Z Erziehungswiss (2020) 23:1019–1052
https://doi.org/10.1007/s11618-020-00972-8

ZfE

ALLGEMEINER TEIL

Determinanten und Akkuratheit von Schülerurteilen über sprachliche Fähigkeiten von Mitschüler(inne)n im Deutschen und den Herkunftssprachen Türkisch und Russisch

Nora Dünkel · Michel Knigge · Jürgen Wilbert

Eingegangen: 10. April 2019 / Überarbeitet: 29. Mai 2020 / Angenommen: 21. September 2020 / Online publiziert: 7. Oktober 2020

Zusammenfassung Modellen der Sprachaneignung zufolge sind für die die Entwicklung sprachlicher Fähigkeiten auch die sprachlichen Fähigkeiten von wichtigen Interaktionspartnern (z. B. Peers) bedeutsam. Da objektive Kompetenzmaße von Interaktionspartnern selten verfügbar sind, könnten alternativ Fremdeinschätzungen der sprachlichen Fähigkeiten erhoben werden. Im Beitrag wurden daher Schülerurteile über sprachliche Fähigkeiten von Mitschüler(inne)n im Deutschen und den Herkunftssprachen Türkisch und Russisch als potentielle Indikatoren tatsächlicher Sprachfähigkeiten untersucht. Mit Hilfe von Mehrebenenmodellen wurde analysiert, welchen Einflussfaktoren die Fremdeinschätzungen unterlagen, wie akkurat diese ausfielen und unter welchen Bedingungen akkuratere Einschätzungen gelangen. In allen Sprachen ergaben sich moderate Zusammenhänge zwischen den Fremdeinschätzungen und objektiven Leistungsmaßen, wobei die Urteilsakkuratheit insbesondere von Merkmalen auf der Beziehungsebene zur eigeschätzten Person (gemeinsamer Unterricht, geteilter Sprachhintergrund, Beziehungsqualität) moderiert wurde. Im Deutschen zeigten sich negative leistungsbezogene Vorurteile gegenüber Jugendlichen mit türkischer und russischer Herkunftssprache. Die Ergebnisse wer-

N. Dünkel (✉)
Allgemeine, Interkulturelle und International Vergleichende Erziehungswissenschaft, Universität Hamburg, 20146 Hamburg, Deutschland
E-Mail: nora.duenkel@uni-hamburg.de

Prof. Dr. M. Knigge
Kultur-, Sozial- und Bildungswissenschaftliche Fakultät, Institut für Rehabilitationswissenschaften, Abteilung Rehabilitationspsychologie, Humboldt-Universität zu Berlin, 10099 Berlin, Deutschland
E-Mail: michel.knigge@hu-berlin.de

Prof. Dr. J. Wilbert
Zentrum für Empirische Inklusionsforschung, Arbeitsbereich Inklusionspädagogik mit dem Schwerpunkt Forschungsmethoden und Diagnostik, Universität Potsdam, 14476 Potsdam, Deutschland
E-Mail: juergen.wilbert@uni-potsdam.de

den in Bezug auf Möglichkeiten und Grenzen von Fremdeinschätzungen sprachlicher Fähigkeiten sowie auf Konsequenzen leistungsbezogener Vorurteile diskutiert.

Schlüsselwörter Fremdeinschätzungen · Peers · Sprachfähigkeiten · Urteilsakkuratheit

Determinants and accuracy of students' ratings of their peers' German and heritage language abilities

Abstract Models of language acquisition suggest that the development of language abilities is influenced by the language skills of relevant interaction partners (e.g. peers). Because objective measures of interaction partners' language skills are rarely available, third party ratings may be an alternative measure. Therefore, the present study investigates students' ratings of their fellow students' language skills as indicators of actual language performance in German and the heritage languages Turkish and Russian. Multilevel models were applied to address the following questions: Which factors influence students' ratings of peers' language skills? How accurate are these ratings and what influences the accuracy of ratings? In all languages, students' ratings were moderately related to peers' test performance and the accuracy of ratings was positively moderated if the students had class together, shared the same language background and with increasing relationship quality. The ratings for German language abilities further revealed negative performance related stereotypes towards peers with Turkish and Russian language backgrounds. The results are discussed with respect to possibilities and boundaries of assessing peers' language skills through student ratings and implications of negative performance related stereotypes.

Keywords judgement accuracy · language abilities · other ratings · peers

1 Einleitung

Sprachlichen Fähigkeiten kommen als Voraussetzung für das Lernen sowie Gegenstand von Leistungstests für den Bildungserfolg eine besondere Rolle zu (vgl. Gogolin 2009; Kempert et al. 2016). Für viele Schüler(innen) mit Migrationshintergrund gehören neben dem Deutschen auch Herkunftssprachen zu ihrem sprachlichen Repertoire. Die Aneignung und Weiterentwicklung sprachlicher Fähigkeiten erfolgt in sozialen Kontexten und in der Interaktion mit Menschen. Obwohl Kontakte zu Gleichaltrigen, sog. Peerbeziehungen, in der Adoleszenz von zunehmender Bedeutung sind (z. B. Heyer et al. 2012; Hannover und Zander 2016), wurde ihre Rolle für die Entwicklung verkehrs- und herkunftssprachlicher Kompetenzen von Jugendlichen noch unzureichend untersucht. Modellen der Sprachaneignung (z. B. Esser 2006) zufolge spielt u. a. die Qualität des sprachlichen Inputs eine wichtige Rolle. Demnach sollten die sprachlichen Fähigkeiten von wichtigen Interaktionspartnern, wie den Peers, im Zusammenhang mit den individuellen sprachlichen Fähigkeiten stehen. Allerdings stellt die Prüfung dieser Annahme hohe Anforderungen bei der

Datenerhebung, da sowohl die Erfassung von Netzwerken, in denen Jugendliche interagieren, als auch umfassende Testungen der befragten Jugendlichen und ihrer Peers in verschiedenen Sprachen notwendig wären. Bisher liegen solche Daten weitestgehend getrennt voneinander vor. Eine ökonomische Möglichkeit zur Erfassung sprachlicher Fähigkeiten relevanter Peers wäre die Erhebung von Fremdeinschätzungen ihrer sprachlichen Fähigkeiten. Dabei könnten Befragungspersonen je nach Forschungsinteresse bestimmte Peers aus ihrem (egozentrierten) Netzwerk benennen (für eine Beschreibung des Erhebungsverfahrens egozentrierter Netzwerke siehe Herz 2012) und die sprachlichen Fähigkeiten dieser Peers einschätzen. Bei solchen Erhebungsverfahren stellt sich jedoch die Frage nach der Akkuratheit der Einschätzung von Eigenschaften Dritter (vgl. Wolf 2010, S. 474 f.). Akkurate Fremdeinschätzungen der sprachlichen Fähigkeiten wären aber eine zentrale Voraussetzung, wenn diese als ökonomisches Verfahren eingesetzt werden sollten, um Aussagen über die Sprachkenntnisse von Netzwerkpersonen und deren Bedeutung für die individuelle Entwicklung sprachlicher Fähigkeiten treffen zu können. Über Fremdeinschätzungen unter Jugendlichen und deren Qualität ist bisher jedoch kaum etwas bekannt.

Vor diesem Hintergrund besteht das Ziel des vorliegenden Beitrags in der Analyse von Fremdeinschätzungen der sprachlichen Fähigkeiten von Mitschüler(inne)n im Deutschen sowie den Herkunftssprachen Türkisch und Russisch, um zu klären, inwieweit diese als Indikatoren für die tatsächlichen Sprachfähigkeiten der Peers herangezogen werden könnten. Durch den Einbezug der unterschiedlichen Sprachen kann bewertet werden, inwieweit Fremdeinschätzungen auch in Sprachen gelingen können, die in der Regel nicht im schulischen Kontext verankert sind und von den Urteiler(inne)n ggf. nicht selbst beherrscht werden. Die Ergebnisse sind somit insbesondere für Studiendesigns relevant, für die in Ermangelung objektiver Indikatoren die Erfassung der sprachlichen Fähigkeiten von Personen aus dem sozialen (Peer-)Umfeld von Studienteilnehmer(inne)n mittels Fremdeinschätzungen gewinnbringend wären. Unter Rückgriff auf Modelle der Urteilsakkuratheit von Lehrerurteilen akademischer Fähigkeiten von Schüler(innen) (vgl. Südkamp et al. 2012) wird im Rahmen des Beitrags untersucht, welchen Einflussfaktoren Schülerurteile über sprachliche Fähigkeiten von Mitschüler(inne)n unterliegen, wie akkurat diese ausfallen und unter welchen Bedingungen akkuratere Einschätzungen gelingen. Dabei stehen Eigenschaften der urteilenden Personen, der beurteilten Personen und Merkmale der Beziehung zwischen beiden im Fokus. Das Vorgehen ermöglicht zudem, potentielle leistungsbezogene Vorurteile unter Schüler(inne)n aufzudecken. Solche Vorurteile können sich nachteilig auf die Leistungsentwicklung auswirken, was insbesondere für ethnische Minderheiten in sprachlichen Leistungsdomänen nachgewiesen werden konnte (für eine Metaanalyse siehe Nadler und Clark 2011).

2 Fremdeinschätzungen akademischer Leistungen von Schüler(inne)n

Urteile über akademische Leistungen und Fähigkeiten von Schüler(inne)n werden im schulischen Kontext regelmäßig durch Lehrkräfte vorgenommen. Im Gegensatz zur formellen Diagnostik (z. B. psychometrische Intelligenztests), stehen bei Diagnoseleistungen von Lehrkräften informelle Urteile im Vordergrund (vgl. Schrader

2013). Mit der Frage, wie gut Lehrkräfte „Merkmale der Schülerinnen und Schüler, aber auch Lern- und Aufgabenanforderungen zutreffend einzuschätzen“ können, beschäftigt sich die pädagogisch-psychologische Forschung unter dem Stichwort ‚*Diagnostische Kompetenz*‘ (vgl. Karing und Artelt 2013, S. 167). Im Fokus dabei steht häufig die Urteilsakkuratheit als messbares Produkt diagnostischer Kompetenz, bei der die Übereinstimmung zwischen einem Lehrerurteil und über Tests oder Fragebögen erhobener Schülermerkmale untersucht werden (vgl. Praetorius und Südkamp 2017, S. 14 f.; Herppich et al. 2018). Ein gängiges Maß der Urteilsakkuratheit stellt die Rangkomponente[1] dar, welche den Zusammenhang zwischen Urteil und tatsächlicher Merkmalsausprägung korrelativ überprüft (vgl. Schrader 2013). In den Übersichtsarbeiten von Hoge und Coladarci (1989) und Südkamp et al. (2012) betrugen die gefundenen Korrelationen zwischen Lehrerurteilen akademischer Fähigkeiten und den erbrachten Schülerleistungen im Median $r = 0{,}66$ bzw. $r = 0{,}53$ und zeigen, dass Lehrer(innen) die Rangfolge der Schülerleistungen in ihren Klassen im Mittel recht gut einschätzen können (vgl. Hoge und Coladarci 1989, S. 303; Südkamp et al. 2012, S. 755). Insgesamt gelingt es ihnen dabei besser, leistungsbezogene Merkmale als andere Schülermerkmale (z. B. kognitive Fähigkeiten, akademische Selbstkonzepte, Lernmotivation) zu bewerten (siehe Südkamp et al. 2012; Machts et al. 2016; Spinath 2005). Zugleich zeigen sich auch in Bezug auf die Urteilsakkuratheit akademischer Fähigkeiten große Unterschiede zwischen Lehrkräften: Südkamp et al. (2012) berichten eine Spannweite der Korrelationen zwischen $r = -0{,}03$ und $r = 0{,}84$. Dies wirft die Frage nach Moderatoren der Urteilsakkuratheit auf (siehe Abschn. 3).

Im Kontext von *Schüler*urteilen akademischer Leistungen liegt der Forschungsfokus zumeist auf Zusammenhängen zwischen *Selbst*einschätzungen und leistungsbezogenen Merkmalen von Schüler(inne)n. Nach Brown et al. (2015) fallen diese bei Korrelationen zwischen $r = 0{,}20$ und $r = 0{,}80$ (wobei nur wenige Studien höhere Korrelationen als $r = 0{,}60$ berichten) überwiegend schwach bis moderat aus (vgl. Brown et al. 2015, S. 446). In ähnlichem Umfang bewegen sich auch Zusammenhänge zwischen Selbsteinschätzungen sprachlicher Fähigkeiten und Sprachtests (siehe Edele et al. 2015). *Fremd*einschätzungen von Fähigkeiten zwischen Schüler(innen) wurden bisher jedoch selten untersucht. Eine der Ausnahmen bildet eine Studie von Pohlmann et al. (2004), in welcher bei Gymnasiast(inn)en der siebten und neunten Klasse Fremdeinschätzungen von Schülerselbstkonzepten durch Lehrkräfte und Mitschüler(innen) untersucht wurden. Die berichtete Korrelation zwischen den Selbstkonzepten und Fremdeinschätzungen durch Mitschüler(innen) lag für das Fach Deutsch bei $r = 0{,}37$ und muss aufgrund der identischen Metrik der beiden Skalen als eher schwach bewertet werden (vgl. Pohlmann et al. 2004, S. 161 f.). Höher fiel der Zusammenhang zwischen den durch die Mitschüler(innen) eingeschätzten Selbstkonzepten und der Deutschnote ($r = 0{,}46$) aus; dieser übertraf sogar den der Deutschnote mit dem Fähigkeitsselbstkonzept der Schüler(innen) ($r = 0{,}39$;

[1] Auch andere Maße, wie z. B. die Niveau- und Differenzierungskomponenete sind in der Forschung geläufig (vgl. Karst 2017). Diese lassen sich jedoch nur ermitteln, wenn Urteil und Kriterium auf derselben Skala verortet sind. Da dies im vorliegenden Beitrag nicht gegeben war, wird aus Platzgründen auf eine Darstellung empirischer Befunde zu diesen Maßen verzichtet.

vgl. Pohlmann et al. 2004). Diesen Umstand erklären die Autoren mit der Tendenz, bei der Bewertung anderer Personen „leicht zugängliche Informationen" heranzuziehen (Pohlmann et al. 2004, S. 165). Ähnlich zeigte sich bei Lehrereinschätzungen von Schülerselbstkonzepten, dass diese stärker durch die Schülerleistungen als deren Selbstkonzepte vorhergesagt werden (vgl. Praetorius et al. 2010, S. 128): Das Fähigkeitsselbstkonzept einer Person speist sich aus der individuellen Interpretation der eigenen Fähigkeiten und entzieht sich damit der direkten Beobachtbarkeit, wohingegen Leistungen im Zentrum der Unterrichtswahrnehmung von Lehrkräften stehen.

