Frenzel (2010) erfasst (z. B. *„Ich finde den Deutschunterricht langweilig.";* $\alpha = 0.87$).

Zur Erfassung des *Leseverhaltens der Schülerinnen und Schüler* wurde eine Skala aus drei Items einer am IQB in Anlehnung an Gattenmeier (2004) entwickelten Skala gebildet (1 = *„Trifft gar nicht zu"* bis 4 = *„Trifft genau zu"*), wobei sich die Operationalisierung spezifisch auf die Vermeidung schwieriger Texte bezieht (*„Texte mit langen Sätzen lese ich nicht gern.", „Viele Fremdwörter und Fachausdrücke in einem Text stören mich.", „Umfangreiche Bücher schrecken mich ab.";* $\alpha = 0.71$). Höhere Werte bilden dabei eine höhere Vermeidungstendenz ab. Die Auswahl der Items für diese Skala begründet sich durch das Vorliegen gleichlautender Lehrkräfteitems, die zur Operationalisierung geschlechtsspezifischer Überzeugungen genutzt wurden (siehe nachfolgender Abschnitt *Lehrkraftüberzeugungen*). Zusätzlich wurde mit einer explorativen Hauptachsenanalyse der Gesamtskala mit Oblimin-Rotation mit dem Programm SPSS 25 sichergestellt, dass die Items auf einem gemeinsamen

Faktor laden (standardisierte Faktorladungen ≥ .65) und nur geringe Nebenladungen (zwischen - .06 und .09) vorliegen. (Die vollständige Skala *Lesemotivation und -präferenz,* die im IQB-Bildungstrend 2015 eingesetzt wurde, kann dem Skalenhandbuch entnommen werden; Schipolowski et al., 2018.)

Kontrollvariablen auf Ebene der Schülerinnen und Schüler

Als Kontrollvariablen dienten die kognitiven Grundfähigkeiten, der sozioökonomische Status der Familie (HISEI; Ganzeboom, 2010) sowie die Schülerangabe der zu Hause gesprochenen Sprache (Referenzkategorie: *„immer Deutsch"* versus *„meistens, manchmal oder nie Deutsch"*). Das eingesetzte Instrument zur Erfassung der kognitiven Grundfähigkeiten ist ein Untertest zum Schlussfolgernden Denken (figurale Aspekte) des Berliner Tests zur Erfassung fluider und kristalliner Intelligenz für die 8. bis 10. Jahrgangsstufe (BEFKI; Wilhelm, Schroeders & Schipolowski, 2014).

Lehrkraftüberzeugung

Die Operationalisierung der geschlechtsspezifischen Überzeugung der Lehrkräfte geschah in zwei Schritten. Die Lehrkräfte wurden mit getrennten Skalen zum Leseverhalten der Mädchen und Jungen ihrer Klasse befragt. Die Items erfassten, inwiefern die Lehrkraft den Eindruck hat, dass Jungen bzw. Mädchen schwierigere Lesetexte vermeiden. Diese evaluative Einschätzung durch die Lehrkräfte impliziert eine Wertung bezüglich des Leseverhaltens. Hierbei würde eine Vermeidungseinschätzung auf eine wahrgenommene niedrigere Lesekompetenz oder ein niedrigeres Leseinteresse – eine für Jungen stereotype Einschätzung – hinweisen. Für diese Skalen wurden die gleichen drei Items (in Anlehnung an Gattenmaier, 2004) mit vierstufiger Skala wie im Schülerfragebogen verwendet *(„Wie schätzen Sie die folgenden Aussagen für die [Jungen/Mädchen] in Ihrer Klasse ein?" – „Umfangreiche Bücher schrecken sie ab.", „Sie lesen nicht gerne Texte mit langen Sätzen.", „Viele Fremdwörter und Fachausdrücke in einem Text stören sie.").* Eine Hauptachsenanalyse der Gesamtskala mit Oblimin-Rotation ergab, dass die drei ausgewählten Items auf einem gemeinsamen Faktor laden (Mädchen: standardisierte Faktorladungen ≥ .75, Jungen: standardisierte Faktorladungen ≥ .70) und nur geringe Nebenladungen (Mädchen: zwischen - .03 und .00, Jungen: zwischen - .03 und .02) vorliegen. Die Reliabilitäten der Lehrkräfte-Skalen waren gut (Mädchen $\alpha = 0.81$, Jungen $\alpha = 0.84$). Diese separate Befragung zum Leseverhalten von Mädchen und Jungen wurde genutzt, um geschlechtsspezifische Lehrkraftüberzeugungen zu operationalisieren. Durch die Geschlechterdifferenzen betonende Art der Fragestellung (nach der Einschätzung aller Mädchen bzw. aller Jungen der Klasse) werden, so die Prämisse, ggf. vorhandene geschlechtsspezifische Überzeugungen im Vergleich der Skalenwerte für Mädchen und Jungen quantifizierbar. Zur Sichtbarmachung differenzieller Lehrkraftüberzeugungen zum Leseverhalten von Mädchen und Jungen wurden im zweiten Schritt die Mittelwerte der beiden Lehrkraftskalen zum Leseverhalten der Schülerinnen und Schüler voneinander abgezogen ($M_{Jungen} - M_{Mädchen}$), um für jede Lehrkraft einen Differenzwert zu bilden. Ein Differenzwert von 0 gibt an, dass die Lehrkraft keinen Unterschied zwischen dem Leseverhalten der Jungen und der Mädchen wahrnimmt, also keine geschlechtsspezifischen Überzeugungen erkennen lässt. Ein Differenzwert mit negativem Vorzeichen gibt an, dass die Lehrkraft die Lesevermeidung der Jungen geringer einschätzt als die der Mädchen. Ein Differenzwert mit positivem Vorzeichen gibt an, dass die Lehrkraft die Vermeidung der Mädchen geringer einschätzt als die der Jungen.

Kontrollvariablen auf Ebene der Lehrkräfte

Als Kontrollvariablen auf Ebene der Lehrkräfte wurden weiterhin das Geschlecht und die Berufserfahrung in Jahren einbezogen. Zusätzlich wurde die Schulart (*„Gymnasium"* vs. *„nicht-gymnasiale Schulformen"*) dummy-kodiert einbezogen.

Analysestrategie

Zunächst wurden Mittelwertvergleiche berechnet, um deskriptiv zu prüfen, ob sich eine Noten-, Leistungs- und Motivationsdifferenz zwischen Mädchen und Jungen im Fach Deutsch zeigt (Tabelle 1). Aufgrund der geclusterten Datenstruktur wurde für die weiteren Analysen ein Modell mit zwei Ebenen (Schülerebene und Lehrkräfteebene) angenommen. Es wurden drei Random-Intercept-Mehrebenenmodelle spezifiziert (siehe Tabelle 2). Zur Beantwortung der ersten Forschungsfrage wurde nur das Geschlecht als Prädiktor der Deutschnote untersucht (Modell 1). Danach wurde analysiert, ob die angenommene Notendifferenz zwischen Jungen und Mädchen auch unter Kontrolle der Testleistungen im Fach Deutsch (Modell 2, Forschungsfrage II) sowie unter Kontrolle

der Testleistungen, der motivationalen und soziodemografischen Merkmale sowie der kognitiven Fähigkeiten der Schülerinnen und Schüler (Modell 3, Forschungsfrage III) bestehen bleibt. Abschließend wurde der Zusammenhang von Lehrkraftmerkmalen sowie der Lehrkraftüberzeugung mit der Deutschnote untersucht. Dazu wurde ein Random-Slope-Modell mit Crosslevel-Interaktion (Modell 4, Forschungsfrage IV) spezifiziert. Es wird ein Moderatoreffekt angenommen, da ein Haupteffekt der geschlechtsspezifischen Lehrkraftüberzeugung unabhängig vom Geschlecht der Jugendlichen unplausibel erscheint. Daher sollte eine signifikante Crosslevel-Interaktion zwischen geschlechtsstereotypen Lehrkraftüberzeugungen und dem Schülergeschlecht sichtbar werden. Das Vorliegen von geschlechtsstereotypen Überzeugungen zum Leseverhalten (Jungen vermeiden eher schwierige Texte als Mädchen) sollte, wenn diese tatsächlich Benotung bedingen, unter Kontrolle von Kompetenzen, mit einem Notenvorteil für Mädchen und einem Notennachteil für Jungen einhergehen.