Ferner fielen bei Pohlmann et al. (2004) die Schülerselbstkonzepte höher aus als die Fremdeinschätzungen durch Lehrkräften und Mitschüler(innen). Als mögliche Ursache nennen die Autoren den für Selbstberichte bekannten „self-serving bias", die Tendenz die eigenen Fähigkeiten zu Zwecken des Selbstschutzes zu überhöhen (Pohlmann et al. 2004, S. 164). Somit könnten Fremdeinschätzungen im Vergleich zu Selbsteinschätzungen sogar akkurater sein. Eine Studie von Stipek (1981) zeigte, dass Kinder zu Beginn der Schulzeit die Leistungen ihrer Peers akkurater einschätzten als ihre eigenen. Peers, die von Lehrkräften als leistungsstark bewertet wurden, wurden auch von Mitschüler(inne)n als intelligent eingeschätzt, wogegen ein Zusammenhang mit Lehrereinschätzungen bei Selbsteinschätzungen von Intelligenz erst ab der zweiten und dritten Klasse nachgewiesen werden konnte (vgl. Stipek 1981). Auch Befunde aus der Persönlichkeitsforschung zeigen, dass Peerratings der „Big-Five"-Persönlichkeitsmerkmale inkrementelle Validität über Selbsteinschätzungen hinaus besitzen oder sogar validere Prädiktoren für akademische Leistungen darstellen können (z. B. Bratko et al. 2006; Connelly und Ones 2010; Ziegler et al. 2010).

Dagegen zeigt eine Studie von Zander et al. (2014) anhand einer Stichprobe von Neuntklässlern, dass Fremdeinschätzungen durch negative leistungsbezogene Vorurteile gegenüber Schüler(inne)n mit Migrationshintergrund verzerrt sind. In der Studie bewerteten Schüler(innen) nach der Bearbeitung eines standardisierten Leistungstests in Mathematik ihre eigene Leistung und die eines selbstgewählten Peers, sodass tatsächliche sowie wahrgenommene Leistungsdifferenzen ermittelt werden konnten. Die Jugendlichen mit Migrationshintergrund neigten dazu, die Leistung ihrer Mitschüler(innen) ohne Migrationshintergrund zu überschätzen und ihre eigenen zu unterschätzen (vgl. Zander et al. 2014). Besonders deutlich wurde dieser Bias bei Schüler(innen) mit Migrationshintergrund, die einen leistungsschwächeren Peer ohne Migrationshintergrund wählten, diesen aber als genauso gut oder sogar besser einschätzten (vgl. Zander et al. 2014). Obwohl die Performanz von befreundeten Vergleichspartnern tendenziell zu optimistisch bewertet wurde, blieb dieser „pal effect" aus, wenn Schüler(innen) ohne Migrationshintergrund Peers mit Migrationshintergrund bewerteten (vgl. Zander et al. 2014, S. 66 f.).

Im Gegensatz zu Leistungsbeurteilungen von Lehrern handelt es sich bei Fremdeinschätzungen unter Schüler(inne)n um weniger klar umrissene Bewertungssituationen. Lehrer(innen) dokumentieren Kompetenzen und Lernfortschritte ihrer Schüler(innen) regelmäßig, verfügen über umfassende Erfahrungen in der Fähigkeitseinschätzung, können dabei auf verschiedene Informationsquellen zurückgreifen und die Art und Weise der Informationsverarbeitung an das Ziel der

Beurteilung anpassen (z. B. Dünnebier et al. 2009). Basierend auf dem Kontinuum-Modell von Fiske und Neuberg (1990) beschreiben Herppich et al. (2018) Informationsverarbeitungsprozesse bei Leistungsurteilen von Lehrkräften auf einem Kontinuum zwischen zwei distinkten Modi: Bei der Bildung eines ersten Eindrucks wird mit größerer Wahrscheinlichkeit der erste Modus aktiviert, bei dem die Urteilsbildung auf Grundlage weniger und leicht beobachtbarer Merkmale heuristisch abläuft, wogegen bei Urteilen mit weitreichenden Folgen (z. B. Schullaufbahnempfehlungen) gemäß dem zweiten Modus eine kontrollierte und individualisierte Informationsverarbeitung wahrscheinlicher wird. Dagegen fällt bei informellen Urteilen akademischer Fähigkeiten unter Schüler(innen) die Motivation für eine kriteriengeleitete und individualisierte Bewertung aufgrund fehlender Expertise und Urteilskonsequenzen womöglich gering aus. In der Folge könnten kategoriengeleitete Informationsverarbeitungsprozesse aktiviert werden, die anfälliger für Verzerrungen sind. Auch sind objektive Leistungsrückmeldungen nicht immer öffentlich oder gar verfügbar (z. B. im Falle der Herkunftssprachen), sodass sich Schüler(innen) vermutlich stärker auf Leistungseindrücke aus informellen Kontexten berufen und Bewertungen aus leicht beobachtbaren Verhaltensweisen ableiten müssen. Dies wirft die Frage auf, von welchen Faktoren die Urteile sowie deren Akkuratheit beeinflusst sein könnten.

3 Determinanten von Urteilen und Moderatoren der Urteilsakkuratheit

Mit den Bedingungen akkurater Einschätzungen durch Außenstehende beschäftigt sich das heuristische Modell von Südkamp et al. (2012, S. 756 f.), welches auf Grundlage theoretischer und empirischer Befunde Moderatoren der Urteilsakkuratheit für Lehrerurteile akademischer Leistungen von Schüler(inne)n beschreibt. Die Urteilsakkuratheit ergibt sich aus der Übereinstimmung des Lehrerurteils mit der

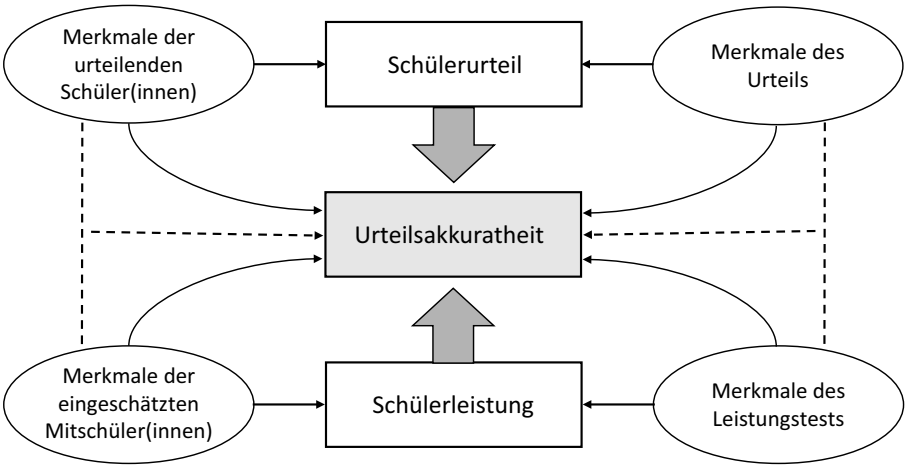

Abb. 1 Heuristische Modell der Urteilsakkuratheit nach Südkamp et al. (2012, S. 756) modifiziert auf das Anwendungsbeispiel von Fremdeinschätzungen unter Schüler(inne)n

Schülerleistung (meist gemessen als Korrelation) und wird beeinflusst von Merkmalen der Lehrkraft, Merkmalen der Schüler(innen), Eigenschaften des Leistungstests und des Urteils. Abb. 1 zeigt das Modell übertragen auf den Kontext von Fremdeinschätzungen unter Schüler(inne)n.

Auf der Ebene der einschätzenden Personen werden Personeneigenschaften als Moderationseffekte der Urteilsakkuratheit diskutiert, die die Art und Weise beeinflussen, wie gut urteilsrelevante Informationen aufgenommen, wahrgenommen und interpretiert werden können (vgl. Südkamp et al. 2012, S. 746; Förster und Böhmer 2017, S. 49). Dabei wird beispielsweise vermutet, dass höhere kognitive Grundfähigkeiten von Lehrkräften vorteilhaft für die Bewältigung intelligenter Denk- und Wahrnehmungsprozesse im Zusammenhang des Urteilens sein sollten: es müssen komplexe Informationen wahrgenommen, entschlüsselt, relevante Informationen erinnert und in einer angemessenen Beurteilung zusammengeführt werden (vgl. Kaiser et al. 2012, S. 253). In einer experimentellen Studie fanden Kaiser et al. (2012) einen positiven Moderationseffekt der kognitiven Fähigkeiten von Lehramtsstudierenden auf die Urteilsakkuratheit im computersimulierten Klassenraum. Zudem stehen kognitive Fähigkeiten eng im Zusammenhang mit fachlichen Leistungen und Kompetenzen (Schrader und Helmke 2008), die ebenfalls in Verbindung mit der Urteilsakkuratheit gebracht werden können. Einer Studie von Kruger und Dunning (1999) zufolge gelang es Studierenden mit geringen Testleistungen in einem Grammatiktest weder ihre eigenen noch die Leistungen ihrer Mitstudierenden im selben Test gut einzuschätzen. Bei leistungsstarken gegenüber leistungsschwachen Teilnehmer(inne)n fielen die Korrelationen zwischen den eingeschätzten und den tatsächlichen Leistungen ihrer Peers im Mittel fast doppelt so hoch aus ($r = 0{,}66$) (Kruger und Dunning 1999, S. 1127). Zu ähnlichen Ergebnissen kamen auch Analysen mit Selbsteinschätzungen (vgl. Alderson 2005; Brantmeier und Vanderplank 2008).

Hinsichtlich von Merkmalen der eingeschätzten Schüler(innen) liegen verschiedene Hinweise auf Moderationen der Urteilsakkuratheit vor. Während Befunde mehrheitlich auf positiv verzerrte Lehrerurteile literaler Fähigkeiten von Mädchen auch unter Kontrolle objektiver Leistungen deuteten (z. B. Hinnant et al. 2009; Ready und Wright 2011; Kuhl und Hannover 2012; Lorenz et al. 2016; Meissel et al. 2017), fanden andere Studien keine geschlechtsbezogenen Unterschiede von Leistungsurteilen (z. B. Karing et al. 2011; Zhu und Urhahne 2015). Nicht hinreichend belegt ist zudem, ob die höheren Leistungsurteile bei Mädchen auch im Sinne der Rangkomponente inakkurater ausfallen oder tatsächlich akkurater sind. Wiederholt konnte belegt werden, dass in Bewertungen akademischer Leistungen auch motivational-affektive Merkmale von Schüler(inne)n eingehen. Für Lehrkräfte zeigte sich, dass das Fähigkeitsselbstkonzept der Schüler(innen), ihre Unterrichtsbeteiligung sowie niedrigere Erfolgserwartungen und Leistungsangst mit dem Leistungsurteil zusammenhingen (vgl. Schrader und Helmke 1990; Rakoczy et al. 2008; Urhahne et al. 2010; Kaiser et al. 2013). Ähnliches ergab sich im Kontext von Nominierungen kompetenter Peers unter Grundschüler(inne)n. Diese wurden häufig mit Arbeitseinstellungen und -weisen, Persönlichkeitsmerkmalen und dem Sozialverhalten der eingeschätzten Mitschüler(innen) begründet (vgl. Stipek 1981). In einigen Studien konnte zudem ein über objektive Leistungen hinausgehender Einfluss des ethnischen Hintergrunds von Schüler(innen) auf die Leistungsbewertungen (z. B. bei Ready und

Wright 2011; Glock et al. 2015; Meissel et al. 2017) sowie die Leistungserwartungen von Lehrkräften (z. B. Tenenbaum und Ruck 2007; van den Bergh et al. 2010; Tobisch und Dresel 2017) zum Nachteil von Schüler(inne)n mit Migrationshintergrund ermittelt werden. Widersprüchliche Ergebnisse liefern Studien dabei jedoch in Bezug auf die Akkuratheit von Einschätzungen. Bei Zander et al. (2014, S. 67) ergaben sich z. B. akkuratere Urteile bei Schüler(innen) ohne Migrationshintergrund, die (stereotypengemäß) einen leistungsschwächeren Peer mit Migrationshintergrund bewertet hatten. Auch Kaiser et al. (2017) berichten akkuratere Lehrerurteile in Bezug auf Schüler(innen), die einer Minderheit im Klassenkontext angehörten (unabhängig davon, ob das Geschlecht oder der ethnische Hintergrund das Minderheitsmerkmal darstellte). Möglicherweise stellen sich in solchen Fällen größere Herausforderungen bei Bewertung, die Lehrerurteile aufgrund einer individualisierten Informationsverarbeitung akkurater werden lassen (vgl. Kaiser et al. 2017). Dagegen fanden z. B. Breidebach und Gruber (2018) sowie Glock et al. (2015) inakkuratere Lehrerurteile für Schüler(inne)n mit Migrationshintergrund.

Südkamp et al. (2012) unterstellen in ihrem Modell auch einen Einfluss der Beziehung zwischen Merkmalen der einschätzenden und eingeschätzten Personen, welcher aber bisher selten untersucht wurde. Eine Ausnahme bildet eine Studie von Zhu und Urhahne (2015), in der der Umgang der Lehrkräfte mit den Schüler(inne)n (aus Schülersicht) untersucht wurde. Jedoch moderierten keine der einbezogenen Beziehungsmerkmale (darunter: Lernunterstützung, Erreichbarkeit, Bevorzugung und gerechte Notenvergabe) die Akkuratheit der Lehrerurteile in der Fremdsprache Englisch. Unterschiedliche Befunde belegen den Einfluss des Bekanntheitsgrads zur eingeschätzten Person auf die Urteilsakkuratheit. In einer Studie von Ready und Wright (2011) zeigte sich, dass Urteilsverzerrungen literaler Fähigkeiten durch nicht leistungsbezogene Merkmale der Schüler(innen) (z. B. ethnische Herkunft) im Verlauf eines Schuljahres abnahmen, wenn die Lehrkräfte ihre Schüler(innen) besser kennengelernt hatten. Auch in der Persönlichkeitsforschung gilt als gesichert, dass die Qualität der Einschätzungen von Außenstehenden mit steigendem Bekanntheitsgrad zunimmt (vgl. Marsh und Craven 1991; Borkenau und Liebler 1993; Connelly und Ones 2010). Hayes und Dunning (1997) berichten beispielsweise höhere Übereinstimmungen zwischen Selbst- und Fremdeinschätzungen bezüglich verschiedener Persönlichkeitsmerkmale unter Freunden. Connelly und Ones (2010) betonen, dass dabei nicht allein die Kontakthäufigkeit entscheidend ist, sondern vor allem die emotionale Nähe Einblicke in Einstellungen und Gefühle einer anderen Person erlaubt, die zur verbesserten Einschätzung insbesondere von Persönlichkeitsmerkmalen führen, die eher im Verborgenen liegen (siehe auch Funder und Dobroth 1987; Clovin und Funder 1991; Vazire 2010). In Kontrast dazu stehen die Ergebnisse von Zander et al. (2014), wonach negative leistungsbezogene Vorurteile gegenüber Schüler(inne)n mit Migrationshintergrund auch unter befreundeten Jugendlichen bestehen.

Weitere Einflüsse auf die Urteilsakkuratheit können ferner von Merkmalen des Leistungsurteiles sowie des zum Vergleich herangezogenen Leistungstests ausgehen (siehe Abb. 1). Dabei ergaben sich höhere Korrelationen zwischen Lehrerurteilen und den Leistungen von Schüler(inne)n in Studien, in denen die Lehrkräfte über den als Vergleichsmaßstab herangezogenen Leistungstest in Kenntnis gesetzt wurden,

sowie in Studien, bei denen Leistungstests und Lehrerurteile kongruente Konstrukte erfassten (Südkamp et al. 2012). Da diese Moderatoren nicht Gegenstand des vorliegenden Beitrags sind, wird aus Platzgründen auf eine detailliertere Darstellung des Forschungsstands verzichtet.

4 Ziel der Studie und Fragestellungen

Ziel des Beitrags ist die Prüfung, inwieweit Fremdeinschätzungen der sprachlichen Fähigkeiten von Mitschüler(inne)n als Indikatoren für deren tatsächliche Sprachfähigkeiten herangezogen werden könnten. Dazu werden Determinanten sowie die Akkuratheit von Schülerurteilen über sprachliche Fähigkeiten von Mitschüler(inne)n im Deutschen und den Herkunftssprachen Türkisch und Russisch untersucht. Fokussiert werden dabei Eigenschaften der urteilenden Personen, der beurteilten Personen und Merkmale der Beziehung zwischen beiden. Im Rahmen der Analysen sollen erste Antworten auf folgende Fragen gefunden werden:

Fragestellung I Welche Merkmale der eingeschätzten Mitschüler(innen) beeinflussen die Fremdeinschätzungen sprachlicher Fähigkeiten über Testleistungen hinaus?

Fragestellung II Wie akkurat sind Fremdeinschätzungen der sprachlichen Fähigkeiten von Mitschüler(inne)n im Deutschen und den Herkunftssprachen Türkisch und Russisch?

Fragestellung III Von welchen Merkmalen der (a) urteilenden und (b) beurteilten Schüler(innen) sowie (c) deren Beziehung wird die Urteilsakkuratheit moderiert?

Die erste Forschungsfrage bezieht sich auf Faktoren, die die Fremdeinschätzungen über objektive Maßstäbe hinaus determinieren. Der Forschungsstand lässt erwarten, dass Schüler(inne)n, die zu Hause auch Türkisch oder Russisch sprechen, bei gleicher Leistung im Deutschen schlechtere Bewertungen ihrer Fähigkeiten erhalten als monolingual deutsch aufgewachsene Jugendliche (*H1*) (siehe Ready und Wright 2011; Zander et al. 2014; Meissel et al. 2017; Tobisch und Dresel 2017). Aufgrund der Hinweise auf eine Überschätzung literaler Fähigkeiten von Mädchen (siehe Hinnant et al. 2009; Ready und Wright 2011; Kuhl und Hannover 2012; Lorenz et al. 2016; Meissel et al. 2017) erwarten wir zudem, dass Mädchen im Vergleich zu Jungen positivere Fähigkeiten in den untersuchten Sprachen zugeschrieben werden (*H2*). Ferner zeigte sich bei Lehrer- und Schülerurteilen, dass diese von motivational-affektiven Schülermerkmalen beeinflusst werden (siehe Stipek 1981; Schrader und Helmke 1990; Rakoczy et al. 2008; Urhahne et al. 2010; Kaiser et al. 2013). Motivational-affektive Verhaltensweisen, wie Ängstlichkeit, Unterrichtsbeteiligung und Fleiß, sind relativ einfach zu beobachten und daher als Urteilskriterien schnell verfügbar. Leider standen solche Verhaltensindikatoren nicht zur Verfügung. Anzunehmen ist jedoch ein enger Zusammenhang zwischen solchen Verhaltensweisen und den Selbsteinschätzungen sprachlicher Fähigkeiten, da sich letztere in leicht

beobachtbarem Verhalten manifestieren könnten: Schüler(innen), die ihre eigenen Fähigkeiten in einer Sprache als hoch ansehen, verhalten sich womöglich offensiver, z. B. indem sie sich häufiger im Unterricht melden, die betreffende Sprache häufiger nutzen oder Mitschüler(innen) korrigieren. Daher soll geprüft werden, ob die Selbsteinschätzungen sprachlicher Fähigkeiten der eingeschätzten Mitschüler(innen) die Fremdeinschätzungen über objektive Leistungsmaße hinaus vorhersagen (*H3*).

Die Urteilsakkuratheit (Forschungsfrage II) wird im Sinne der Rangkomponente operationalisiert, als die Akkuratheit mit der die Rangfolge der Fähigkeiten von Peers korrekt eingeschätzt wird (vgl. Schrader 2013; Karst 2017). Dabei lässt sich angesichts des bisherigen Forschungsstandes eine signifikante, aber moderate Vorhersage der Fremdeinschätzungen durch die testbasierten Leistungsindikatoren sprachlicher Fähigkeiten erwarten (*H4*). Als potentielle Moderatoren der Urteilsakkuratheit (Forschungsfrage III) werden Merkmale der beurteilten Schüler(innen) (Geschlecht, Sprachhintergrund) und der urteilenden Schüler(innen) (kognitive Fähigkeiten, Geschlecht) sowie deren Beziehung (Beziehungsqualität, gleicher Sprachhintergrund, gemeinsamer Unterricht) untersucht. Aufgrund der engen Verbindungen zwischen kognitiven Fähigkeiten und Informationsverarbeitungsprozessen sowie fachlichen Kompetenzen (siehe Südkamp et al. 2012; Kaiser et al. 2012; Förster und Böhmer 2017), lässt sich ein positiver Moderationseffekt der kognitiven Fähigkeiten der einschätzenden Personen auf die Urteilsakkuratheit erwarten (*H5*). Ferner sollte die Genauigkeit von Fremdeinschätzungen beeinflusst werden durch die Qualität der Beziehung zur beurteilten Person (z. B. Marsh und Craven 1991; Hayes und Dunning 1997) und die Gelegenheiten, die zu bewertenden Eigenschaften dieser zu beobachten (Funder 1995). Erwartet werden daher akkuratere Fremdeinschätzungen mit zunehmender Beziehungsqualität zwischen urteilenden und beurteilten Schüler(inne)n sowie im Falle eines gemeinsamen Unterrichts (*H6*). Schließlich gehen wir auch der Frage nach, ob Fremdurteile unter Schüler(inne)n mit gleichem Sprachhintergrund besser gelingen. Insbesondere für die Herkunftssprachen Türkisch und Russisch liegt die Annahme auf der Hand, dass akkuratere Einschätzungen sprachlicher Fähigkeiten gelingen, wenn die Urteilenden selbst Kenntnisse in der betreffenden Sprache haben (*H7*). Da die Befunde zur Wirkung der Merkmale Geschlecht und des ethnischen bzw. sprachlichen Hintergrunds in Bezug auf die Urteils*akkuratheit* bisher noch nicht zufriedenstellend geklärt werden konnten (siehe Abschn. 3), beziehen wir diese ungerichtet in die Analysen ein.

5 Methoden

5.1 Stichprobe

Die Datengrundlage bildet der erste Messzeitpunkt (2016) des Forschungsprojekts *Mehrsprachigkeitsentwicklung im Zeitverlauf (MEZ)*. In diesem wurden Schüler(innen) der siebten und neunten Klasse befragt und Daten zu rezeptiven (Leseverstehen) und produktiven (Schreibfähigkeiten) sprachlichen Fähigkeiten in der Unterrichtssprache Deutsch sowie den Herkunftssprachen Türkisch und Russisch erhoben (siehe Gogolin et al. 2017). In einem Fragebogen waren die Jugendli-

chen zudem aufgefordert, jeweils bis zu drei andere Studienteilnehmer(innen) aus schulinternen, jahrgangsübergreifenden Teilnehmerlisten auszuwählen und diese hinsichtlich ihrer sprachlichen Fähigkeiten im Deutschen und sofern zutreffend auch deren herkunftssprachlichen Fähigkeiten im Russisch bzw. Türkisch einzuschätzen. Ferner machten die Befragten Angaben zu ihrer Beziehung zu den eingeschätzten Peers. Die Vorgabe lautete, schulintern drei andere Studienteilnehmer(innen) zu wählen, die sie kennen, wobei möglichst eine Person darunter sein sollte, mit der sie keinen gemeinsamen Unterricht hatten. So sollte eine ausreichende Bekanntschaft und zugleich eine Variabilität hinsichtlich der Beziehungsmerkmale erreicht werden.

Insgesamt hatten 1166 Schüler(innen) den Fragebogen zur Erfassung der Fremdeinschätzungen sprachlicher Fähigkeiten ausgefüllt. Die Teilnehmer(innen) waren im Durschnitt 14,6 Jahre alt, entstammten zu etwa gleichen Teilen den Jahrgängen sieben und neun und waren zu rund 63 % weiblich. Ein Gymnasium wurde von 56 % der Teilnehmer(innen) besucht. Rund 38 % der Jugendlichen waren monolingual deutsch, 24 % deutsch-russisch und 38 % deutsch-türkisch aufgewachsen. Die Stichprobe der Teilnehmer(innen) war zugleich jene, aus der die Peers zur Einschätzung entstammten, d. h. die Teilnehmer(innen) konnten zugleich Urteilende und Beurteilte sein. Hieraus ergab sich eine Datenstruktur, bei der Teilnehmer(innen) im Datensatz mehrfach auftauchen konnten: Nämlich als Urteilende (je nach Anzahl der vorgenommenen Einschätzungen) sowie als Beurteilte (abhängig davon, von wie vielen Teilnehmer(inne)n sie zur Bewertung ausgewählt wurden). Aufgrund der daraus resultierenden Abhängigkeiten in den Daten war eine besondere Analysemethode notwendig, die im Folgenden beschrieben wird.

5.2 Analysemethoden

Die Vorhersage der Fremdeinschätzungen sprachlicher Fähigkeiten im Deutschen und den Herkunftssprachen Türkisch und Russisch erfolgte mittels kreuzklassifizierter Mehrebenenmodelle (siehe Hox et al. 2018), um die genestete Datenstruktur der Fremdeinschätzungen (L1) innerhalb von Schulen nach urteilenden Schüler(inne)n (L2) und beurteilten Mitschüler(inne)n (L2) angemessen zu berücksichtigen.[2] Die Vorgehensweise orientiert sich an einem Anwendungsbeispiel von Karst et al. (2017) zur Bestimmung von Kennwerten diagnostischer Kompetenz mittels linearer Mischmodelle. Im Unterschied zu streng hierarchischen Modellen wurden im vorliegenden Beitrag zwei Zufallseffekte modelliert, die zulassen, dass die Mittelwerte der Fremdeinschätzungen nach urteilenden Schüler(inne)n und eingeschätzten Mitschüler(inne)n variieren können (siehe Hox et al. 2018, S. 161 f.). Auf dem untersten Level (L1) befinden sich die Fremdeinschätzungen (i), die sich aus einer spezifischen Kombination von Urteiler(inne)n (j) und Beurteilten (k) ergeben. Diese Einschätzungen bilden eine Kreuzklassifizierung, da sie innerhalb der übergeordneten Gruppierungsebene (L2) von Schulen gleichzeitig bestimmten Urteiler(inne)n und Beurteilten zugeordnet werden können. Für das Nullmodell kann die Fremdeinschätzung (i) einer urteilenden Person (j) und einem eingeschätz-

[2] Auf eine Berücksichtigung der hierarchischen Ebene der Schulklasse wurde aufgrund der zu geringen Fallzahlen innerhalb dieser verzichtet.

ten Peer (k) modelliert werden durch die folgende Formel (vgl. Hox et al. 2018, S. 167 f.):

$$Y_{i(jk)} = \beta_0 + u_{0j} + v_{0k} + e_{i(jk)}.$$

Dabei wird die abhängige Variable (hier die Fremdeinschätzung) modelliert durch einen Intercept β_0 sowie die Residualfehlerterme u_j für Urteiler(innen) und v_{0k} für Beurteilte in Schulen und das Regressionsresiduum $e_{i(jk)}$. Ermittelt wird dabei ein fester Effekt für einen Gesamt-Intercept sowie die als Zufallseffekte modellierten Varianzkomponenten σ_e^2 für die Varianz der Fremdeinschätzungen nach Urteiler(innen) σ_u^2 und Beurteilten σ_v^2. Dieses Grundmodell kann nun z. B. durch die Aufnahme von Merkmalen der Urteiler(innen) und Beurteilten als Prädiktoren erweitert werden (z. B. $\text{Fremdeinschätzung}_{i(jk)} = \beta_0 + \beta_1 \textit{Geschlecht-Urteiler}_j + \beta_2 \textit{Geschlecht-Beurteilter}_k + u_{0j} + v_{0k} + e_{i(jk)}$).

Alle Analysen erfolgten in R (R Core Team 2017) mit dem Paket „lme4“ (Bates et al. 2015). Berechnet wurden getrennte Modelle zur Vorhersage der Fremdeinschätzungen sprachlicher Fähigkeiten im Deutschen, im Türkischen und im Russischen. Zur korrekten Interpretation des Intercepts und der Interaktionsterme wurden die abhängigen Variablen sowie alle metrischen Prädiktoren jeweils am Gesamtmittelwert z-standardisiert. Um die Frage zu beantworten, welche Merkmale den Zusammenhang zwischen den Fremdeinschätzungen und den Testleistungen der Peers (also die Urteilsakkuratheit) moderieren, wurden Interaktionsterme der potentiellen Moderatoren mit den Testleistungen der Peers in die Modelle aufgenommen. Zur Prüfung der Signifikanz der Parameter wurden Modelle mittels Likeliehood-Ratio-Tests (siehe Hox et al. 2018) gegeneinander getestet, die die entsprechenden Parameter enthalten bzw. nicht enthalten. Zudem wurden zur Ermittlung der Effektstärken marginale (fixed effects) und konditionale (Gesamtmodell) R-Quadrat Schätzer berechnet (nach Nakagawa und Schielzeth 2013). Erstere können zudem genutzt werden, um Effekte der Testleistungen der eingeschätzten Peers auf die Fremdeinschätzungen mit den quadrierten Korrelationen aus anderen Studien zu vergleichen.

5.3 Variablen

5.3.1 Fremdeinschätzungen sprachlicher Fähigkeiten (Abhängige Variablen)

Die verwendeten Fremdeinschätzungsskalen sind angelehnt an die Erfassung leistungsbezogener Selbstkonzepte (nach Wagner et al. 2009). Für die Sprachen Deutsch, Türkisch und Russisch wurden mit Hilfe von jeweils drei Items Einschätzungen der Fähigkeiten[3] der gewählten Peers im *Texte lesen und verstehen, Texte schreiben* und *Wortschatz/Vokabeln* auf einer sechs-stufigen Notenskala erfasst. Dabei wurden die Items jeweils so rekodiert, dass ein hoher Skalenwert für gute sprachliche Fähigkeiten steht. Die gebildeten Einschätzungsskalen für das Deutsche (M = 4,8; SD = 0,8; Min = 1; Max = 6) und die Herkunftssprachen Türkisch

[3] Für alle Sprachen wurde gefragt: „Wie schätzt du die Fähigkeiten der Personen in folgenden Sprachen ein?“.

(M = 4,5; SD = 0,9) und Russisch (M = 4,8; SD = 0,9) wiesen eine hohe interne Konsistenz auf (Deutsch $\alpha = 0{,}85$; Russisch $\alpha = 0{,}89$; Türkisch $\alpha = 0{,}90$). Fälle, bei denen Schüler(innen) angaben, die Fähigkeiten des gewählten Peer nicht einschätzen zu können, wurden aus den Analysen ausgeschlossen.