Die Mittelwertvergleiche und das Effektstärkemaß Cohens *d* (Cohen, 1988) wurden unter der Verwendung von Populationsgewichten in R (RCore-Team, 2017) mit dem Paket eatRep (Weirich & Hecht, 2018) berechnet. Für die Ermittlung der Standardfehler wurde die geclusterte Datenstruktur nach dem Jackknife-2-Verfahren (Wolter, 2007) berücksichtigt (für Details siehe Sachse et al., 2016). Die Modellschätzungen erfolgten mit der Software Mplus Version 7.11 (Muthén & Muthén, 1998–2017) unter Verwendung der Populationsgewichte auf Schülerebene. Die akzeptierte α-Fehlerwahrscheinlichkeit wurde aufgrund der Stichprobengröße auf 1 % festgelegt. Alle kontinuierlichen Prädiktorvariablen gingen *z*-standardisiert (grand-mean-zentriert) in die Analysen ein. Für die Schülerdaten wurden für alle unabhängigen Variablen, die fehlende Werte aufwiesen, 15 Imputationen mit dem R-Paket mice (van Buuren & Groothuis-Oudshoorn, 2011) vorgenommen. Fälle mit fehlenden Werten auf der abhängigen Variable wurden, wie in der Stichprobenbeschreibung erläutert, aus der Analyse ausgeschlossen. Für die fehlenden Werte im Lehrkräftedatensatz wurde das in Mplus implementierte Full-Information-Maximum-Likelihood-Verfahren (FIML, Enders, 2010) angewendet, da der Anteil fehlender Werte in den Analysevariablen gering war (eingeschätztes Leseverhalten Mädchen 4,1 %, eingeschätztes Leseverhalten Jungen 5,5 %, geschlechtsspezifische Lehrkraftüberzeugung 6,2 %, Geschlecht der Lehrkraft 3 %, Berufsjahre 2 %).

Ergebnisse

Deskriptive Ergebnisse

Die Mittelwerte und Standardabweichungen der Noten, der Kompetenztests, des Tests der kognitiven Grundfähigkeiten und der motivationalen Merkmale der Jugendlichen sowie der Lehrkraftüberzeugung zum Leseverhalten sind getrennt für Jungen und Mädchen in Tabelle 1 zu finden.

Tab. 1: Mittelwerte, Standardabweichungen und Intraklassenkorrelationen der Schülermerkmale und der Lehrkraftüberzeugung zum Leseverhalten (Vermeidung) getrennt für Mädchen und Jungen.

	$M_{Mädchen}$ (*SD*)	M_{Jungen} (*SD*)	Cohens *d* ($M_{Jungen} - M_{Mädchen}$)	ICC
Deutschnote	2.74 (0.83)	3.21 (0.83)	-0.56*	0.21
Lesekompetenz	528 (90)	502 (93)	0.28*	0.46
Zuhörkompetenz	528 (92)	504 (94)	0.25*	0.51
Orthografiekompetenz	540 (88)	496 (91)	0.49*	0.48
Anstrengung	3.49 (0.83)	3.40 (0.85)	0.11*	0.03
Selbstkonzept Deutsch	3.36 (0.54)	3.18 (0.60)	0.33*	0.06
Interesse Deutsch	2.55 (0.68)	2.29 (0.69)	0.38*	0.08
Langeweile Deutsch	2.16 (0.80)	2.36 (0.86)	-0.24*	0.13
Leseverhalten Selbsteinschätzung	2.15 (0.78)	2.10 (0.81)	0.06	0.05
Leseverhalten Lehrkraftüberzeugung	2.95 (0.73)	3.25 (0.67)	-0.43*	–

Anmerkungen: Die Lehrkräfteeinschätzung des Leseverhaltens wurde für Jungen und Mädchen getrennt erfragt. Die Effektstärke wurde auf Basis dieser beiden Einschätzungen (derselben Lehrkräfte) berechnet. Höhere Werte auf der Variable Leseverhalten kennzeichnen eine stärkere Vermeidung von schwierigen Lesetexten. $N_{Schülerinnen\ und\ Schüler} = 21.432$, $N_{Lehrkräfte} = 945$.
*$p < 0.01$

Mädchen erhielten, wie aufgrund der Literatur zu erwarten war, eine um etwa eine halbe Note bessere Bewertung im Fach Deutsch als Jungen ($M_{Mädchen} = 2.74$, $M_{Jungen} = 3.21$, $d = 0.56$, $p < .01$, siehe Tab. 1). Dazu passend erzielten die Mädchen auch in den Kompetenzbereichen Lesen ($d = 0.28$), Zuhören ($d = 0.25$) und insbesondere in der Orthografie ($d = 0.49$) signifikant höhere Leistungen. Im Vergleich der motivationalen Merkmale im Fach Deutsch zeigten sich die Mädchen signifikant interessierter ($d = 0.39$) und weniger gelangweilt ($d = 0.24$). Ihr Selbstkonzept in diesem Fach war erwartungsgemäß signifikant höher als das der Jungen ($d = 0.33$). Im Vergleich dazu war der Unterschied zwischen den Schülerinnen und Schülern in der selbstberichteten Anstrengungsbereitschaft zwar signifikant ($d = 0.11$) aber weniger ausgeprägt. Beim selbstberichteten Leserverhalten zeigte sich bei den Schülerinnen und Schülern kein signifikanter Unterschied ($d = 0.06$). Die Lehrkräfte schätzten hingegen das Leseverhalten ihrer Schülerinnen und Schüler im Mittel als stärker vermeidend ein als die Jugendlichen selbst. Weiterhin sahen die Lehrkräfte einen signifikanten Unterschied bei der Vermeidung schwieriger Texte zwischen den Jungen und den Mädchen ihrer Klasse ($d = -0.43$).

Tab. 2: Prädiktion der Deutschnote

	Null-modell	Modell 1		Modell 2		Modell 3		Modell 4	
		b	*SE*	*b*	*SE*	*b*	*SE*	*b*	*SE*
Individualmerkmale									
Geschlecht (1 = weiblich)		0.46*	0.01	0.29*	0.02	0.23*	0.01	0.23*	0.01
Lesen				0.15*	0.02	0.12*	0.01	0.12*	0.01
Zuhören				0.12*	0.02	0.08*	0.01	0.08*	0.01
Orthografie				0.29*	0.01	0.22*	0.01	0.22*	0.01
Leseverhalten						-0.01	0.01	0.01	0.01
Selbstkonzept Deutsch						0.19*	0.01	0.18*	0.01
Langeweile Deutsch						-0.03*	0.01	-0.03*	0.01
Interesse Deutsch						0.06*	0.01	0.06*	0.01
Anstrengung						0.07*	0.01	0.07*	0.01
kog. Grundfähigkeiten						0.05*	0.01	0.05*	0.01
Sozioöko. Hintergrund						0.04*	0.01	0.06*	0.01
Sprache zu Hause (1 = immer deutsch)						0.01	0.01	0.01	0.01
Lehrkraftebene									
Lehrkraftüberzeugung								-0.02	0.01
Schülergeschlecht × Lehrkraft-überzeugung								0.03*	0.01
Geschlecht (1 = weiblich)								0.01	0.01
Berufserfahrung (Jahre)								0.06	0.01
Schulart (1 = Gymnasium)								0.12*	0.01
(Residual)Varianzen									
Within	0.62*	0.57*	0.01	0.45*	0.01	0.40*	0.01	0.39*	0.01
Between	0.17*	0.16*	0.01	0.14*	0.01	0.11*	0.01	0.11*	0.01
Within R^2	–	0.08	–	0.27	–	0.35	–	0.37	–
Between R^2	–	0.06	–	0.18	–	0.35	–	0.35	–