5.3.2 Testbasierte Leistungsindikatoren für Sprachfähigkeiten der eingeschätzten Mitschüler(innen)

Zur Abbildung sprachlicher Fähigkeiten der eingeschätzten Mitschüler(innen) im Deutschen und den Herkunftssprachen Russisch und Türkisch wurden Maße für rezeptive sowie produktive sprachliche Fähigkeiten herangezogen. Bei ersterem handelt es sich um einen Lesegeschwindigkeits- und Leseverständnistest (LGVT 5–12+ nach Schneider et al. 2017; sowie in Kooperation entwickelte Parallelversionen für das Russische und Türkische). Der Test wurde speziell für die Klassenstufen 5–12 entwickelt und „weist eine hinreichende Korrelation mit dem PISA 2000-Leseverständnistest auf" (Gogolin et al. 2017, S. 16). Dabei handelt es sich um einen Fließtext, bei dem an verschiedenen Stellen aus drei Alternativen das Wort unterstrichen werden sollte, welches am besten in den Textzusammenhang passt. Verwendet wurden jeweils die erzielten Rohwerte für das *Leseverstehen* im Deutschen (Min = 0; Max = 85; M = 29,2; SD = 10,4; $\alpha = 0{,}82$) sowie den Herkunftssprachen Türkisch (Min = –12; Max = 46; M = 8,5; SD = 9,0; $\alpha = 0{,}64$) und Russisch (Min = –8; Max = 57; M = 9,4; SD = 12,3; $\alpha = 0{,}86$). In die Rohwertzählung gingen korrekte Unterstreichungen mit zwei Punkten, falsche oder mehrere Unterstreichungen mit einem Minuspunkt und Auslassungen mit null Punkten ein.

Die schriftliche Sprach*produktion* wurde mit dem Schreibimpuls *Lebkuchenhaus* (basierend auf dem „FörMig-Bumerang" nach Dirim und Döll 2009) erfasst. Bei dieser Aufgabe waren die Schüler(innen) aufgefordert, auf Grundlage einer Bilderfolge die darin dargestellte Anfertigung eines Lebkuchenhauses als Artikel für ein Jugendmagazin schriftlich wiederzugeben. In die Analysen ging die Punktzahl für die *Aufgabenbewältigung* im Deutschen (Min = 9; Max = 27; M = 17,9; SD = 3,8; $\alpha = 0{,}84$) sowie den Herkunftssprachen Türkisch (Min = 0; Max = 24; M = 12,7; SD = 4,6; $\alpha = 0{,}90$) und Russisch ein (Min = 0; Max = 26; M = 10,6; SD = 6,1; $\alpha = 0{,}95$). Dazu bewerten geschulte Auswerter(innen) für jeden der neun abgebildeten Arbeitsschritte, ob dieser „nicht", „angedeutet", „einfach" oder „differenziert, ausführlich" durch die Schüler(innen) beschrieben wurde. Je nach dem konnten jeweils Null, ein, zwei oder drei und insgesamt maximal 27 Punkte vergeben werden.

5.3.3 Weitere Merkmale der eingeschätzten Mitschüler(innen)

In allen Modellen wurden das *Geschlecht* (0 = männlich, 1 = weiblich), die besuchte *Klassenstufe* (0 = Jahrgang 7; 1 = Jahrgang 9), der besuchte *Bildungsgang* (0 = andere, 1 = Gymnasium) sowie die *selbsteingeschätzten Sprachfähigkeiten* einbezogen. Die Selbsteinschätzungen sprachlicher Fähigkeiten (nach Wagner et al. 2009) wurden mit Hilfe von rekodierten Notenskalen (sechsstufig) erfasst. In diese gingen je sechs Items zur Bewertung der Fähigkeiten und Kenntnisse in den Bereichen „Aussprache", „Texte schreiben", „Rechtschreibung und Zeichensetzung", „richtige Gram-

matik beim Sprechen“, „richtige Grammatik beim Schreiben“ und „Wortschatz und Vokabeln“ ein[4]. Die Mittelwerte der hoch intern konsistenten Skalen lagen bei 5,1 Punkten im Deutschen (Min = 1,3; Max = 6,0; SD = 0,7; $\alpha = 0{,}87$), 4,2 Punkten im Türkischen (Min = 1,2; Max = 6,0; SD = 1,0; $\alpha = 0{,}92$) und 4,1 Punkten im Russischen (Min = 1,8; Max = 6,0; SD = 1,0; $\alpha = 0{,}90$). Für die Modelle im Deutschen wurden darüber hinaus Dummyvariablen generiert, die den *Sprachhintergrund* der eingeschätzten Mitschüler(innen) abbildeten (25 % deutsch-russisch; 36 % deutsch-türkisch).

5.3.4 Merkmale der urteilenden Schüler(innen)

In allen Modellen wurden das *Geschlecht* und die *kognitiven Fähigkeiten* der urteilenden Schüler(innen) berücksichtigt. Letztere wurden mit Hilfe des nonverbalen Untertests (N2) des kognitiven Fähigkeitstests (KFT 4–12 + R) von Heller und Perleth (2000) erhoben. Der Test besteht aus 25 Multiple-Choice-Aufgaben, bei denen die Schüler(innen) figurale Analogien erkennen sollten. Um eine Vergleichbarkeit der Testwerte zwischen Schüler(inne)n der Klassen sieben und neun zu gewährleisten, wurden die auf Basis der Summe richtiger Antworten gebildeten jahrgangsnormierten T-Wertsummen verwendet ($0{,}85 < \alpha < 0{,}92$).

5.3.5 Beziehungsmerkmale

Auf der Beziehungsebene wurden die *Beziehungsqualität*, der *gemeinsame Unterrichtsbesuch* und das *Vorhandensein eines identischen Sprachhintergrunds* (Dummy) als Moderatoren für die Urteilsakkuratheit berücksichtigt. Zur Erfassung der Beziehungsqualität wurden Angaben zum Bekanntheitsgrad[5] sowie der subjektiven Wichtigkeit[6] der einzuschätzenden Zielpersonen für die Befragten zu einem Mittelwert verrechnet ($r = 0{,}79$). Der Mittelwert aus diesen zwei fünfstufigen Items kann im Falle einer hohen Punktzahl als eine hohe Beziehungsqualität zwischen einschätzender und eingeschätzter Person interpretiert werden (Modelle Deutsch: M = 3,5; SD = 1,2; Türkisch: M = 3,6; SD = 1,2; Russisch: M = 3,4; SD = 1,3). Die befragten Schüler(innen) waren ferner aufgefordert anzugeben, ob sie mit der einzuschätzenden Person (irgendeinen) gemeinsamen Unterricht hatten. In den Modellen im Deutschen beziehen sich 79 % der Fremdeinschätzungen auf Personen, die die urteilenden Schüler(innen) aus dem Unterricht kennen (Türkisch 69 %; Russisch 75 %). Der Anteil von Fremdeinschätzungen unter Jugendlichen mit geteiltem Sprachhintergrund lag in den Modellen im Deutschen bei rund 36 % (Türkisch 67 %; Russisch 57 %).

[4] Wortlaut der Fragestellung: „Wie schätzt du deine Beherrschung der deutschen Sprache ein?“ bzw. „Wie schätzt du deine Fähigkeiten und Kenntnisse in der [russischen/türkischen] Sprache ein?“.

[5] Frage: „Wie gut kennst du diese Personen?“; Fünfstufige Skala: „weniger gut“ (1), „sehr gut“ (5).

[6] Frage: „Wie wichtig sind dir diese Personen?“; Fünfstufige Skala: „gar nicht wichtig“ (1), „sehr wichtig“ (5).

6 Ergebnisse

6.1 Fremdeinschätzungen sprachlicher Fähigkeiten im Deutschen

Tab. 1 zeigt die Modelle zur Vorhersage der Fremdeinschätzungen sprachlicher Fähigkeiten im Deutschen mit zwei random intercepts zur Berücksichtigung der genesteten Datenstruktur nach urteilenden und beurteilten Schüler(innen) in Schulen. Der Varianzanteil, der auf die Clusterung nach urteilenden Schüler(innen) zurückzuführen ist, lag bei ICC = 28 %, für die eingeschätzten Mitschüler(innen) sind es ICC = 35 % (*D0*). Die Testleistungen der Peers im Schreiben und Lesen erklärten im Modell *D1* gemeinsam 12 % der Varianz der Fremdeinschätzungen sprachlicher Fähigkeiten. Erreichten die eingeschätzten Peers eine um eine Standardabweichung höhere Testleistung in der Schreibaufgabe, fiel das Schülerurteil um b = 0,18 Standardabweichungen höher aus. Ein Anstieg der Testleistung im Lesen war mit einer Zunahme der Fremdeinschätzung um b = 0,25 Standardabweichungen assoziiert. Auch getrennt erwiesen sich die beiden Prädatoren als signifikant, wobei die Varianzaufklärungen durch die Testleistungen im Schreiben bei 6 % und durch das Leseverständnis bei 9 % lagen (siehe Tab. 4 im Anhang). Im Modell *D2* wurden die weiteren Merkmale der eingeschätzten Mitschüler(innen) aufgenommen. Während die Fremdeinschätzungen bei Mädchen und Peers mit höheren Selbsteinschätzungen sprachlicher Fähigkeiten im Deutschen unter Kontrolle des Bildungsgangs und der Jahrgangsstufe höher ausfielen, wurden Jugendliche mit deutsch-russischem und deutsch-türkischem Sprachhintergrund signifikant schlechter bewertet. Im Modell *D3* zeigte sich ein signifikanter Geschlechtereffekt, wonach weibliche Urteilende die sprachlichen Fähigkeiten ihrer Peers um b = 0,24 Standardabweichungen höher einschätzten. Der positive Haupteffekt der kognitiven Fähigkeiten ist aufgrund des geringen Regressionsgewichts dagegen eher zu vernachlässigen (*D3*). Auf der Beziehungsebene ergab sich lediglich für die Beziehungsqualität ein signifikanter Haupteffekt (siehe *D4*), wonach die Schülerurteile mit steigender Beziehungsqualität zur eingeschätzten Person höher ausfielen. Während alle anderen Effekte weitestgehend stabil blieben, ergaben sich unter Berücksichtigung der Beziehungsmerkmale nunmehr keine positiv verzerrten Bewertungen weiblicher Peers.

Die durchgeführten Likelihood-Ratio-Tests verwiesen auf signifikante Modellverbesserungen durch die aufgenommenen Prädiktoren. Im Modell *D4* wurden 26 % der Varianz der Fremdeinschätzungen durch die Prädiktoren gemeinsam erklärt. Zur Abschätzung, welche der Prädiktoren den stärksten Effekt auf die Fremdeinschätzungen sprachlicher Fähigkeiten hatten, wurden für das Modell *D4* zusätzlich standardisierte Regressionskoeffizienten[7] berechnet. Dabei erwiesen sich ein türkischer Sprachhintergrund der Peers ($\beta = -0{,}20$), die Selbsteinschätzungen ($\beta = 0{,}17$) sowie die Leseleistungen ($\beta = 0{,}16$), als stärkste Prädiktoren.

Zur Überprüfung, ob das Geschlecht und der Sprachhintergrund der eingeschätzten Peers, die Urteilermerkmale Geschlecht und kognitive Fähigkeiten sowie die Merkmale auf Beziehungsebene (Beziehungsqualität, gemeinsamer Unterricht, geteilter Sprachhintergrund) die Urteilsakkuratheit moderieren, wurden aufbauend auf

[7] Auf eine Darstellung der Koeffizienten in den Ergebnistabellen wurde aus Platzgründen verzichtet.

Tab. 1 Kreuzklassifizierte Mehrebenenmodelle für Fremdeinschätzungen im Deutschen (z-standardisiert)

		Modell D0		Modell D1		Modell D2		Modell D3		Modell D4		Modell D4a		Modell D4b	
		b	(se)	b	(se)	b	(se)	b	(se)	b	(se)	b	(se)	b	(se)
Fixe Parameter															
Konstante		–0,01	(0,03)	–0,01	(0,03)	0,13*	(0,06)	–0,46**	(0,12)	–0,50**	(0,12)	–0,50**	(0,12)	–0,49**	(0,12)
Prädiktoren eingeschätzte Personen															
a	Testleistung Schreibaufgabe	–	–	0,18**	(0,02)	0,10**	(0,02)	0,09**	(0,02)	0,09**	(0,02)	0,11**	(0,02)	0,09**	(0,02)
b	Testleistung Leseverständnis	–	–	0,25**	(0,02)	0,17**	(0,02)	0,16**	(0,02)	0,16**	(0,02)	0,16**	(0,02)	0,22**	(0,03)
c	Selbsteinschätzung Deutsch	–	–	–	–	0,17**	(0,02)	0,17**	(0,02)	0,17**	(0,02)	0,17**	(0,02)	0,17**	(0,02)
d	Geschlecht: weiblich	–	–	–	–	0,17**	(0,04)	0,10*	(0,04)	0,08	(0,04)	0,08	(0,04)	0,08	(0,04)
e	Sprachhintergrund: deutsch-russisch	–	–	–	–	–0,30**	(0,05)	–0,29**	(0,05)	–0,28**	(0,06)	–0,28**	(0,06)	–0,28**	(0,06)
f	Sprachhintergrund: deutsch-türkisch	–	–	–	–	–0,43**	(0,05)	–0,39**	(0,05)	–0,41**	(0,05)	–0,39**	(0,05)	–0,41**	(0,05)
g	Bildungsgang: Gymnasium	–	–	–	–	0,08	(0,05)	0,00	(0,05)	–0,01	(0,05)	–0,01	(0,05)	–0,01	(0,05)
h	Jahrgang: Klasse 9	–	–	–	–	–0,11*	(0,05)	–0,12*	(0,05)	–0,11*	(0,05)	–0,11*	(0,05)	–0,10*	(0,05)
Prädiktoren urteilende Personen															
i	Kognitive Fähigkeiten[a]	–	–	–	–	–	–	0,01**	(0,00)	0,01**	(0,00)	0,01**	(0,00)	0,01**	(0,00)
j	Geschlecht: weiblich	–	–	–	–	–	–	0,24**	(0,04)	0,23**	(0,04)	0,23**	(0,04)	0,23**	(0,04)
Prädiktoren Beziehungsmerkmale															
k	Beziehungsqualität	–	–	–	–	–	–	–	–	0,14**	(0,02)	0,14**	(0,02)	0,14**	(0,02)
l	Gemeinsamer Unterricht	–	–	–	–	–	–	–	–	–0,00	(0,04)	–0,00	(0,04)	–0,00	(0,04)
m	Geteilter Sprachhintergrund	–	–	–	–	–	–	–	–	0,03	(0,04)	0,03	(0,04)	0,03	(0,04)
Interaktionen[b]															
e Deutsch-russisch * a Testleistung Schreibaufgabe		–	–	–	–	–	–	–	–	–	–	–0,10*	(0,04)	–	–
j Urteiler Geschlecht * b Testleistung Leseverständnis		–	–	–	–	–	–	–	–	–	–	–	–	–0,08*	(0,04)

Tab. 1 (Fortsetzung)

	Modell D0		Modell D1		Modell D2		Modell D3		Modell D4		Modell D4a		Modell D4b	
	b	(se)	b	(se)	b	(se)	b	(se)	b	(se)	b	(se)	b	(se)
Varianzkomponenten [Varianzanteile]														
Urteiler in Schulen	0,2738	[28 %]	0,2605	[30 %]	0,2416	[31 %]	0,2218	[29 %]	0,2261	[31 %]	0,2248	[30 %]	0,2258	[31 %]
Beurteilte in Schulen	0,3440	[35 %]	0,2336	[27 %]	0,1489	[19 %]	0,1472	[19 %]	0,1441	[19 %]	0,1435	[19 %]	0,1446	[20 %]
Residuum	0,3640	[37 %]	0,3720	[43 %]	0,3871	[50 %]	0,3874	[51 %]	0,3688	[50 %]	0,3690	[50 %]	0,3675	[50 %]
Gütekriterien														
R^2 Marginal/Conditional	0,00/0,63		0,12/0,62		0,21/0,61		0,24/0,61		0,26/0,63		0,26/0,63		0,26/0,63	
Log-Likelihood	–3484,9		–3377,2**		–3282,9**		–3256,7**		–3216,6**		–3214,3*		–3213,8*	