Anmerkungen: Die Modelle 1 – 3 sind Random-Intercept-Modelle und Modell 4 ein Random-Slope-Modell. Alle kontinuierlichen Variablen, außer der AV (Deutschnote), gingen *z*-standardisiert in die Analysen ein. Daher wird der unstandardisierte Regressionskoeffizient *b* angegeben. Die Deutschnote wurde zur besseren Interpretation der Ergebnisse invertiert. Höhere Werte auf der Variable Leseverhalten kennzeichnen eine stärkere Vermeidung von schwierigen Lesetexten. R^2 wurde berechnet, indem jeweils die Residualvarianz von der Gesamtvarianz des Nullmodells abgezogen und durch diese dividiert wurde.
* $p < 0.01$

Vorhersage der Deutschnote durch Merkmale der Schülerinnen und Schüler

Die Ergebnisse der Mehrebenenmodelle zur Prädiktion der Deutschnote der Schülerinnen und Schüler sind in Tabelle 2 dargestellt.

In Modell 1 zeigte sich erwartungsgemäß ein signifikanter Zusammenhang des Geschlechts mit der Deutschnote in Höhe von $b = 0.46$ Notenpunkten (invertierter Wert). Dies spiegelt den in Tabelle 1 bereits beschriebenen Mittelwertunterschied in den Noten der Mädchen und Jungen wider. Zur Forschungsfrage I lässt sich somit festhalten, dass in der bundesweiten Stichprobe von Jugendlichen der neunten Jahrgangsstufe ein Notenvorteil der Mädchen im Fach Deutsch von einer halben Note besteht. In Modell 2 zeigte sich unter Kontrolle der Testleistung weiterhin ein signifikanter Zusammenhang des Geschlechts mit der Deutschnote von $b = 0.29$ Notenpunkten. Die Kompetenzen der Schülerinnen und Schüler im Lesen ($b = 0.15$, Zuhören ($b = 0.12$) und insbesondere in der Orthografie ($b = 0.29$) trugen erwartungsgemäß ebenfalls zur Vorhersage der Deutschnote bei. Somit blieb der Notenvorteil der Mädchen im Fach Deutsch auch dann noch bestehen, wenn ihre höheren Kompetenzen in diesem Fach berücksichtigt wurden (Forschungsfrage II). Im Modell 3 wurden das selbstberichtete Leseverhalten, die Anstrengungsbereitschaft, das Selbstkonzept, die Langeweile und das Interesse am Fach sowie die kognitiven Fähigkeiten, der sozioökonomische Hintergrund der Familie und die zu Hause verwendete Sprache eingeschlossen. Auch hier zeigte sich weiterhin ein signifikanter Effekt des Geschlechts ($b = 0.23$) und der Kompetenzen (Lesen: $b = 0.12$, Zuhören: $b = 0.08$, Orthografie: $b = 0.22$) auf die Deutschnote. Das fachspezifische Selbstkonzept ($b = 0.19$) war der stärkste Prädiktor unter den motivationalen Merkmalen der Schülerinnen und Schüler. Langeweile im Deutschunterricht ($b = -0.03$), Interesse ($b = 0.06$) sowie Anstrengungsbereitschaft ($b = 0.07$) trugen signifikant aber nur in geringem Maße zur Vorhersage der Deutschnote bei (siehe Tabelle 2). Zusammenfassend zeigten die Analysen, dass motivationale Merkmale einen zusätzlichen Beitrag zur Erklärung von Deutschnoten leisten, jedoch weiterhin Geschlechterdisparitäten in der Benotung bestehen bleiben (Forschungsfrage III).

Lehrkraftmerkmale und geschlechtsspezifische Lehrkraftüberzeugung

In Modell 4 (Tabelle 2) wurden schließlich Lehrkraftmerkmale und die geschlechtsspezifische Lehrkraftüberzeugung zum Leseverhalten sowie die Schulart einbezogen. Weder die Berufserfahrung (Jahre) der Lehrkräfte noch das Geschlecht der Lehrkraft standen in signifikantem Zusammenhang mit der Deutschnote. Für beide Merkmale wurde in weiteren Modellen (ohne Abbildung) geprüft, ob signifikante Crosslevel-Interaktionseffekte mit dem Schülergeschlecht vorliegen. Dies war nicht der Fall. Es zeigte sich weiterhin, dass Lehrkräfte an Gymnasien signifikant bessere Noten erteilten als Lehrkräfte nicht-gymnasialer Schularten ($b = 0.12$). Erwartungsgemäß zeigte sich bei der geschlechtsspezifischen Lehrkraftüberzeugung kein signifikanter Haupteffekt ($b = -0.02$). Um zu testen, ob ein differenzieller Effekt der Lehrkraftüberzeugung für die Gruppe der Mädchen oder der Jungen vorliegt, wurde im Modell 4 ein Random-Slope-Modell mit Crosslevel-Interaktion spezifiziert (Tabelle 2 und Abbildung 1).

Dabei zeigte sich ein kleiner signifikanter Interaktionseffekt von Schülergeschlecht und geschlechtsspezifischer Lehrkraftüberzeugung bezüglich des Leseverhaltens ($b = 0.03$) auf die Note im Fach Deutsch. Das heißt, wenn Lehrkräfte Jungen eine stärkere Vermeidung schwieriger Lesetexte zuschreiben als Mädchen, benoten sie nach Kontrolle von Kompetenzen, motivationalen und soziodemografischen Merkmalen sowie kognitiven Fähigkeiten Jungen im Fach Deutsch etwas schlechter als Mädchen (Forschungsfrage IV). Die Größe dieses Notennachteils ist jedoch gering.

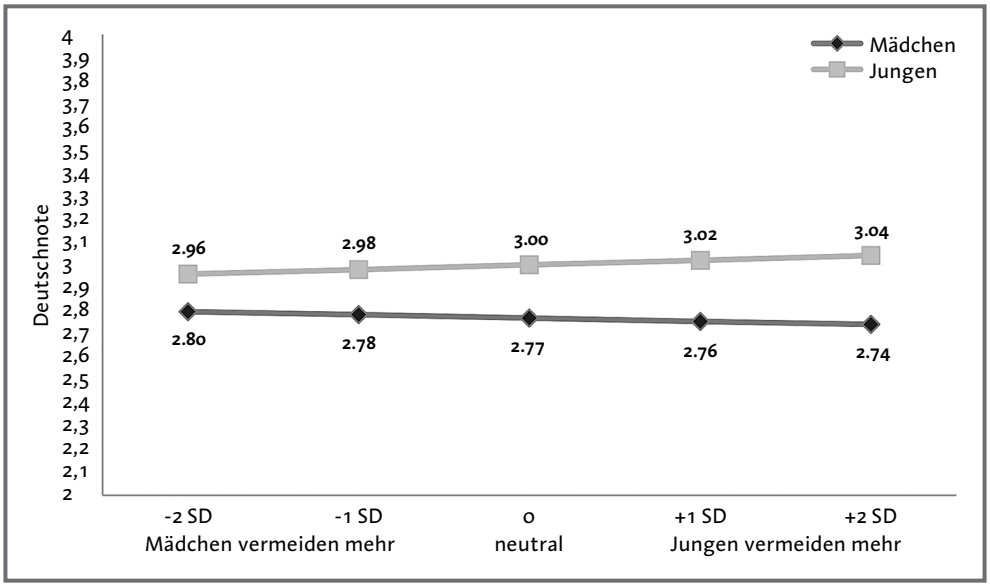

Abb. 1: Crosslevel-Interaktion: Deutschnote und geschlechtsspezifische Lehrkraftüberzeugung bezüglich des Leseverhaltens (Differenzwert $M_{Jungen} - M_{Mädchen}$) basierend auf Modell 5 in Tabelle 2. Zur besseren Interpretierbarkeit wurden die Noten für diese Abbildung nicht invertiert. Der Differenzwert wurde *z*-standardisiert.

Diskussion

Geschlechterunterschiede in der Benotung zugunsten von Mädchen sind lange bekannt (Voyer & Voyer, 2014; Brookhart et al., 2016) und werden in den letzten Jahren verstärkt diskutiert (Helbig, 2012; Kuhl & Hannover, 2012; Han et al., 2017). Dieser Beitrag prüfte, ob die erwartete geschlechtsdifferenzielle Benotung im Fach Deutsch repliziert werden kann und welchen Erklärungsbeitrag schulische Kompetenzen, motivationale und soziodemografische Merkmale sowie Lehrkraftfaktoren für (geschlechtsdifferenzielle) Benotung leisten.

Rolle von Individualmerkmalen

In der Datenanalyse zeigte sich im ersten Modell ein Notenunterschied von etwa einer halben Note zugunsten der Mädchen. Durch die schrittweise Hinzunahme weiterer Individualmerkmale reduzierte sich der Unterschied auf etwa eine Viertel Note im finalen Modell. Als wesentliches Ergebnis dieser Studie wurde ein Benotungsnachteil für Jungen im Fach Deutsch erneut bestätigt (Replikation von Han et al., 2017; Helbig, 2010; Kuhl & Hannover, 2012; Lehmann et al., 1997; Zinn & Bayer, 2018). Die höheren Kompetenzen und weitere relevante Faktoren konnten den Geschlechterunterschied in der Benotung im Fach Deutsch, wie auch in der Studie von Zinn und Bayer (2018), teilweise aber nicht vollständig erklären. Auch in Untersuchungen für das Fach Mathematik von Hochweber (2010), Lintorf (2012) und Lehmann et al. (2000) blieb nach Kontrolle von Kompetenzen und weiteren Merkmalen ein Vorteil für Mädchen zurück. Das fachliche Selbstkonzept zeigte wie bei Lehmann et al. (1997) deutliche Zusammenhänge mit der Zeugnisnote im Fach Deutsch, ähnlich den Befunden für das Fach Mathematik von Hochweber (2010) und Lehmann et al. (2000).

Da die drei Kompetenzbereiche Lesen, Zuhören und Orthografie getrennt voneinander in die Analysen eingingen, war es möglich, die Gewichtung dieser Kompetenzen bei der Notenvergabe nachzuvollziehen. Es zeigte sich, dass

insbesondere die Orthografieleistung der Jugendlichen mit der Deutschnote zusammenhängt. Dies wurde bereits von Valtin, Badel, Löffler, Meyer-Schepers und Voss (2003) für die vierte Jahrgangsstufe festgestellt. Als Gründe für diese Gewichtung der Rechtschreibkompetenz sind stärker kriterial orientierte Zensierungsmaßstäbe sowie eine Dominanz der Benotung von schriftlichen Leistungen denkbar. Dieser Befund verdeutlicht, dass weitere Forschung zur Gewichtung der einzelnen Kompetenzbereiche bei der Benotung notwendig ist. Auch wenn in dieser Studie bereits Kompetenztests in drei Bereichen enthalten waren, könnten zukünftige Studien noch weitere für die Benotung im Fach Deutsch relevante Kompetenzen wie z. B. kreatives Schreiben erfassen.

Rolle von Lehrkraftmerkmalen und der geschlechtsspezifischen Lehrkraftüberzeugungen

Als weiteres Ergebnis dieser Studie konnte zudem ein kleiner Benotungsnachteil für Jungen im Fach Deutsch im Zusammenhang mit geschlechtsspezifischen Überzeugungen von Lehrkräften gezeigt werden. Wenn eine Lehrkraft annimmt, dass Jungen mehr als Mädchen solche Texte präferieren, deren Handlung man leicht folgen kann, die kurz sind und keine Fachbegriffe enthalten, entsteht daraus womöglich eine etwas stärker negative Einschätzung der Kompetenzen oder motivationaler Merkmale der Jungen. Diese scheint in geringem Umfang in Zusammenhang mit geschlechterdifferenzieller Notenvergabe zu stehen. Das bedeutet, unsere Studie liefert weitere Hinweise darauf, dass Lehrkrafturteile in Form von Ziffernnoten mit Geschlechterstereotypen zusammenhängen können. Für Erwartungen und Einschätzungen von Lehrkräften, jedoch nicht für Ziffernnoten, zeigten diesen Zusammenhang mit Geschlechterstereotypen bereits Holder und Kessels (2017), Lorenz et al. (2016) sowie Muntoni und Retelsdorf (2018). Die Größe des Zusammenhangs von Benotungstendenzen und Lehrkraftüberzeugung in der vorliegenden Studie ist jedoch so gering, dass seine praktische Relevanz infrage gestellt und seine Replizierbarkeit in weiteren Studien untersucht werden muss. Angesichts einer möglichen Akkumulation von Benotungsnachteilen über die Zeit oder aus verschiedenen Quellen könnte aber auch dieser kleine Zusammenhang bedeutsam werden. Es gibt erste Hinweise darauf, dass Noten Ankereffekten, also einer Urteilsverzerrung aufgrund von zuvor erteilten Noten, unterliegen können (Dünnebier, Gräsel & Krolak-Schwerdt, 2009). Somit wäre eine sukzessive Notenverschlechterung durch den Einfluss geschlechtsspezifischer Lehrkraftüberzeugungen möglich. Hierzu wären längsschnittliche Untersuchungen zur Notenentwicklung innerhalb der Schullaufbahn notwendig, wie bereits von Dresel et al. (2006) sowie Han et al. (2017) für ausgewählte Jahrgangsstufen umgesetzt.

Limitationen und Forschungsbedarf

Trotz der methodischen Stärken der Studie, wie der großen Stichprobe und der Verwendung von Kompetenztests in drei für das Fach Deutsch maßgeblichen Kompetenzbereichen, gibt es Limitationen. Da zahlreiche theoretische Bedingungsfaktoren der Notengebung existieren (vgl. Hochweber, 2010), könnte die Interpretation der Ergebnisse durch mögliche Einflüsse weiterer unberücksichtigter Prädiktoren limitiert sein. Auch wenn die Testleistung in den Bereichen Lesen, Zuhören und Orthografie die Fachleistung umfangreicher abbildet als sie in anderen Large-Scale-Assessment-Studien (z. B. PISA oder IGLU) erfasst wird, blieben wesentliche Aspekte des Curriculums und der im Deutschunterricht zu erwerbenden Kompetenzen im IQB-Bildungstrend 2015 unberücksichtigt. Dazu gehören insbesondere produktive sprachliche Kompetenzen wie Schreibkompetenzen über Orthografie hinaus sowie Sprachgebrauch und Sprechen, aber auch die Rezeption von Literatur über die Lesekompetenz hinaus. Zum anderen wird die Leistungsbeurteilung der Lehrkräfte auch vom Arbeitsverhalten, wie z. B. der Mitarbeit der Schülerinnen und Schüler im Unter-