Geringfügige Abweichungen von 100 % bei den Varianzanteilen [% der Gesamtvarianz] aufgrund von Rundungen möglich
Die Prädiktoren a, b, c und k wurden z-standardisiert
n (Einschätzungen) = 2662, *n* (Urteilende) = 1075, *n* (Beurteilte) = 1035
b Unstandardisierte Regressionsgewichte, *se* Standardfehler, *sd* Standardabweichung
[a]Jahrgangsspezifische T-Werte
[b]Es wurden Modelle mit je einem Interaktionsterm für die Prädiktoren d, e, f, i, j, k, l und m mit den Testleistungen a und b berechnet. In der Tabelle dargestellt sind nur diejenigen Modelle mit mindestens marginalen Effekten für die berechneten Interaktionen
* $p < 0{,}05$; ** $p < 0{,}01$

dem Modell *D4* weitere Modelle mit jeweils einer Interaktion für die genannten Prädiktoren mit den testbasierten Leistungsindikatoren im Schreiben und Lesen berechnet. Dabei zeigte sich, dass die negativ verzerrten Fremdeinschätzungen gegenüber Schüler(inne)n mit russischem Sprachhintergrund in Bezug auf die Testleistungen im Schreiben im Sinne der Rangkomponente zudem inakkurater ausfielen (Modell D4a: b = –0,10). Eine signifikante Moderation der Urteilsakkuratheit in Bezug auf die Testleistungen im Lesen ergab sich darüber hinaus für das Geschlecht der Urteilenden (Modell D4b: b = –0,08). Demnach gelang es Schülerinnen schlechter die Rangordnung der Leistungen ihrer Peers im Lesen einzuschätzen. Weder die kognitiven Fähigkeiten der Urteiler(innen), das Geschlecht der Beurteilten noch die Merkmale auf Beziehungsebene moderierten die Urteilsakkuratheit, weshalb auf eine Darstellung der Modelle mit diesen Interaktionseffekten in Tab. 1 verzichtet wurde. Einer Tendenz nach ließ sich eine positive Moderation der Urteilsakkuratheit in Bezug auf Testleistung im Schreiben durch einen gemeinsamen Unterrichtsbesuch annehmen (b = 0,06; $p = 0{,}085$). Da dies jedoch nur auf einem 10 %-Signifikanzniveau gegen den Zufall abgesichert werden konnte, wurde auf die Darstellung dieses Modells verzichtet.

6.2 Fremdeinschätzungen sprachlicher Fähigkeiten im Türkischen

Die Ergebnisse der Regressionsanalysen zur Vorhersage der Fremdeinschätzungen sprachlicher Fähigkeiten im Türkischen unter Berücksichtigung der Clusterung (siehe *T0*) nach urteilenden (ICC = 0,40) und beurteilten Schüler(innen) (ICC = 0,26) sind in Tab. 2 dargestellt. Die Testleistungen der eingeschätzten Peers waren sowohl einzeln (siehe Tab. 4) als auch gemeinsam (Modell *T1*) prädiktiv für das Schülerurteil sprachlicher Fähigkeiten, wobei höhere Testleistungen im Schreiben und Lesen mit einem Anstieg der Fremdeinschätzungen sprachlicher Fähigkeiten im Türkischen assoziiert waren. Im Modell *T2* zeigte sich, dass die Fremdeinschätzungen über objektive Testleistungen hinaus auch von den Selbsteinschätzungen der Peers vorhergesagt wurden (b = 0,26). Bei gleicher Leistung wurden Mädchen zudem um b = 0,19 Standardabweichungen positiver bewertet, wobei dieser Effekt mit Aufnahme der Merkmale der Urteilenden im Modell *T3* seine Signifikanz verlor und sich nun positivere Leistungsurteile für weibliche Urteilende ergaben. Ein signifikanter Haupteffekt zeigte sich ferner im Modell *T4* für einen geteilten Sprachhintergrund: Waren die Urteilenden selbst türkischsprachig, fielen die Fähigkeitsurteile um b = 0,45 Standardabweichungen geringer aus. Eine Berechnung von standardisierten Regressionskoeffizienten für das Modell *T4* ergab, dass der stärkste Einfluss auf die Fremdeinschätzungen sprachlicher Fähigkeiten von den Selbsteinschätzungen sprachlicher Fähigkeiten ($\beta = 0{,}27$) ausging, gefolgt von einem geteilten Sprachhintergrund ($\beta = -0{,}21$) und den erzielten Testleistungen der Peers im Lesen ($\beta = 0{,}12$) und Schreiben ($\beta = 0{,}10$).

Bei den Analysen zur Urteilsakkuratheit ergaben sich drei signifikante Moderationseffekte. Mit steigender Beziehungsqualität nahm der Zusammenhang zwischen den getesteten Fähigkeiten im Schreiben und den fremdeingeschätzten Fähigkeiten geringfügig zu (siehe *T4a*). Im Vergleich zu Urteilenden mit anderem Sprachhintergrund gelang es Urteilenden mit türkischem Sprachhintergrund deutlich besser,

Tab. 2 Kreuzklassifizierte Mehrebenenmodelle für Fremdeinschätzungen im Türkischen (z-standardisiert)

		Modell T0		Modell T1		Modell T2		Modell T3		Modell T4		Modell T4a		Modell T4b		Modell T4c	
		b	(se)	b	(se)	b	(se)	b	(se)	b	(se)	b	(se)	b	(se)	b	(se)
Fixe Parameter																	
Konstante		0,02	(0,04)	0,04	(0,04)	–0,01	(0,08)	–0,19	(0,17)	0,37*	(0,19)	0,37*	(0,19)	0,35	(0,19)	0,36	(0,19)
Prädiktoren eingeschätzte Personen																	
a	Testleistung Schreibaufgabe	–	–	0,20**	(0,04)	0,10*	(0,04)	0,10*	(0,04)	0,10**	(0,04)	0,10**	(0,04)	–0,05	(0,05)	0,11**	(0,04)
b	Testleistung Leseverständnis	–	–	0,19**	(0,04)	0,12**	(0,04)	0,12**	(0,04)	0,12**	(0,04)	0,12**	(0,04)	0,12**	(0,04)	0,03	(0,06)
c	Selbsteinschätzung Türkisch	–	–	–	–	0,26**	(0,04)	0,27**	(0,04)	0,27**	(0,03)	0,27**	(0,03)	0,27**	(0,03)	0,27**	(0,03)
d	Geschlecht: weiblich	–	–	–	–	0,19**	(0,07)	0,12	(0,07)	0,12	(0,07)	0,13	(0,07)	0,12	(0,07)	0,13	(0,07)
e	Bildungsgang: Gymnasium	–	–	–	–	–0,11	(0,07)	–0,13	(0,08)	–0,11	(0,08)	–0,11	(0,07)	–0,12	(0,07)	–0,11	(0,07)
f	Jahrgang: Klasse 9	–	–	–	–	–0,01	(0,07)	–0,01	(0,07)	0,00	(0,07)	–0,00	(0,07)	0,01	(0,07)	0,00	(0,07)
Prädiktoren urteilende Personen																	
g	Kognitive Fähigkeiten[a]	–	–	–	–	–	–	0,00	(0,00)	–0,00	(0,00)	–0,00	(0,00)	–0,00	(0,00)	–0,00	(0,00)
h	Geschlecht: weiblich	–	–	–	–	–	–	0,20*	(0,08)	0,17*	(0,07)	0,16*	(0,07)	0,18*	(0,07)	0,18*	(0,07)
Prädiktoren Beziehungsmerkmale																	
i	Beziehungsqualität	–	–	–	–	–	–	–	–	0,05	(0,03)	0,05	(0,03)	0,05	(0,03)	0,05	(0,03)
j	Gemeinsamer Unterricht	–	–	–	–	–	–	–	–	–0,08	(0,06)	–0,08	(0,06)	–0,08	(0,06)	–0,08	(0,06)
k	Geteilter Sprachhintergrund	–	–	–	–	–	–	–	–	–0,45**	(0,07)	–0,45**	(0,07)	–0,44**	(0,07)	–0,45**	(0,07)
Interaktionen[b]																	
i Beziehungsqualität * a Testleistung Schreibaufgabe		–	–	–	–	–	–	–	–	–	–	0,06*	(0,03)	–	–	–	–
k Geteilter Sprachhintergrund * a Testleistung Schreibaufgabe		–	–	–	–	–	–	–	–	–	–	–	–	0,23**	(0,06)	–	–
k Geteilter Sprachhintergrund * b Testleistung Leseverständnis		–	–	–	–	–	–	–	–	–	–	–	–	–	–	0,13*	(0,06)

Tab. 2 (Fortsetzung)

	Modell T0		Modell T1		Modell T2		Modell T3		Modell T4		Modell T4a		Modell T4b		Modell T4c	
	b	(se)	b	(se)	b	(se)	b	(se)	b	(se)	b	(se)	b	(se)	b	(se)
Varianzkomponenten [Varianzanteile]																
Urteiler in Schulen	0,3852	[40 %]	0,3873	[45 %]	0,3509	[44 %]	0,3428	[43 %]	0,2974	[40 %]	0,2982	[40 %]	0,2914	[40 %]	0,2956	[40 %]
Beurteilte in Schulen	0,2532	[26 %]	0,1455	[17 %]	0,0905	[11 %]	0,0915	[12 %]	0,0807	[11 %]	0,0791	[11 %]	0,0793	[11 %]	0,0789	[11 %]
Residuum	0,3359	[34 %]	0,3363	[39 %]	0,3556	[45 %]	0,3552	[45 %]	0,3654	[49 %]	0,3625	[49 %]	0,3609	[49 %]	0,3631	[49 %]
Gütekriterien																
R^2 Marginal/Conditional	0,00/0,66		0,12/0,66		0,18/0,63		0,19/0,64		0,23/0,62		0,24/0,63		0,25/0,63		0,24/0,63	
Log-Likelihood	–1300,0		–1256,0**		–1225,6**		–1222,1*		–1201,7**		–1198,9*		–1194,2**		–1199,4*	

Geringfügige Abweichungen von 100 % bei den Varianzanteilen [% der Gesamtvarianz] aufgrund von Rundungen möglich
Die Prädiktoren a, b, c und i wurden z-standardisiert
n (Einschätzungen) = 983, *n* (Urteilende) = 586, *n* (Beurteilte) = 373
b Unstandardisierte Regressionsgewichte, *se* Standardfehler, *sd* Standardabweichung
[a]Jahrgangsspezifische T-Werte
[b]Es wurden Modelle mit je einem Interaktionsterm für die Prädiktoren d, g, h, i, j und k mit den Testleistungen a und b berechnet. In der Tabelle dargestellt sind nur diejenigen Modelle mit signifikanten Interaktionseffekten
* $p < 0{,}05$; ** $p < 0{,}01$

die Leistungen der von ihnen eingeschätzten Peers in Bezug auf die Testleistungen im Schreiben (b = 0,23) und Lesen (b = 0,13) in die richtige Rangfolge zu bringen (siehe *T4b* und *T4c*). Weder die Merkmale der Urteiler(innen) noch das Geschlecht der eingeschätzten Peers oder ein gemeinsamer Unterrichtsbesuch moderierten die Akkuratheit der Fremdeinschätzungen im Türkischen.

Für alle Modelle im Türkischen ergaben sich signifikante Modellverbesserungen durch die aufgenommenen Prädiktoren. Die Testleistungen der eingeschätzten Peers erklärten 12 % der Varianz der Fremdeinschätzungen. Der beste Modellfit ergab sich für das Modell *T4c*, bei dem 25 % der Varianz der Fremdeinschätzungen durch die Prädiktoren erklärt wurden.

6.3 Fremdeinschätzungen sprachlicher Fähigkeiten im Russischen

Auch für das Russische wurden Regressionsmodelle mit zwei Zufallseffekten (random intercepts) berechnet, die Mittelwertunterschiede der Fremdeinschätzungen sprachlicher Fähigkeiten auf Ebene der urteilenden (ICC = 0,08) und beurteilten Schüler(inne)n (ICC = 0,31) modellieren (siehe Tab. 3). Die Schreib- und Lesefähigkeiten der eingeschätzten Mitschüler(innen) im Russischen waren sowohl für sich als auch gemeinsam prädiktiv für die Fremdeinschätzungen, wobei höhere Fähigkeiten mit höheren Leistungsurteilen assoziiert waren (siehe Modell *R1* und ergänzend Tab. 4). Mit Aufnahme der Selbsteinschätzungen sprachlicher Fähigkeiten im Modell *R2*, die die Fremdeinschätzungen signifikant vorhersagten, verloren sich jedoch die signifikanten Einflüsse der Testleistungen. In den Modellen *R3* und *R4* zeigte sich, dass weibliche Urteilende signifikant positivere Leistungsurteile fällten. Für die kognitiven Fähigkeiten der urteilenden Jugendlichen sowie die Beziehungsmerkmale ergaben sich keine signifikanten Haupteffekte.

Bei der Prüfung von Moderatoren der Urteilsakkuratheit ergab sich in Bezug auf die Testleistungen im Lesen ein positiver Effekt im Falle eines gemeinsamen Unterrichtsbesuchs (Modell *R4a*: b = 0,22). Dieses Modell wies zugleich mit 25 % erklärter Varianz den besten Modellfit auf. Nicht ausreichend gegen den Zufall abgesichert werden konnten drei weitere Interaktionsterme, die aus diesem Grund nicht in die Ergebnistabelle aufgenommen wurden. Bei diesen deutete sich an, dass die Fremdeinschätzungen im Russischen womöglich in Bezug auf die Testleistungen im Lesen von weiblichen Peers (b = 0,23; $p = 0{,}085$) sowie mit steigender Beziehungsqualität (b = 0,10; $p = 0{,}079$) und in Bezug auf die Testleistungen im Schreiben im Falle eines gemeinsamen Unterrichtsbesuchs (b = 0,20; $p = 0{,}087$) akkurater ausfallen könnten.

Die für das Modell *R4* zusätzlich berechneten standardisierten Regressionskoeffizienten ergaben, dass die Fremdeinschätzungen sprachlicher Fähigkeiten im Russischen am stärksten durch die Selbsteinschätzungen ($\beta = 0{,}34$) der Russischherkunftssprecher beeinflusst wurden, gefolgt vom Geschlecht der Urteilenden ($\beta = 0{,}15$).