richt, dem Fleiß und der Hausaufgabenbearbeitung (Brookhart et al., 2016) beeinflusst. Dies konnte in der vorliegenden Studie nur näherungsweise über motivationale Merkmale berücksichtigt werden. Weiterhin konnte, wie in vielen Querschnittserhebungen üblich, bei denen die Notenvergabe und der Erhebungszeitraum der Studie nicht zusammenfallen, nur die Halbjahresnote aus dem Testjahr in den Analysen genutzt werden. Das bedeutet, dass die Note einige Wochen vor der Kompetenzmessung und Befragung erteilt wurde. Da sich Noten und motivationale Merkmale, insbesondere das Fähigkeitsselbstkonzept, reziprok beeinflussen (Marsh, Trautwein, Lüdtke, Köller & Baumert, 2005; Retelsdorf, Köller & Möller, 2014), ist es plausibel, dass eine vor Kurzem erteilte Note die motivationalen Prädiktoren bereits beeinflusst haben könnte. Auch wenn dieser Mechanismus die Schätzung der Geschlechterunterschiede nicht direkt beeinflusst haben sollte, wäre es in zukünftigen Erhebungen sinnvoll, die motivationalen Maße kurz vor der Notengebung zu erheben. Auch sollte die konkrete Praxis der Notenvergabe genauer untersucht werden, um bislang unberücksichtigte Faktoren ermitteln zu können. So sollten Lehrkräfte zu den von ihnen genutzten impliziten und expliziten Richtlinien zur Ermittlung der Zeugnisnoten sowie Schulleiterinnen und Schulleiter zu den schulinternen Regelungen in den jeweiligen Jahrgangsstufen befragt werden.

Auch ist die Operationalisierung der geschlechtsspezifischen Lehrkraftüberzeugungen – ein in theoretischer Hinsicht vielversprechendes Konstrukt – mit einer Kurzskala aus Items zur Vermeidung schwieriger Texte nicht optimal. Obwohl die Messgüte des Konstrukts gut ist, wäre es im Sinne der Vergleichbarkeit und Replikation mit anderen Daten besser, ein anderes Instrument zur Erfassung geschlechtsspezifischer Lehrkraftüberzeugungen zu nutzen (beispielsweise von Retelsdorf, Schwartz & Asbrock, 2015). Das Antwortverhalten der Lehrkräfte bei den geschlechtsspezifischen Items könnte zudem durch die Absicht, keine sozial unerwünschten stereotypen Angaben zu machen, beeinflusst worden sein. Allerdings ist anzumerken, dass trotzdem empirisch eine Verschiebung der Lehrkraftüberzeugung zugunsten der Mädchen zu beobachten war, auch wenn diese aufgrund von Erwünschtheitseffekten ggf. noch unterschätzt wird. Soziale Erwünschtheit stellt ein generelles Problem bei der Erfassung von gruppenbezogenen Einschätzungen dar (Stocké, 2004). Eine Möglichkeit, sozial erwünschtes Antwortverhalten von Lehrkräften in zukünftigen Studien zur differenziellen Benotung zu reduzieren, wären implizite Einstellungstestverfahren (Nosek, Greenwald & Banaji, 2005). Glock und Klapproth (2017) und Carlana (2019) nutzten diese beispielweise für andere lehrkraftbezogene Fragestellungen. Das Antwortverhalten der Lehrkräfte könnte allerdings auch durch einen Demand-Effekt (Orne, 1962) der Items beeinflusst worden sein. Durch die direkte Gegenüberstellung der Items für Mädchen und Jungen im Fragebogen könnten gruppenbezogene Kontraste besonders betont worden sein (Cinnirella, 1998). Dies hätte schließlich eher eine Über- statt Unterschätzung der stereotypen Lehrkrafteinschätzungen bedingt. Somit wäre es auch möglich, dass die geschlechtsspezifischen Lehrkrafteinschätzungen für einen Teil der befragten Personen unter- und für einen anderen Teil überschätzt wurden, sodass sich beide Effekte womöglich neutralisiert haben.

Ein interessanter Nebenbefund unserer Studie war, dass sich Geschlechtsunterschiede nur in der Lehrkraft- aber nicht der Selbsteinschätzung des Leseverhaltens der Schülerinnen und Schüler zeigten und diese Selbsteinschätzung keinen Effekt auf die Note hatte. Lehrkraft- und Schülerperspektive scheinen hier also unterschiedliche Aspekte zu erfassen. Es wäre daher interessant, sich in zukünftigen Studien detaillierter mit der Validität von Schüler- vs. Lehrkrafteinschätzungen zu befassen. Bisherige Studien zu Schüler- vs. Lehrkrafteinschätzungen von Noten weisen darauf hin, dass Schülerinnen und Schüler im Selbstbericht ihre Noten akkurat wiedergeben können (Dickhäuser & Plenter, 2005; Schneider & Sparfeldt, 2015).

Aktuelle Forschung weist auch zunehmend darauf hin, dass das Geschlecht als binärer Indikator nur ein grobes Bild des geschlechtsspezifischen Verhaltens und Erlebens der Jugendlichen, im Sinne von deren selbsteingeschätzter Feminität vs. Maskulinität, gibt. Alternativ kann Geschlecht auch auf einem Kontinuum beschrieben werden (Döring, 2013). Es wäre vielversprechend, solche Indikatoren zusätzlich zum binären Geschlecht in zukünftigen Studien zur Benotung zu integrieren. Schließlich sollten für eine umfassende Untersuchung geschlechtsdifferenzieller Benotung Befunde aus Beobachtungsstudien mit experimentellen Studien kombiniert werden. Denkbar wäre etwa der Einsatz von Vignetten. Beispielsweise haben Heyder und Kessels (2015, 2016) sowie Holder und Kessels (2017) bereits Vignetten zu geschlechtsbezogenen Fragestellungen im schulischen Kontext entwickelt und eingesetzt.

Praktische Implikationen

Als praktische Implikationen lassen sich eine stetige Sensibilisierung von Lehrkräften in Aus- und Fortbildung für verbreitete Geschlechterstereotype und die Reflexion der eigenen Notengebungspraxis ableiten. Hierbei sollten empirische Befunde zu Geschlechterunterschieden im schulischen Erfolg und Verhalten rezipiert und gleichzeitig über Grenzen der Übertragbarkeit von Forschungsergebnissen aus Schulleistungsstudien auf Individuen informiert werden.

Bei der Leistungsbeurteilung von Schülerinnen und Schülern könnten alternative Verfahren möglicherweise Geschlechterdisparitäten reduzieren. Machts und Möller (2019) kamen beim Vergleich von Kompetenzrastern, Noten und Leistungstests zu dem Ergebnis, dass eine differenzierte Leistungsrückmeldung zu fachlichen, überfachlichen und sozialen Kompetenzen im Rasterformat das Potenzial für eine Reduzierung von Geschlechtsdisparitäten haben kann. In ihrer Studie zeigten sie, dass das Geschlecht sowie das Sozialverhalten der Schülerinnen und Schüler die Noten im Fach Deutsch, nicht aber die mit Kompetenzrastern erhobene Leistungseinschätzung bedingte. Denkbar wäre zudem die Erprobung anonymer Benotung von schriftlichen Arbeiten – ähnlich den bereits in der Personalauswahl eingesetzten anonymisierten Bewerbungsverfahren. Auch nicht-anonyme alternative Formen der Beurteilung wie Berichtszeugnisse (Valtin et al., 2002) sollten analog zum Vorgehen von Machts und Möller (2019) hinsichtlich ihres möglichen Beitrags zur Verringerung von Geschlechtsunterschieden in der Leistungsbeurteilung evaluiert werden.

Literatur

Arens, A. K. (2019). Wertfacetten im Grundschulalter in drei Fächern: Differenzierung, Entwicklung, Geschlechtseffekte und Zusammenhänge zu Noten. *Zeitschrift für Pädagogische Psychologie,* 1–21. https://doi.org/10.1024/1010-0652/a000257

Artelt, C., Naumann, J. & Schneider, W. (2010). Lesemotivation und Lernstrategien. In E. Klieme, C. Artelt, J. Hartig, N. Jude, O. Köller, M. Prenzel, … P. Stanat (Hrsg.), *PISA 2009. Bilanz nach einem Jahrzehnt* (S. 73–112). Münster: Waxmann.

Baudson, T. G. & Preckel, F. (2013). Teachers' implicit personality theories about the gifted: an experimental approach. *School Psychology Quarterly, 28,* 37–46. https://doi.org/10.1037/spq0000011

Becker-Mrotzek, M., Böhme, K., Bulut, N., Hunger, S., Jost, J., Mörs, M., … Stanat, P. (2016). Integrierte Kompetenzstufenmodelle im Fach Deutsch. In P. Stanat, K. Böhme, S. Schipolowski, & N. Haag (Hrgs.), *IQB Bildungstrend 2015 Sprachliche Kompetenzen am Ende der 9. Jahrgangsstufe im zweiten Ländervergleich* (S. 95–126). Münster, New York: Waxmann.

Böhme, K., Sebald, S., Weirich, S. & Stanat, P. (2016). Geschlechtsbezogene Disparitäten. In P. Stanat, K. Böhme, S. Schipolowski, & N. Haag (Hrgs.), *IQB Bildungstrend 2015 Sprachliche Kompetenzen am Ende der 9. Jahrgangsstufe im zweiten Ländervergleich* (S. 377–402). Münster, New York: Waxmann.

Brookhart, S. M., Guskey, T. R., Bowers, A. J., McMillan, J. H., Smith, J. K., Smith, … Welsh, M. E. (2016). A Century of Grading Research: Meaning and Value in the Most Common Educational Measure. *Review of Educational Research, 86,* 803–848. https://doi.org/10.3102/0034654316672069

Carlana, M. (2019). Implicit Stereotypes: Evidence from Teachers' Gender Bias. *The Quarterly Journal of Economics, 134*, 1163–1224. https://doi.org/10.1093/qje/qjz008

Cinnirella, M. (1998). Manipulating Stereotype Rating Tasks: Understanding Questionnaire Context Effects on Measures of Attitudes, Social Identity and Stereotypes. *Journal of Community & Applied Social Psychology, 8*, 345–362. https://doi.org/10.1002/(SICI)1099-1298(1998090)8:5<345::AID-CASP441>3.0.CO;2-F

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* New York: Erlbaum.

Dickhäuser, O. & Plenter, I. (2005). „Letztes Halbjahr stand ich zwei." Zur Akkuratheit selbst berichteter Noten. *Zeitschrift für Pädagogische Psychologie, 19,* 219–224. https://doi.org/10.1024/1010-0652.19.4.219

Diefenbach, H. (2011). "Bringing Boys Back in" revisited. Ein Rückblick auf die bisherige Debatte über die Nachteile der Jungen im deutschen Bildungssystem. In A. Hadjar (Hrsg.), *Geschlechtsspezifische Bildungsungleichheiten* (S. 333–366). Wiesbaden: Springer. https://doi.org/10.1007/978-3-531-92779-4_14

Döring, N. (2013). Zur Operationalisierung von Geschlecht im Fragebogen: Probleme und Lösungsansätze aus Sicht von Mess-, Umfrage-, Gender- und Queer-Theorie. *GENDER – Zeitschrift für Geschlecht, Kultur und Gesellschaft, 5*(2), 94–113.

Dresel, M., Stöger, H. & Ziegler, A. (2006). Klassen- und Schulunterschiede im Ausmaß von Geschlechtsdiskrepanzen bei Leistungsbewertung und Leistungsaspiration: Ergebnisse einer Mehrebenenanalyse. *Psychologie in Erziehung und Unterricht, 52,* 46–61.

Duckworth, A. L. & Seligman, M. E. P. (2006). Self-Discipline Gives Girls the Edge: Gender in Self-Discipline, Grades, and Achievement Test Scores. *Journal of Educational Psychology, 98,* 198–208. https://doi.org/10.1037/0022-0663.98.1.198

Dünnebier, K., Gräsel, C. & Krolak-Schwerdt, S. (2009). Urteilsverzerrungen in der schulischen Leistungsbeurteilung. *Zeitschrift für Pädagogische Psychologie, 23,* 187–195. https://doi.org/10.1024/1010-0652.23.34.187

Enders, C. K. (2010). *Applied missing data analysis.* New York: Guilford Press.

Fuchs, G. & Brunner, M. (2017). Wie gut können bildungsstandardbasierte Tests den schulischen Erfolg von Grundschulkindern vorhersagen? *Zeitschrift für Pädagogische Psychologie, 31,* 27–39. https://doi.org/10.1024/1010-0652/a000195

Ganzeboom, H. (2010). *A new international socio-economic index [ISEI] of occupational status for the international standard classification of occupation 2008 [ISCO-08] constructed with data from the ISSP 2002–2007.* Paper presented at the Annual Conference of International Social Survey Programme, Lisbon.

Gattenmaier, K. (2004). *Literaturunterricht und Lesesozialisation: Eine empirische Untersuchung zum Lese- und Medienverhalten von Schülern und zur lesesozialisatorischen Wirkung ihrer Deutschlehrer.* Regensburg: edition vulpes.

Gentrup, S., Rjosk, C., Stanat, P. & Lorenz, G. (2018). Einschätzungen der schulischen Motivation und des Arbeitsverhaltens durch Grundschullehrkräfte und deren Bedeutung für Verzerrungen in Leistungserwartungen. *Zeitschrift für Erziehungswissenschaft, 21,* 867–891. https://doi.org/10.1007/s11618-018-0806-2

Glock, S. & Klapproth, F. (2017). Bad boys, good girls? Implicit and explicit attitudes toward ethnic minority students among elementary and secondary school teachers. *Studies in Educational Evaluation, 53,* 77–86. https://doi.org/10.1016/j.stueduc.2017.04.002

Han, M., Elsäßer, S., Lang, V. & Ditton, H. (2017). Geschlechtsspezifische Benotung? Der Einfluss der von Lehrkräften eingeschätzten Verhaltensmerkmale auf die Notengebung. *Zeitschrift für Soziologie der Erziehung und Sozialisation, 37,* 174–195.

Hannover, B. & Kessels, U. (2011). Sind Jungen die neuen Bildungsverlierer? Empirische Evidenz für Geschlechterdisparitäten zuungunsten von Jungen und Erklärungsansätze. *Zeitschrift für Pädagogische Psychologie, 25,* 89–103. https://doi.org/10.1024/1010-0652/a000039

Hannover, B., Wolter, I. & Zander, L. (2017). Geschlechtergerechtigkeit im Klassenzimmer. In T. Eckert & B. H. Gniewosz (Hrgs.), *Bildungsgerechtigkeit* (S. 201–213). Wiesbaden: VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-658-15003-7_12

Helbig, M. (2010). Sind Lehrerinnen für den geringeren Schulerfolg von Jungen verantwortlich? *Kölner Zeitschrift für Soziologie und Sozialpsychologie, 62,* 93–111. https://doi.org/10.1007/s11577-010-0095-0

Helbig, M. (2012). Warum bekommen Jungen schlechtere Schulnoten als Mädchen? Ein sozialpsychologischer Erklärungsansatz. *Zeitschrift für Bildungsforschung, 2,* 41–54. https://doi.org/10.1007/s35834-012-0026-4

Heyder, A. & Kessels, U. (2013). Is School Feminine? Implicit Gender Stereotyping of School as a Predictor of Academic Achievement. *Sex Roles, 69,* 605–617. https://doi.org/10.1007/s11199-013-0309-9

Heyder, A. & Kessels, U. (2015). Do teachers equate male and masculine with lower academic engagement? How students' gender enactment triggers gender stereotypes at school. *Social Psychology of Education, 18,* 467–485. https://doi.org/10.1007/s11218-015-9303-0

Heyder, A. & Kessels, U. (2016). Boys Don't Work? On the Psychological Benefits of Showing Low Effort in High School. *Sex Roles, 77,* 72–85. https://doi.org/10.1007/s11199-016-0683-1

Heyder, A., van Hek, M. & van Houtte, M. (2020). When Gender Stereotypes Get Male Adolescents into Trouble: A Longitudinal Study on Gender Conformity Pressure as a Predictor of School Misconduct. *Sex Roles, published online 16. 4. 2020.* https://doi.org/10.1007/s11199-020-01147-9

Hochweber, J. (2010). *Was erfassen Mathematiknoten? Korrelate von Mathematik-Zeugniszensuren auf der Schüler- und Schulklassenebene in Primar- und Sekundarstufe.* Münster, New York, München, Berlin: Waxmann.