Tab. 3 Kreuzklassifizierte Mehrebenenmodelle für Fremdeinschätzungen im Russischen (z-standardisiert)

		Modell R0		Modell R1		Modell R2		Modell R3		Modell R4		Modell R4a	
		b	(se)	b	(se)	b	(se)	b	(se)	b	(se)	b	(se)
Fixe Parameter													
Konstante		–0,05	(0,07)	–0,01	(0,06)	–0,03	(0,15)	–0,04	(0,31)	0,03	(0,32)	–0,04	(0,32)
Prädiktoren eingeschätzte Personen													
a	Testleistung Schreibaufgabe	–	–	0,24**	(0,08)	0,10	(0,08)	0,10	(0,08)	0,10	(0,08)	0,10	(0,08)
b	Testleistung Leseverständnis	–	–	0,19*	(0,08)	0,10	(0,07)	0,10	(0,07)	0,10	(0,07)	–0,05	(0,10)
c	Selbsteinschätzung Russisch	–	–	–	–	0,33**	(0,07)	0,33**	(0,07)	0,34**	(0,07)	0,34**	(0,07)
d	Geschlecht: weiblich	–	–	–	–	0,04	(0,13)	–0,10	(0,14)	–0,10	(0,14)	–0,10	(0,14)
e	Bildungsgang: Gymnasium	–	–	–	–	0,02	(0,11)	0,05	(0,12)	0,05	(0,12)	0,04	(0,12)
f	Jahrgang: Klasse 9	–	–	–	–	–0,02	(0,11)	–0,02	(0,11)	–0,01	(0,11)	–0,00	(0,01)
Prädiktoren urteilende Personen													
g	Kognitive Fähigkeiten[a]	–	–	–	–	–	–	–0,00	(0,01)	–0,00	(0,01)	–0,00	(0,01)
h	Geschlecht: weiblich	–	–	–	–	–	–	0,32**	(0,12)	0,32**	(0,12)	0,32**	(0,12)
Prädiktoren Beziehungsmerkmale													
i	Beziehungsqualität	–	–	–	–	–	–	–	–	0,02	(0,05)	0,02	(0,05)
j	Gemeinsamer Unterricht	–	–	–	–	–	–	–	–	–0,08	(0,12)	–0,09	(0,12)
k	Geteilter Sprachhintergrund	–	–	–	–	–	–	–	–	–0,07	(0,10)	–0,06	(0,10)
Interaktionen[b]													
j Gemeinsamer Unterricht * b Testleistung Leseverständnis		–	–	–	–	–	–	–	–	–	–	0,22*	(0,11)

Tab. 3 (Fortsetzung)

	Modell R0		Modell R1		Modell R2		Modell R3		Modell R4		Modell R4a	
	b	(se)	b	(se)	b	(se)	b	(se)	b	(se)	b	(se)
Varianzkomponenten [Varianzanteile]												
Urteiler in Schulen	0,0769	[8 %]	0,1273	[15 %]	0,1274	[16 %]	0,1017	[13 %]	0,1004	[13 %]	0,1009	[13 %]
Beurteilte in Schulen	0,3129	[31 %]	0,1343	[16 %]	0,0610	[8 %]	0,0608	[8 %]	0,0624	[8 %]	0,0596	[8 %]
Residuum	0,6106	[61 %]	0,5854	[69 %]	0,5916	[76 %]	0,5980	[79 %]	0,5959	[79 %]	0,5885	[79 %]
Gütekriterien												
R^2 Marginal/Conditional	0,00/0,39		0,16/0,42		0,22/0,41		0,24/0,40		0,24/0,41		0,25/0,41	
Log-Likelihood	–450,1		–431,9**		–421,1**		–417,3*		–416,9		–414,8*	

Geringfügige Abweichungen von 100 % bei den Varianzanteilen [% der Gesamtvarianz] aufgrund von Rundungen möglich
Die Prädiktoren a, b, c und i wurden z-standardisiert
n (Einschätzungen) = 327, *n* (Urteilende) = 239, *n* (Beurteilte) = 133
b Unstandardisierte Regressionsgewichte, *se* Standardfehler, *sd* Standardabweichung
[a]Jahrgangsspezifische T-Werte.
[b]Es wurden Modelle mit je einem Interaktionsterm für die Prädiktoren d, g, h, i, j und k mit den Testleistungen a und b berechnet. In der Tabelle dargestellt sind nur diejenigen Modelle mit signifikanten Interaktionseffekten
* $p < 0{,}05$; ** $p < 0{,}01$

7 Diskussion

Im vorliegenden Beitrag wurden Schülerurteile über die sprachlichen Fähigkeiten von Mitschüler(inne)n im Deutschen sowie den Herkunftssprachen Türkisch und Russisch mit dem Ziel in den Blick genommen, zu bewerten, ob und unter welchen Bedingungen sich diese als Indikatoren sprachlicher Fähigkeiten von Peers (z. B. im Rahmen egozentrierter Netzwerke) eignen könnten. Mit Hilfe solcher Indikatoren wäre es mit relativ geringem Aufwand möglich, die Qualität des sprachlichen Inputs durch relevante Interaktionspartner und deren Bedeutung für die individuelle Sprachentwicklung und den Bildungserfolg Jugendlicher weiter zu erforschen. Der Beitrag fokussierte vor diesem Hintergrund erstens auf Determinanten der Fremdeinschätzungen der sprachlichen Fähigkeiten von Mitschüler(inne)n, wobei es um die Prüfung von Merkmalen der eingeschätzten Peers ging, die die Fremdeinschätzungen über Testleistungen hinaus beeinflussen könnten (Fragestellung I). Zweitens wurde die Rangkomponente in Bezug auf Testleistungen im Lesen und Schreiben herangezogen, um die Frage nach der Urteilsakkuratheit der Fremdeinschätzungen zu klären (Fragestellung II). Und schließlich wurden Merkmale der urteilenden und beurteilten Schüler(innen) sowie deren Beziehung als Moderatoren der Urteilsakkuratheit untersucht (Fragestellung III).

7.1 Fragestellung I: Einflussfaktoren auf die Fremdeinschätzungen unter Kontrolle der objektiven Leistungsindikatoren

Für das Deutsche wurden leistungsbezogene Vorurteile gegenüber Schüler(inne)n sprachlicher Minderheiten (*H1*) erwartet (vgl. Zander et al. 2014). In der Tat zeigte sich, dass gegenüber monolingual deutsch aufgewachsenen Schüler(inne)n Peers mit anderer Herkunftssprache signifikant geringere Leistungsurteile erhielten, wobei diejenigen mit türkischem Sprachhintergrund am schlechtesten bewertet wurden. Dieser Effekt entspricht in seiner Richtung den möglicherweise aus den Medien bekannten Befunden den PISA-Studien (vgl. Stanat et al. 2010) und belegt nach Sprachhintergrund variierende leistungsbezogene Vorurteile bei den Jugendlichen.

Die Erwartung, dass die sprachlichen Fähigkeiten von Mädchen im Vergleich zu Jungen höher bewertet werden (wie z. B. bei Hinnant et al. 2009) bestätigte sich nicht (*H2*): Während sich im Russischen von vornherein kein Geschlechtereffekt zeigte, waren im Deutschen und Türkischen zunächst positivere Leistungsurteile bei der Einschätzung von Mädchen zu beobachten, wobei sich der Effekt unter Kontrolle von Urteiler- und Beziehungsmerkmalen verlor. Dagegen ergab sich in allen Sprachen, ein systematischer Einfluss des Geschlechts der Urteilenden, wonach Mädchen positivere Leistungsurteile vergaben. Vermutlich war der Effekt des Geschlechts der eingeschätzten Mitschüler(innen) mit dem Geschlechtereffekt der Urteilenden einerseits und der Beziehungsqualität andererseits konfundiert, denn Mädchen gaben zudem im Mittel eine höhere Beziehungsqualität zu den von ihnen gewählten (und zu über 78 % weiblichen) Peers an.

Vermutet wurde ferner, dass sich Schüler(innen) bei der Fremdeinschätzung an leicht beobachtbaren Verhaltensweisen orientieren, die mit motivationalen oder affektiven Merkmalen zusammenhängen (z. B. Rakoczy et al. 2008; Urhahne et al.

2010; Kaiser et al. 2013). Da keine direkten Indikatoren für solche Verhaltensweisen vorlagen, wurden die Selbsteinschätzungen sprachlicher Fähigkeiten der beurteilten Peers einbezogen, von denen erwartet werden kann, dass sie sich in Verhaltensweisen manifestieren (wie z. B. Unterrichtsbeteiligung, Gebrauch der Herkunftssprache), die sich leicht in Bezug auf sprachliche Kompetenzen interpretieren lassen. Gemäß der Erwartung ergaben sich über objektive Leistungsindikatoren hinaus Effekte der Selbsteinschätzungen sprachlicher Fähigkeiten der Peers auf die Fremdeinschätzungen (*H3*). Eine Limitation der Studie besteht allerdings darin, dass die Überprüfung der Annahme eines indirekten Zusammenhangs der Selbsteinschätzungen über beobachtbares Verhalten auf die Fremdeinschätzungen aufgrund der fehlenden objektiven Verhaltensindikatoren nicht möglich war. In allen untersuchten Sprachen war der Einfluss der Selbsteinschätzungen auf die Fremdeinschätzungen sogar stärker als der der Testleistungen. Dies galt insbesondere für das Russische, für das nach Kontrolle der Selbsteinschätzungen kein zusätzlicher Effekt der Testleistungen auf die Fremdeinschätzungen mehr zu finden war. Zu berücksichtigen ist jedoch, dass die Selbst- und Fremdeinschätzungen auf derselben Skala und zum Teil mit Hilfe derselben Items gemessen wurden. Der stärkere Zusammenhang zwischen diesen Indikatoren ist somit auch auf die höhere inhaltliche wie konzeptionelle Kongruenz der Maße zurückzuführen. Dafür sprechen die höheren Kriteriums-Urteils-Korrelationen bei Lehrerurteilen, wenn dieses und der zum Vergleich herangezogene Leistungstest auf derselben Metrik verortet ist und dieselbe Leistungsdomäne abbilden (z. B. Hoge und Coladarci 1989; Südkamp et al. 2012).

7.2 Fragestellung II: Urteilsakkuratheit der Fremdeinschätzungen

Hypothesenkonform zeigte sich eine signifikante, aber moderate Vorhersage der Fremdeinschätzungen sprachlicher Fähigkeiten durch die objektiven Testleistungen der eingeschätzten Peers (*H4*). Die Testleistungen im Lesen und Schreiben erklärten gemeinsam zwischen 12 und 16 % der Varianz der Fremdeinschätzungen im Deutschen, Türkischen und Russischen. Für das Leseverständnis allein lagen die Varianzaufklärungen zwischen 8 und 12 %, für Testleistungen im Schreiben zwischen 6 und 13 % (siehe Tab. 4 im Anhang). Als Vergleichsgröße zur Beurteilung der Urteilsakkuratheit können die in anderen Studien gefundenen Korrelationen zwischen Leistungsurteilen und Leistungstests quadriert herangezogen werden, die dem Anteil erklärter Varianz entsprechen. Für die Urteilsakkuratheit von Leistungsbeurteilungen durch Lehrkräfte ergaben sich in der Metaanalyse von Südkamp et al. (2012) durchschnittlich 28 % erklärter Varianz (bei r = 0,53), was deutlich höher ausfällt als in der vorliegenden Studie. Auch Pohlmann et al. (2004) fanden mit einem R-Quadrat von 0,21 höhere Varianzaufklärungen der durch Mitschüler(innen) eingeschätzten Schülerselbstkonzepte im Deutschen durch die Deutschnote. Vergleichbar scheinen die im vorliegenden Beitrag ermittelten Varianzaufklärungen mit Zusammenhängen zwischen Selbstberichten sprachlicher Fähigkeiten und Testleistungen: Edele et al. (2015) berichten Varianzaufklärungen von 5–12 % für die Selbsteinschätzungen im Deutschen durch Schülerleistungen im Leseverstehen sowie 6–19 % bei Selbsteinschätzungen in den Herkunftssprachen Türkisch und Russisch durch einen Hörverständnistests.

Obwohl sich die Fremdeinschätzungen und verwendeten Leistungsindikatoren inhaltlich auf kongruente Konstrukte bezogen, fallen die Zusammenhänge eher moderat bis gering aus. Analog zu Selbsteinschätzungen sprachlicher Fähigkeiten lässt sich für Fremdeinschätzungen vermuten, dass diese als Prädiktoren für individuelle sprachliche Fähigkeiten von Schüler(innen) womöglich plausible Ergebnisse erzeugen, aber die Effekte sprachlicher Fähigkeiten von Peers aufgrund der Ungenauigkeit der Maße potentiell unterschätzt werden könnten (vgl. Edele et al. 2015). Vergleiche zwischen Modellen in denen der Einfluss sprachlicher Fähigkeiten von Peers mittels Fremdeinschätzungen gemessen wird und solchen, in denen testbasierte Leistungsinformationen verwendet werden, könnten hier weitere Erkenntnisse liefern.

7.3 Fragestellung III: Moderatoren der Urteilsakkuratheit

Als Moderatoren der Urteilsakkuratheit wurden Merkmale der eingeschätzten Peers (Sprachhintergrund und Geschlecht), Merkmale der urteilenden Schüler(innen) (kognitive Fähigkeiten und Geschlecht) sowie Merkmale auf Beziehungsebene (Beziehungsqualität, gemeinsamer Unterricht, geteilter Sprachhintergrund) geprüft. In Bezug auf Fremdeinschätzungen im Deutschen hatten sich negativ verzerrte Leistungsurteile gegenüber Schüler(inne)n mit nicht-deutscher Herkunftssprache gezeigt. In Bezug auf die Testleistungen von russischsprachigen Peers im Schreiben ergab sich nun, dass diese auch im Sinne der Rangkomponente inakkurater ausfielen.

Sowohl im Deutschen als auch in den Herkunftssprachen Türkisch und Russisch vergaben Mädchen höhere Fremdeinschätzungen. Eine Moderation der Urteilsakkuratheit durch das Geschlecht der Urteilenden ließ sich jedoch lediglich für das Deutsche nachweisen, wonach die überhöhten Fähigkeitsurteile von Mädchen in Bezug auf das Leseverständnis im Sinne der Rangkomponente inakkurater ausfielen. Das Geschlecht der *beurteilten* Mitschüler(innen) moderierte die Akkuratheit der Fremdeinschätzungen dagegen in keiner der untersuchten Sprachen. Die aus Theorie und Forschung abgeleitete Vermutung, dass die kognitiven Fähigkeiten der Urteilenden die Urteilsakkuratheit positiv moderieren könnte (z. B. Brantmeier und Vanderplank 2008; Südkamp et al. 2012; Kaiser et al. 2012), konnte weder für das Deutsche noch die Herkunftssprachen bestätigt werden (*H5*).

Die Analysen lieferten verschiedene Anhaltspunkte für die Annahme, dass die Urteilsakkuratheit von der Beziehung zur eingeschätzten Person moderiert wird (z. B. Marsh und Craven 1991; Funder 1995; Hayes und Dunning 1997). Erwartet wurden akkuratere Fremdeinschätzungen sprachlicher Fähigkeiten mit steigender Beziehungsqualität sowie im Falle eines gemeinsamen Unterrichtsbesuchs (*H6*). Im Russischen wurde die Rangkomponente in Bezug auf die Leseleistungen positiv von einem gemeinsamen Unterrichtsbesuch moderiert. Möglicherweise gilt dies auch in Bezug auf die Schreibfähigkeiten von Peers im Russischen sowie im Deutschen, allerdings konnten die gefunden Interaktionseffekte hier nur auf einem 10 %-Signifikanzniveau gegen den Zufall abgesichert werden. Eine positive Moderation der Urteilsakkuratheit mit steigender Beziehungsqualität ergab sich für Fremdeinschätzungen in Bezug auf die Testleistungen von Peers im Schreiben im Türkischen und in der Tendenz auch im Russischen (bei $p = 0{,}079$). Eine positive Moderation der Ur-

teilsakkuratheit durch das Vorhandensein eines identischen Sprachhintergrunds (*H7*) ergab sich im Türkischen sowohl in Bezug auf die Testleistungen von Peers im Lesen als auch Schreiben. Dabei zeigte sich, dass die Leistungsurteile von Schüler(innen) ohne türkischen Sprachhintergrund tendenziell zu optimistisch ausfielen, während diese bei türkischsprachigen Urteiler(inne)n strenger, aber zugleich akkurater waren. Im Russischen blieb ein solcher Moderationseffekt erwartungswidrig aus, was daran liegen könnte, dass in der Studie auch Schüler(innen) (mit und ohne russischem Sprachhintergrund) vertreten waren, die Russisch als Fremdsprache lernten. Dies könnte erklären, warum sich Effekte des gemeinsamen Unterrichtsbesuchs bei den Fremdeinschätzungen im Russischen deutlicher zeigten als im Türkischen. Sollte dies zutreffen, sprächen die gefundenen Effekte in den Herkunftssprachen beide dafür, dass die Einschätzungen besser gelingen, wenn die Urteilenden selbst Kenntnisse in den betreffenden Sprachen haben. Allerdings kann mit den vorliegenden Daten nicht geklärt werden, ob der Effekt im Russischen tatsächlich auf einen gemeinsamen Fremdsprachenunterricht zurückzuführen ist.

7.4 Fazit

Über Fremdeinschätzungen der sprachlichen Fähigkeiten unter Schüler(inne)n war bisher kaum etwas bekannt. Im Hinblick auf die Einschätzungen herkunftssprachlicher Fähigkeiten von Peers lagen gar keine Befunde vor. Im vorliegenden Beitrag konnten systematische und signifikant positive Zusammenhänge zwischen fremdeingeschätzten Sprachfähigkeiten und den gemessenen Testleistungen von Jugendlichen anhand zweier Leistungsindikatoren im Deutschen, Russischen und Türkischen nachgewiesen werden. Die eingangs formulierte Frage, ob in Ermangelung objektiver Maße Fremdeinschätzungen als Proxys für sprachliche Fähigkeiten von Peers eingesetzt werden könnten, muss auf Grundlage der vorliegenden Daten jedoch abgelehnt werden. Dies ist mit der relativ geringen Urteilsakkuratheit im Sinne der Rangkomponente einerseits und andererseits den systematischen Verzerrungen der Urteile nach Merkmalen der eingeschätzten Peers, dem Geschlecht der Urteilenden und Beziehungsmerkmalen zu begründen. Für die Frage nach der Rolle der Qualität des sprachlichen Inputs von Peers für die individuelle Entwicklung sprachlicher Fähigkeiten bedeutet dies, dass entweder objektive Leistungsmaße für die Peers erfasst werden sollten oder andere Proxis für den sprachlichen Input herangezogen werden müssen (z. B. über die Erfassung der Sprachnutzung oder sprachbezogener Aktivitäten mit Peers).

Besonders bedenklich erscheint der Befund, dass bei gleicher Leistung die sprachlichen Fähigkeiten im Deutschen von herkunftssprachlich russischen sowie türkischen Schüler(inne)n signifikant schlechter bewertet wurden. Dies galt unabhängig von der Qualität der Beziehung zwischen urteilender und beurteilter Person und trotz der Tatsache, dass die Teilnehmenden überwiegend bereits in Deutschland geboren wurden und alle seit mindestens der dritten Klasse eine Schule in Deutschland besuchten (siehe Gogolin et al. 2017). Solche Vorurteile können negative Folgen für die Leistungsentwicklungen von Jugendlichen mit nicht-deutscher Herkunftssprache haben, da leistungsbezogene Selbstkonzepte wesentlich durch Rückmeldungen signifikanter Anderer geformt werden (vgl. Marsh 1990). Dies belegen Studien,

die eine Leistungsbeeinträchtigungen durch Stereotypenbedrohungen für verschiedene ethnische Minderheiten insbesondere in sprachlichen Domänen nachgewiesen haben (siehe Nadler und Clark 2011 für eine Metaanalyse). Die Befunde sprechen deutlich dafür, dass Anstrengungen unternommen werden sollten, die negativen und unzutreffenden leistungsbezogenen Stereotype gegenüber lebensweltlich mehrsprachigen Schüler(inne)n abzubauen. In diesem Zusammenhang wären Studien wünschenswert, die Wirkungszusammenhänge zwischen Fremdeinschätzungen und Selbstkonzepten sprachlicher Fähigkeiten sowie die Konsequenzen von negativ verzerrten Leistungsurteilen gegenüber Schüler(inne)n mit sprachlichem Migrationshintergrund weiter fokussieren (wie z. B. bei Sander et al. 2017).

Die vorliegenden Ergebnisse deuten insgesamt auf eine heuristische Informationsverarbeitung bei der Beurteilung sprachlicher Fähigkeiten von Peers hin (siehe Herppich et al. 2018), wobei im Sinne eines Halo-Effekts (Nisbett und Wilson 1977) der Gesamteindruck zu einer Person sowie Leistungsstereotype eingehen. Welche Prozesse genau bei der Urteilsbildung abliefen, kann mit den vorliegenden Daten jedoch nicht geklärt werden. Interessant wären hier Studiendesigns, die weitere Merkmale der urteilenden Schüler(innen), wie z. B. motivationale Aspekte der Leistungsbeurteilung und Einstellungen berücksichtigen. Für ein besseres Verständnis darüber, welche Informationen bei der Bewertung sprachlicher Fähigkeiten von Peers genutzt werden, könnten ergänzend Einflüsse von Merkmalen der Peers, wie sprachprosodische Merkmale (z. B. der Akzent, siehe Anderson et al. 2007), motivationale Merkmale (z. B. Unterrichtsbeteiligung) sowie andere beobachtbare Verhaltensweisen (z. B. Leseverhalten, Sprachnutzung) untersucht werden.

Eine Limitation des Studiendesigns betrifft die geringen Fallzahlen innerhalb der Gruppierungsebene der urteilenden Schüler(innen). In Studien zur Urteilsakkuratheit von Lehrkräften werden Rangkorrelationen für die Urteile in der Regel auf Basis der Einschätzungen aller Schüler(innen) einer Klasse erfasst. In unserem Falle konnten aus befragungsökonomischen Gründen maximal drei Peers eingeschätzt werden, was sowohl die Ermittlung von Moderationseffekten einschränkt als auch dazu führen kann, dass sich Fehleinschätzungen stärker auf die ermittelten Regressionskoeffizienten auswirken. In Bezug auf die Beziehungsmerkmale konnten erwartungsgemäße Hinweise darauf gefunden werden, dass Gelegenheiten, die Eigenschaften von Peers zu beobachten, sich positiv auf die Urteilsakkuratheit auswirken (Funder 1995). Zugleich variierten diese je nach Sprache und dem für die Bewertung der Urteilsakkuratheit herangezogenen Leistungskriterium (Lesen, Schreiben). Zur weiteren Klärung könnte z. B. berücksichtigt werden, wie sich der Besuch eines gemeinsamen herkunftssprachlichen (ggf. auch fremdsprachlichen) Unterrichts oder auch die gemeinsame Nutzung der Herkunftssprache auf die Urteilsakkuratheit auswirkt. Zudem müssten Effekte der Beziehungsqualität mit größeren Stichproben untersucht werden, bei denen Urteiler(innen) mehr als nur drei Peers einschätzen. Es lässt sich vermuten, dass die Jugendlichen ohnehin ihre engsten Peers wählten (vgl. Dijkstra et al. 2008), sodass Unterschiede nach der Beziehungsqualität in der vorliegenden Studie nicht ausreichend gegeben waren.

Mögliche Verbesserungen schließen auch die Fremdeinschätzungen selbst ein. Diese stießen als Notenskala bei höheren Fähigkeitsurteilen an ihre Grenzen. Eine Erfassungsmethode, bei der Urteil und Kriterium auf derselben Skala verortet

sind, hätte den Vorteil, dass über die Rangkomponente hinaus auch Aussagen über die Urteilsakkuratheit hinsichtlich des Leistungsniveaus sowie die Heterogenität der Schülerleistungen möglich wären (vgl. Karst 2017). Neuere Erkenntnisse zur Akkuratheit von Lehrerurteilen legen nahe, dass globalere Urteile geeignet sind, um die Rangfolge von Schülerleistungen einzuschätzen, wogegen die Beurteilung des Niveaus und der Streuung von Schülerleistungen differenziertere Maße erfordern (vgl. Karst et al. 2018). Ferner hätte es sinnvoll sein können, die Schüler(innen) darüber zu informieren, dass ihre Einschätzungen mit den erhobenen Tests im Lesen und Schreiben verglichen werden. Eine entsprechende Informiertheit moderierte nach Südkamp et al. (2012) die Korrelationen zwischen Urteil und Kriterium positiv.

Trotz der benannten Limitationen konnten mit der durchgeführten Studie wichtige Erkenntnisse gewonnen und die Möglichkeiten und Grenzen von Fremdeinschätzungen zur Erhebung sprachlicher Fähigkeiten von Peers im Rahmen klassischer (egozentrierter) Fragebogenerhebungen aufgezeigt werden. Einerseits deuten die Befunde auf Bedingungen, unter denen akkuratere Fremdeinschätzungen unter Schüler(innen) gelingen könnten (z. B. durch Urteile innerhalb von Klassengrenzen), andererseits belegen die Ergebnisse deutliche Grenzen von Fremdeinschätzungen als ökonomisches Verfahren zur Erfassung sprachlicher Fähigkeiten. So ist von der Nutzung dieser als Indikatoren tatsächlicher Kompetenzen von Peers in der hier untersuchten Form abzuraten. Nichtsdestotrotz erscheint die Beschäftigung mit Wahrnehmungsprozessen sprachlicher Fähigkeiten unter Schüler(inne)n auch unabhängig von der Frage nach ihrer Validität wichtig. Stereotype Überzeugungen können sich im Handeln manifestieren und damit Lernprozesse negativ beeinflussen (siehe Karst und Bonefeld 2020). So können beispielsweise Leistungserwartungen von Lehrkräften im Fach Deutsch bei gleicher Leistung gegenüber Schüler(innen) ethnischer Minderheiten negativ verzerrt sein (z. B. Lorenz et al. 2016), wobei das Ausmaß negativ verzerrter Leistungserwartungen den Umfang von Leistungsdisparitäten zwischen Schüler(inne)n mit und ohne Migrationshintergrund erklären kann (van den Bergh et al. 2010). Nun bestätigen die vorliegenden Ergebnisse, dass solche negativ verzerrten Leistungserwartungen auch unter (befreundeten) Schüler(inne)n vorzufinden sind (siehe Zander et al. 2014). Vor dem Hintergrund der ungleichen Bildungserfolgschancen von Schüler(inne)n mit und ohne Migrationshintergrund (siehe Stanat et al. 2010; Kempert et al. 2016) sollten leistungsbezogene Vorurteile insbesondere in sprachlichen Domänen ernst genommen werden und Peers als wichtiger sozialer Faktor für das Lernen auch im Hinblick auf negative Folgen von Stigmatisierungs- und Zuschreibungsprozessen weiter untersucht werden.

Funding Open Access funding enabled and organized by Projekt DEAL.

Anhang

Tab. 4 Varianzaufklärung der Fremdeinschätzungen durch die Testleistungen

		Konfidenzintervall (b)			
	b	2,5 %	97,5 %	R^2 marginal	Log-Likelihood
Deutsch					
Schreibaufgabe	0,25**	0,20	0,29	0,06	–3434,3**
Leseverständnis	0,30**	0,25	0,34	0,09	–3405,5**
Türkisch					
Schreibaufgabe	0,30**	0,23	0,37	0,09	–1267,5**
Leseverständnis	0,28**	0,21	0,35	0,08	–1269,0**
Russisch					
Schreibaufgabe	0,37**	0,24	0,49	0,13	–435,0**
Leseverständnis	0,34**	0,22	0,47	0,12	–436,5**

Unstandardisierte Regressionskoeffizienten für die Modelle mit je nur einem Prädiktor für Testleistungen der eingeschätzten Peers
Likelihood-Ratio-Test zum Vergleich der Modelle zum Nullmodell (siehe Tab. 1, 2 und 3: D0; T0; R0)
** $p < 0{,}01$

Literatur

Alderson, C. A. (2005). *Diagnosing foreign language proficiency: the interface between learning and assessment*. New York: Continuum.

Anderson, S., Downs, S. D., Faucette, K., Griffin, J., King, T., & Woolstenhulme, S. (2007). How accents affect perceptions of intelligence, physical attractiveness, and trustworthiness of Middle-Eastern-, Latin-American-, British- and Standard-American-English-Accent speakers. *BYU Undergraduate Journal of Psychology*, *3*(1), 5–11.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

van den Bergh, L., Denessen, E., Hornstra, L., Voeten, M., & Holland, R. W. (2010). The implicit prejudiced attitudes of teachers. *American Educational Research Journal*, *47*(2), 497–527.

Borkenau, P., & Liebler, A. (1993). Convergence of stranger ratings of personality and intelligence with self-ratings, partner-ratings, and measured intelligence. *Journal of Personality and Social Psychology*, *63*(3), 546–553.

Brantmeier, C., & Vanderplank, R. (2008). Descriptive and criterion-referenced self-assessment with L2 readers. *System*, *36*(3), 456–477.

Bratko, D., Chamarro-Premuzic, T., & Saks, Z. (2006). Personality and school performance: incremental validity of self- and peer ratings over intelligence. *Personality and Individual Differences*, *41*, 131–142.

Breidebach, G., & Gruber, N. (2018). Teachers' diagnostic competence in the context of gender and migration related stereotyping. *Journal of Teacher Education and Educators*, *7*(1), 5–16.

Brown, G. T. L., Andrade, H. L., & Chen, F. (2015). Accuracy in student self-assessment: directions and cautions for research. *Assessment in Education: Principles, Policy & Practice*, *22*(4), 444–457.

Clovin, C. R., & Funder, D. C. (1991). Predicting personality and behavior: a boundary on the acquaintanceship effect. *Journal of Personality and Social Psychology*, *60*(6), 884–894.

Connelly, B. S., & Ones, D. S. (2010). An other perspective on personality. Meta-analytic integration of observers' accuracy and predictive validity. *Psychological bulletin*, *136*(6), 1092–1122.

Dijkstra, P., Kuyper, H., van der Werf, G., Buunk, A. P., & van der Zee, Y. G. (2008). Social comparison in the classroom: a review. *Review of Educational Research*, *78*(4), 828–879.

Dirim, I., & Döll, M. (2009). „Bumerang" – Erfassung der Sprachkompetenzen im Übergang von der Schule in den Beruf – vergleichende Beobachtungen zum Türkischen und Deutschen am Beispiel einer Schülerin. In D. Lengyel (Hrsg.), *Von der Sprachdiagnose zur Sprachförderung* (S. 139–146). Münster: Waxmann.