Hoffmann, L. & Richter, D. (2016). Aspekte der Aus- und Fortbildung von Deutsch- und Englischlehrkräften im Ländervergleich. In P. Stanat, K. Böhme, S. Schipolowski, & N. Haag (Hrgs.), *IQB Bildungstrend 2015 Sprachliche Kompetenzen am Ende der 9. Jahrgangsstufe im zweiten Ländervergleich* (S. 95–126). Münster, New York: Waxmann.

Holder, K. & Kessels, U. (2017). Gender and ethnic stereotypes in student teachers' judgments: a new look from a shifting standards perspective. *Social Psychology of Education, 20,* 471–490. https://doi.org/10.1007/s11218-017-9384-z

Jansen, M., Schneider, R., Schipolowski, S. & Henschel, S. (2019). Motivationale Schülermerkmale im Fach Mathematik und in den naturwissenschaftlichen Fächern. In P. Stanat, S. Schipolowski, N. Mahler, S. Weirich, & S. Henschel (Hrsg.), *IQB-Bildungstrend 2018 Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I im zweiten Ländervergleich* (S. 337–354). Münster: Waxmann.

Jones, S. & Myhill, D. (2004). 'Troublesome boys' and 'compliant girls': gender identity and perceptions of achievement and underachievement. *British Journal of Sociology of Education, 25,* 547–561. https://doi.org/10.1080/0142569042000252044

Kessels, U. & Heyder, A. (2017). Die Wertschätzung schulischer Anstrengung als Mediator von Geschlechtsunterschieden in Noten. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 49,* 86 – 97. https://doi.org/10.1026/0049-8637/a000171

Kessels, U. & Heyder, A. (2018). Geschlechtsunterschiede. In D. Rost, J. R. Sparfeldt, & S. Buch (Hrgs.), *Handwörterbuch Pädagogische Psychologie* (S. 209 – 217). Weinheim: Beltz.

Krolak-Schwerdt, S., Böhmer, M. & Gräsel, C. (2009). Verarbeitung von schülerbezogener Information als zielgeleiteter Prozess. *Zeitschrift für Pädagogische Psychologie, 23,* 175 – 186. https://doi.org/10.1024/1010-0652.23.34.175

Kuhl, P. & Hannover, B. (2012). Differenzielle Benotungen von Mädchen und Jungen. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 44,* 153 – 162. https://doi.org/10.1026/0049-8637/a000066

Lauermann, F., Meißner, A. & Steinmayr, R. (2020). Relative Importance of Intelligence and Ability Self-Concept in Predicting Test Performance and School Grades in the Math and Language Arts Domains. *Journal of Educational Psychology, 112*, 364 – 383.

Lehmann, R., Peek, R. & Gänsfuß, R. (1997). *Aspekte der Lernausgangslage und der Lernentwicklung von Schülerinnen und Schülern, die im Schuljahr 1996/97 eine fünfte Klasse an Hamburger Schulen besuchten.* Bericht über die Erhebung im September 1996 (LAU 5). Hamburg: Eigendruck der Behörde für Schule, Jugend und Berufsbildung.

Lehmann, R., Peek, R., Gänsfuß, R., Lutkat, S., Mücke, S. & Barth, I. (2000). *QuaSUM Qualitätsuntersuchung an Schulen zum Unterricht in Mathematik. Ergebnisse einer repräsentativen Untersuchung im Land Brandenburg.* Teltow: Grabow.

Lintorf, K. (2012). *Wie vorhersagbar sind Grundschulnoten? Prädiktionskraft individueller und kontextspezifischer Merkmale.* Wiesbaden: VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-94339-8

Litman, J. A. (2008). Interest and deprivation factors of epistemic curiosity. *Personality and Individual Differences, 44,* 1585 – 1595. https://doi.org/10.1016/j.paid.2008.01.014

Lorenz, G., Gentrup, S., Kristen, C., Stanat, P. & Kogan, I. (2016). Stereotype bei Lehrkräften? Eine Untersuchung systematisch verzerrter Lehrererwartungen. *Kölner Zeitschrift für Soziologie und Sozialpsychologie, 68,* 89 – 111. https://doi.org/10.1007/s11577-015-0352-3

Lüdtke, O. & Robitzsch, A. (2017). Eine Einführung in die Plausible-Values-Technik für die psychologische Forschung. *Diagnostika, 63,* 193 – 205. https://doi.org/10.1026/0012-1924/a000175

Machts, N. & Möller, J. (2019). Geschlechterunterschiede auf Kompetenzrastern. Die Relevanz überfachlicher Kompetenzen. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 51,* 97 – 109. https://doi.org/10.1026/0049-8637/a000211

Marsh, H., Trautwein, U., Lüdtke, O., Köller, O. & Baumert, J. (2005). Academic Self-Concept, Interest, Grades, and Standardized Test Scores: Reciprocal Effects Models of Causal Ordering. *Child Development, 76,* 397 – 416. https://doi.org/10.1111/j.1467-8624.2005.00853.x

Muntoni, F. & Retelsdorf, J. (2018). Gender-specific teacher expectations in reading – The role of teachers' gender stereotypes. *Contemporary Educational Psychology, 54,* 212 – 220. https://doi.org/10.1016/j.cedpsych.2018.06.012

Muthén, L. K. & Muthén, B. O. (1998 – 2017). *Mplus User's Guide.* Eighth Edition. Los Angeles, CA: Muthén & Muthén.

Neugebauer, M., Helbig, M. & Landmann, A. (2010). Unmasking the Myth of the Same-Sex Teacher Advantage. *European Sociological Review, 27,* 669 – 689. https://doi.org/10.1093/esr/jcq038

Nosek, B. A., Greenwald, A. G. & Banaji, M. R. (2005). Understanding and Using the Implicit Association Test: II. Method Variables and Construct Validity. *Personality and Social Psychology Bulletin, 31,* 166 – 180. https://doi.org/10.1177/0146167204271418

Orne, M. T. (1962). On the social psychology of the psychological experiment: with particular reference to demand characteristics and their implications. *American Psychologist, 17,* 776 – 783. https://doi.org/10.1037/h0043424

Preckel, F., Götz, T. & Frenzel, A. (2010). Ability grouping of gifted students: Effects on academic self-concept and boredom. *British Journal of Educational Psychology, 80,* 451 – 472. https://doi.org/10.1348/000709909X480716

Ramm, G., Prenzel, M., Baumert, J., Blum, W., Lehmann, R. & Leutner, D. (Hrgs.). (2006). *PISA 2003: Dokumentation der Erhebungsinstrumente.* Münster: Waxmann.

RCoreTeam. (2017). *R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.* Retrieved from https://R-Forge.R-project.org/projects/eat/.