Dünnebier, K., Gräsel, C., & Krolak-Schwerdt, S. (2009). Urteilsverzerrungen in der schulischen Leistungsbeurteilung: Eine experimentelle Studie zu Ankereffekten. *Zeitschrift für Pädagogische Psychologie*, *23*, 187–195.

Edele, A., Seuring, J., Kristen, C., & Stanat, P. (2015). Why bother with testing? The validity of immigrants' self-assessed language proficiency. *Social Science Research*, *52*, 99–123.

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

Fiske, S. T., & Neuberg, S. L. A. (1990). Continuum of impression formation, from category-based to individuating processes: influences of information and motivation on attention and interpretation. *Advances in Experimental Social Psychology*, *23*, 1–74.

Förster, N., & Böhmer, I. (2017). Das Linsenmodell – Grundlagen und exemplarische Anwendungen in der pädagogisch-psychologischen Diagnostik. In A. Südkamp & A.-K. Praetorius (Hrsg.), *Diagnostische Kompetenz von Lehrkräften. Theoretische und methodische Weiterentwicklungen* (Pädagogische Psychologie und Entwicklungspsychologie, Bd. 94, S. 46–50). Münster: Waxmann.

Funder, D. C. (1995). On the accuracy of personality judgement: a realistic approach. *Psychological Review*, *102*(4), 652–670.

Funder, D. C., & Dobroth, K. M. (1987). Differences between traits: properties associated with interjudge agreement. *Journal of Personality and Social Psychology*, *52*, 409–418.

Glock, S., Krolak-Schwerdt, S., & Cate, P. I. M. (2015). Are school placement recommendations accurate? The effect of students' ethnicity on teachers' judgments and recognition memory. *European Journal of Psychology of Education*, *30*(2), 169–188.

Gogolin, I. (2009). Zweisprachigkeit und die Entwicklung bildungssprachlicher Fähigkeiten. In I. Gogolin & U. Neumann (Hrsg.), *Streitfall Zweisprachigkeit* (S. 263–280). Wiesbaden: VS.

Gogolin, I., Klinger, T., Lagemann, M., & Schnoor, B. (2017). Indikation, Konzeption und Untersuchungsdesign des Projekts Mehrsprachigkeitsentwicklung im Zeitverlauf (MEZ*)* (MEZ Arbeitspapiere 1). Hamburg: Universität Hamburg. http://www.pedocs.de/frontdoor.php?source_opus=14825. Zugegriffen: 30. Sep. 2020.

Hannover, B., & Zander, L. (2016). Die Bedeutung der Peers für die individuelle schulische Entwicklung. In J. Möller, M. Köller & T. Riecke-Baulecke (Hrsg.), *Basiswissen Lehrerbildung. Schule und Unterricht, Lehren und Lernen* (1. Aufl., S. 91–105). Seelze: Friedrich Verlag GmbH.

Hayes, A. F., & Dunning, D. (1997). Construal processes and trait ambiguity: implications for the self-peer agreement in personality judgment. *Journal of Personality and Social Psychology*, *72*, 664–677.

Heller, K. A., & Perleth, C. (2000). *KFT 4–12+ R. Kognitiver Fähigkeitstest für 4. bis 12. Klassen, Revision* (3. Aufl.). Göttingen: Beltz.

Herppich, S., Praetorius, A.-K., Förster, N., Glogger-Frey, I., Karst, K., Leutner, D., Behrmann, L., Böhmer, M., Ufer, S., Klug, J., Hetmanek, A., Ohle, A., Böhmer, I., Karing, C., Kaiser, J., & Südkamp, A. (2018). Teachers' assessment competence: Integrating knowledge-, process-, and product-oriented approaches into a competence-oriented conceptual model. *Journal of Teaching and Teacher Education*, *76*, 181–193.

Herz, A. (2012). Erhebung und Analyse egozentrierter Netzwerke. In S. Kulin, K. Frank, D. Fickermann & K. Schwippert (Hrsg.), *Soziale Netzwerkanalyse: Theorie, Methoden, Praxis* (Netzwerke im Bildungsbereich, Bd. 5, S. 133–150). Münster: Waxmann.

Heyer, R., Palentien, C., & Gürlevik, A. (2012). Peers. In U. Bauer, U. H. Bittlingmayer & A. Scherr (Hrsg.), *Handbuch Bildungs- und Erziehungssoziologie* (S. 983–999). Wiesbaden: Springer VS.

Hinnant, J. O., O'Brien, M., & Ghazarian, S. R. (2009). The longitudinal relations of teacher expectations to achievement in the early school years. *Journal of Educational Psychology*, *101*(3), 662–670.

Hoge, R. D., & Coladarci, T. (1989). Teacher-based judgments of academic achievement: a review of literature. *Review of Educational Research*, *59*(5), 297–313.

Hox, J. J., Moerbeek, M., & Van de Schoot, R. (2018). *Multilevel analysis. Techniques and applications* (3. Aufl.). New York: Routledge.

Kaiser, J., Helm, F., Retelsdorf, J., Südkamp, A., & Möller, J. (2012). Zum Zusammenhang von Intelligenz und Urteilsgenauigkeit bei der Beurteilung von Schülerleistungen im Simulierten Klassenraum. *Zeitschrift für Pädagogische Psychologie*, *26*(4), 251–261.

Kaiser, J., Retelsdorf, J., Südkamp, A., & Möller, J. (2013). Achievement and engagement: How student characteristics influence teacher judgments. *Learning and Instruction*, *28*, 73–84.

Kaiser, J., Südkamp, A., & Möller, J. (2017). The effects of student characteristics on teachers' judgment accuracy: disentangling ethnicity, minority status, and achievement. *Journal of Educational Psychology*, *109*(6), 871–888.

Karing, C., & Artelt, C. (2013). Genauigkeit von Lehrpersonenurteilen und Ansatzpunkte ihrer Förderung in der Aus- und Weiterbildung von Lehrkräften. *Beiträge zur Lehrerinnen- und Lehrerbildung*, *31*(2), 166–173.

Karing, C., Matthäi, J., & Artelt, C. (2011). Genauigkeit von Lehrerurteilen über die Lesekompetenz ihrer Schülerinnen und Schüler in der Sekundarstufe I – Eine Frage der Spezifität? *Zeitschrift für Pädagogische Psychologie*, *25*(3), 159–172.

Karst, K. (2017). Akkurate Urteile – die Ansätze von Schrader (1989) und McElvany et al. (2009). In A. Südkamp & A.-K. Praetorius (Hrsg.), *Diagnostische Kompetenz von Lehrkräften. Theoretische und methodische Weiterentwicklungen* (Pädagogische Psychologie und Entwicklungspsychologie, Bd. 94, S. 21–29). Münster: Waxmann.

Karst, K., & Bonefeld, M. (2020). Stereotype, Urteile und Urteilsakkuratheit von Lehrkräften: Eine Zusammenfassung im Rahmen des Heterogenitätsdiskurses. In S. Glock & H. Kleen (Hrsg.), *Stereotype in der Schule* (S. 281–308). Wiesbaden: Springer VS.

Karst, K., Dotzel, S., & Dickhäuser, O. (2018). Comparing global judgments and specific judgments of teachers about students' knowledge. Is the whole the sum of its parts? *Teaching and teacher education*, *76*, 194–203.

Karst, K., Hartig, J., Kaiser, J., & Lipowsky, F. (2017). Mehrebenenmodelle als Werkzeuge zur Analyse diagnostischer Kompetenz von Lehrkräften – ein lineares Modell (LMM) und seine Anwendung in R. In A. Südkamp & A.-K. Praetorius (Hrsg.), *Diagnostische Kompetenz von Lehrkräften. Theoretische und methodische Weiterentwicklungen* (Pädagogische Psychologie und Entwicklungspsychologie, Bd. 94, S. 153–174). Münster: Waxmann.

Kempert, S., Edele, A., Rauch, D., Wolf, K. M., Paetsch, J., Darsow, A., Maluch, J., & Stanat, P. (2016). Die Rolle der Sprache für zuwanderungsbezogene Ungleichheiten im Bildungserfolg. In C. Diehl,

C. Hunkler & C. Kristen (Hrsg.), *Ethnische Ungleichheiten im Bildungsverlauf* (S. 157–242). Wiesbaden: Springer VS.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personal and Social Psychology*, *77*(6), 1121–1134.

Kuhl, P., & Hannover, B. (2012). Differenzielle Benotungen von Mädchen und Jungen. Der Einfluss der von der Lehrkraft eingeschätzten Kompetenz zum selbstgesteuerten Lernen. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, *44*(3), 153–162.

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*(1), 89–111.

Machts, N., Kaiser, J., & Schmidt, F. T. C. (2016). Accuracy of teachers' judgements of students' cognitive abilities: a meta-analysis. *Educational Research Review*, *19*, 85–103.

Marsh, H. W. (1990). A multidimensional, hierarchical model of self-concept: theoretical and empirical justification. *Educational Psychology Review*, *2*(2), 77–172.

Marsh, H. W., & Craven, R. (1991). Self-other agreement on multiple dimensions of preadolescent self-concept: Inferences by teacher, mothers, and fathers. *Journal of Educational Psychology*, *83*, 393–404.

Meissel, K., Meyer, F., Yao, E. S., & Rubie-Davies, C. M. (2017). Subjectivity of teacher judgments: Exploring student characteristics that influence teacher judgments of student ability. *Teaching and Teacher Education*, *65*, 48–60.

Nadler, J. T., & Clark, M. H. (2011). Stereotype threat: a meta-analysis comparing African Americans to Hispanic Americans. *Journal of Applied Social Psychology*, *41*(4), 872–890.

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*(2), 105–205.

Nisbett, R. E., & Wilson, T. D. (1977). The halo effect: evidence for unconscious alteration of judgements. *Journal of Personality and Social Psychology*, *35*(4), 250–256.

Pohlmann, B., Möller, J., & Streblow, L. (2004). Fremdeinschätzungen von Schülerselbstkonzepten durch Lehrer und Mitschüler. *Zeitschrift für Pädagogische Psychologie*, *18*(3/4), 157–169.

Praetorius, A.-K., & Südkamp, A. (2017). Eine Einführung in das Thema der diagnostischen Kompetenz von Lehrkräften. In A. Südkamp & A.-K. Praetorius (Hrsg.), *Diagnostische Kompetenz von Lehrkräften. Theoretische und methodische Weiterentwicklungen* (Pädagogische Psychologie und Entwicklungspsychologie, Bd. 94, S. 13–18). Münster: Waxmann.

Praetorius, A.-K., Greb, K., Lipowsky, F., & Gollwitzer, M. (2010). Lehrkräfte als Diagnostiker – Welche Rolle spielt die Schülerleistung bei der Einschätzung von mathematischen Selbstkonzepten? *Journal for Educational Research Online*, *2*(1), 121–144.

R Core Team (2017). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. http://www.R-project.org/.

Rakoczy, K., Klieme, E., Bürgermeister, A., & Harks, B. (2008). The interplay between student evaluation and instruction. *Journal of Psychology*, *216*(2), 111–124.

Ready, D. D., & Wright, D. L. (2011). Accuracy and inaccuracy in teachers' perceptions of young children's cognitive abilities. The role of child background and classroom context. *American Educational Research Journal*, *48*(2), 335–360.

Sander, A., & Ohle, A. (2017). Stereotypenbedrohung als Ursache für geringen Wortschatzzuwachs bei Grundschulkindern mit Migrationshintergrund. *Zeitschrift für Erziehungswissenschaft*, *21*, 177–197.

Schneider, W., Schlagmüller, M., & Ennemoser, M. (2017). *LGVT 5–12 – Lesegeschwindigkeits- und -verständnistest für die Klassen 5–12*. Göttingen: Hogrefe.

Schrader, F.-W. (2013). Diagnostische Kompetenz von Lehrpersonen. *Beiträge zur Lehrerinnen- und Lehrerbildung*, *31*(2), 154–165.

Schrader, F.-W., & Helmke, A. (1990). Lassen sich Lehrer bei der Leistungsbeurteilung von sachfremden Gesichtspunkten leiten? Eine Untersuchung zu Determinanten diagnostischer Lehrerurteile. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, *22*, 312–324.

Schrader, F.-W., & Helmke, A. (2008). Determinanten der Schulleistung. In M. K. W. Schweer (Hrsg.), *Lehrer-Schüler-Interaktion. Inhaltsfelder, Forschungsperspektiven und methodische Zugänge* (2. Aufl. Schule und Gesellschaft, Bd. 24, S. 285–302). Wiesbaden: VS, GWV.

Spinath, B. (2005). Akkuratheit der Einschätzung von Schülermerkmalen durch Lehrer und das Konstrukt der diagnostischen Kompetenz. *Zeitschrift für Pädagogische Psychologie*, *19*(1/2), 85–95.

Stanat, P., Rauch, D., & Segeritz, M. (2010). Schülerinnen und Schüler mit Migrationshintergrund. In E. Klieme, C. Artelt, J. Hartig, N. Jude, O. Köller, M. Prenzel, W. Schneider & P. Stanat (Hrsg.), *PISA 2009. Bilanz nach einem Jahrzehnt* (S. 200–230). Münster: Waxmann.

Stipek, D.J. (1981). Children's perceptions of their own and their classmates' ability. *Journal of Educational Psychology*, *73*(3), 404–410.

Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of teachers' judgements of students' academic achievement: a meta-analysis. *Journal of Educational Psychology*, *104*(3), 743–762.

Tenenbaum, H.R., & Ruck, M.D. (2007). Are teachers' expectations different for racial minority than for European American students? A meta-analysis. *Journal of Educational Psychology*, *99*(2), 253–273.

Tobisch, A., & Dresel, M. (2017). Negatively or positively biased? Dependencies of teachers' judgments and expectations based on students' ethnic and social backgrounds. *Social Psychology of Education*, *20*(4), 731–752.

Urhahne, D., Zhou, J., Stobbe, M., Chao, S.-H., Zhu, M., & Shi, J. (2010). Motivationale und affektive Merkmale unterschätzter Schüler. Ein Beitrag zur diagnostischen Kompetenz von Lehrkräften. *Zeitschrift für Pädagogische Psychologie*, *24*(3–4), 275–288.

Vazire, S. (2010). Who knows what about a person? The self–other knowledge asymmetry (SOKA) model. *Journal of Personality and Social Psychology*, *98*(2), 281–300.

Wagner, W., Helmke, A., & Rösner, E. (2009). *Deutsch Englisch Schülerleistungen international: Dokumentation der Erhebungsinstrumente für Schülerinnen und Schüler, Eltern und Lehrkräfte*. Frankfurt a. M.: GFPF, DIPF.

Wolf, C. (2010). Egozentrierte Netzwerke: Datenerhebung und Datenanalyse. In C. Stegbauer & R. Häußling (Hrsg.), *Handbuch Netzwerkforschung* (S. 471–483). Wiesbaden: Springer VS.

Zander, L., Webster, G.D., & Hannover, B. (2014). Better than me?! How adolescents with and without migration background perceive each others' performance in German classrooms. *Psychology*, *5*(1), 62–69.

Zhu, M., & Urhahne, D. (2015). Teachers' judgements of students' foreign-language achievement. *European Journal of Psychology of Education*, *30*(1), 21–39.

Ziegler, M., Danay, E., Schölmerich, F., & Bühner, M. (2010). Predicting academic success with the Big 5 rated from different points of view. Self-rated, Other rated and faked. *European Journal of Personality*, *24*, 341–355.