Reiss, K., Weis M., Klieme, E. & Köller, O. (2019). *PISA 2018. Grundbildung im internationalen Vergleich.* Münster, New York: Waxmann. https://doi.org/10.31244/9783830991007

Retelsdorf, J., Köller, O. & Möller, J. (2014). Reading achievement and reading self-concept – Testing the reciprocal effects model. *Learning and Instruction, 29,* 21 – 30. https://doi.org/10.1016/j.learninstruc.2013.07.004

Retelsdorf, J., Schwartz, K. & Asbrock, F. (2015). "Michael can't read!" – Teachers' gender stereotypes and boys' reading self-concept. *Journal of Educational Psychology, 107,* 186 – 194. https://doi.org/10.1037/a0037107

Sachse, K., Haag, N. & Weirich, S. (2016). Testdesign und Auswertung des IQB-Bildungstrend 2015: Technische Grundlagen. In P. Stanat, K. Böhme, S. Schipolowski & N. Haag (Hrgs.), *IQB-Bildungstrend 2015 Sprachliche Kompetenzen am Ende der 9. Jahrgangsstufe im zweiten Ländervergleich* (S. 509 – 526). Münster, New York: Waxmann.

Schipolowski, S., Haag, N., Böhme, K. & Sachse, K. (2016). Anlage, Durchführung und Auswertung des IQB-Bildungstrends 2015. In P. Stanat, K. Böhme, S. Schipolowski & N. Haag (Hrgs.), *IQB-Bildungstrend 2015 Sprachliche Kompetenzen am Ende der 9. Jahrgangsstufe im zweiten Ländervergleich* (S. 95 – 126). Münster, New York: Waxmann.

Schipolowski, S., Haag, N., Milles, F., Pietz, S. & Stanat, P. (2018). *IQB-Bildungstrend 2015. Skalenhandbuch zur Dokumentation der Erhebungsinstrumente in den Fächern Deutsch und Englisch.* Humboldt-Universität zu Berlin, Institut zur Qualitätsentwicklung im Bildungswesen. Berlin. https://doi.org/10.18452/19997

Schipolowski, S., Wittig, J., Mahler, N. & Stanat, P. (2019). Geschlechtsbezogene Disparitäten. In P. Stanat, S. Schi-

polowski, N. Mahler, S. Weirich & S. Henschel (Hrsg.), *IQB-Bildungstrend 2018 Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I im zweiten Ländervergleich* (S. 237–258). Münster: Waxmann.

Schneider, R. & Sparfeldt, J. R. (2015). Zur (Un-)Genauigkeit selbstberichteter Zensuren bei Grundschulkindern. *Psychologie in Erziehung und Unterricht, 63.* https://doi.org/10.2378/peu2016.art05d

Spiel, C., Wagner, P. & Fellner, G. (2002). Wie lange arbeiten Kinder zu Hause für die Schule? *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 34,* 125–135. https://doi.org/10.1026//0049-8637.34.3.125

Stanat, P., Böhme, K., Schipolowski, S., Haag, N., Weirich, S., Sachse, K., … Federlein, F. (2018). *IQB-Bildungstrend Sprachen 2015 (IQB-BT 2015) Version: 2.* IQB – Institut zur Qualitätsentwicklung im Bildungswesen. Datensatz. http://doi.org/10.5159/IQB_BT_2015_v2

Stanat, P. & Kunter, M. (2001). Geschlechtsunterschiede in den Basiskompetenzen. In J. Baumert, E. Klieme, M. Neubrand, M. Prenzel, U. Schiefele, W. Schneider, … M. Weiß (Hrgs.), *PISA 2000 Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (S. 250–269). Opladen: Verlag Leske + Budrich.

Steinmayr, R. & Spinath, B. (2008). Sex differences in school achievement: what are the roles of personality and achievement motivation? *European Journal of Personality, 22,* 185–209. https://doi.org/10.1002/per.676

Stocké, V. (2004). Entstehungsbedingungen von Antwortverzerrungen durch soziale Erwünschtheit. Ein Vergleich der Prognosen der Rational-Choice Theorie und des Modells der Frame-Selektion. *Zeitschrift für Soziologie, 33,* 303–320. https://doi.org/10.1515/zfsoz-2004-0403

Tent, L. (2001). Zensuren. In D. Rost (Hrsg.), *Handwörterbuch Pädagogische Psychologie* (S. 805–811). Weinheim, Basel: Beltz.

Valtin, R., Badel, I., Löffler, I., Meyer-Schepers, U. & Voss, A. (2003). Orthographische Kompetenzen von Schülerinnen und Schülern der vierten Klasse. In W. Bos, E.-M. Lankes, M. Prenzel, K. Schwippert, G. Walther & R. Valtin (Hrsg.), *Erste Ergebnisse aus IGLU. Schülerleistungen am Ende der vierten Jahrgangsstufe im internationalen Vergleich* (S. 227–264). Münster, New York, München, Berlin: Waxmann.

Valtin, R., Schmude, C., Rosenfeld, H., Darge, K., Ostrup, G., Thiel, … Wagner, C. (2002). *Was ist ein gutes Zeugnis? Noten und verbale Beurteilungen auf dem Prüfstand.* Weinheim: Juventa.

van Buuren, S. & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software, 45*(3), 1–67. https://doi.org/10.18637/jss.v045.i03

van Ophuysen, S. (2006). Vergleich diagnostischer Entscheidungen von Novizen und Experten am Beispiel der Schullaufbahnempfehlung. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 38,* 154–161. https://doi.org/10.1026/0049-8637.38.4.154

van Ophuysen, S. (2008). Zur Veränderung der Schulfreude von Klasse 4 bis 7. Eine Längsschnittanalyse schulformspezifischer Effekte von Ferien und Grundschulübergang. *Zeitschrift für Pädagogische Psychologie, 22,* 293–306. https://doi.org/10.1024/1010-0652.22.34.293

Voyer, D. & Voyer, S. D. (2014). Gender differences in scholastic achievement: a meta-analysis. *Psychological Bulletin, 140,* 1174–1204. https://doi.org/10.1037/a0036620

Wagner, W., Helmke, A. & Rösner, E. (2009). *Deutsch Englisch Schülerleistungen International. Dokumentation der Erhebungsinstrumente für Schülerinnen und Schüler, Eltern und Lehrkräfte.* Frankfurt am Main: Deutsches Institut für Internationale Pädagogische Forschung (DIPF).

Weirich, S. & Hecht, M. (2018). *eatRep: Educational Assessment Tools for Replication Methods.* R package version 0.9.5/r763. https://R-Forge.R-project.org/projects/eat/.

Wendt, H., Steinmayr, R. & Kasper, D. (2016). Geschlechterunterschiede in mathematischen und naturwissenschaftlichen Kompetenzen. In H. Wendt, W. Bos, C. Selter, O. Köller, K. Schwippert & D. Kasper (Hrsg), *TIMSS 2015 Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* (S. 257–298). Münster: Waxmann.

Wilhelm, O., Schroeders, U. & Schipolowski, S. (2014). *BEFKI 8–10. Berliner Test zur Erfassung fluider und kristalliner Intelligenz für die 8. bis 10. Jahrgangsstufe.* Göttingen: Hogrefe.

Wolter, K. M. (2007). *Introduction to variance estimation.* New York: Springer. https://doi.org/10.1007/978-0-387-35099-8_1

Zinn, S. & Bayer, M. (2018). Zur Entwicklung des Zusammenhangs und der Erklärbarkeit von Leistungen und Kompetenzen bei Schülerinnen und Schülern der Sekundarstufe I. *LIfBi Working Paper,* No. 77. https://doi.org/10.13140/RG.2.2.24669.97760/1

Christin Rüdiger
Dr. Malte Jansen
Dr. Camilla Rjosk
Institut zur Qualitätsentwicklung
im Bildungswesen (IQB)
Wissenschaftliche Einrichtung der Länder
an der Humboldt-Universität zu Berlin e.V.
Unter den Linden 6
D-10099 Berlin
E-Mail: christin.ruediger@iqb.hu-berlin.de
malte.jansen@iqb.hu-berlin.de
camilla.rjosk@iqb.hu-berlin.de