

Hübner, Nicolas; Kohler, Britta; Neumann, Marko; Becker, Michael

## **Die Vergessenen. Einblicke in kriteriale Präferenzen und die Gewichtung von mündlichen Schulnoten durch Lehrpersonen**

*Zeitschrift für Erziehungswissenschaft 29 (2026) 3, S. 763-788*



### Quellenangabe/ Reference:

Hübner, Nicolas; Kohler, Britta; Neumann, Marko; Becker, Michael: Die Vergessenen. Einblicke in kriteriale Präferenzen und die Gewichtung von mündlichen Schulnoten durch Lehrpersonen - In: Zeitschrift für Erziehungswissenschaft 29 (2026) 3, S. 763-788 - URN: urn:nbn:de:0111-pedocs-357103 - DOI: 10.25656/01:35710; 10.1007/s11618-026-01381-z

<https://nbn-resolving.org/urn:nbn:de:0111-pedocs-357103>

<https://doi.org/10.25656/01:35710>

in Kooperation mit / in cooperation with:



<https://info.oe-deepgreen.de>

### Nutzungsbedingungen

Dieses Dokument steht unter folgender Creative Commons-Lizenz: <http://creativecommons.org/licenses/by/4.0/deed.de> - Sie dürfen das Werk bzw. den Inhalt vervielfältigen, verbreiten und öffentlich zugänglich machen sowie Abwandlungen und Bearbeitungen des Werkes bzw. Inhaltes anfertigen, solange Sie den Namen des Autors/Rechteinhabers in der von ihm festgelegten Weise nennen.

Mit der Verwendung dieses Dokuments erkennen Sie die Nutzungsbedingungen an.

### Terms of use

This document is published under following Creative Commons-License: <http://creativecommons.org/licenses/by/4.0/deed.en> - You may copy, distribute and render this document accessible, make adaptations of this work or its contents accessible to the public as long as you attribute the work in the manner specified by the author or licensor.

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



### Kontakt / Contact:

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

Mitglied der

  
Leibniz-Gemeinschaft

# Die Vergessenen: Einblicke in kriteriale Präferenzen und die Gewichtung von mündlichen Schulnoten durch Lehrpersonen

Nicolas Hübner  · Britta Kohler · Marko Neumann · Michael Becker

Eingegangen: 17. April 2025 / Überarbeitet: 6. Oktober 2025 / Angenommen: 28. November 2025 /  
Online publiziert: 10. Februar 2026  
© The Author(s) 2026

**Zusammenfassung** Die Genese, Bedeutung und Aussagekraft von Schulnoten sind seit vielen Jahrzehnten Gegenstand gesellschaftspolitischer und wissenschaftlicher Debatten. Neben Forschungsarbeiten, die explizit Schulnoten untersuchen, erfüllen diese in zahlreichen Studien auch eine wichtige Funktion als Kontrollvariable. Somit fußt ein nicht unbeträchtlicher Anteil bildungswissenschaftlicher Forschungsarbeiten und Erkenntnisse, wenn nicht explizit, so zumindest partiell, auf Schulnoten und hier vorrangig auf Zeugnisnoten. Auffällig ist, dass zu mündlichen Noten bisher kaum Forschungsarbeiten existieren, obwohl ihnen rechtlich ein substanzielles Gewicht bei der Bildung von Zeugnisnoten eingeräumt wird. Bisher ist daher noch weitgehend unklar, welche Kriterien Lehrkräfte bei der Bildung mündlicher Noten verwenden, ob sich unterschiedliche LehrerInnentypen bei der Wahl dieser Kriterien identifizieren lassen und welche Bedeutung mündliche Noten für Lehrpersonen bei der Bildung von Zeugnisnoten haben. In der vorliegenden Studie werden diese Forschungsfragen

---

✉ Prof. Dr. Nicolas Hübner  
Bonner Zentrum für Lehrer\*innenbildung (BZL), Universität Bonn, Poppelsdorfer  
Allee 15, 53115 Bonn, Deutschland  
E-Mail: [nhuebner@uni-bonn.de](mailto:nhuebner@uni-bonn.de)

Prof. Dr. Britta Kohler  
Institut für Erziehungswissenschaft, Universität Tübingen, Münzgasse 26, 72070 Tübingen,  
Deutschland  
E-Mail: [britta.kohler@uni-tuebingen.de](mailto:britta.kohler@uni-tuebingen.de)

Dr. Marko Neumann  
Struktur und Steuerung des Bildungswesens (SteuBis), DIPF | Leibniz-Institut für Bildungsforschung  
und Bildungsinformation, Rostocker Str. 6, 60323 Frankfurt am Main, Deutschland  
E-Mail: [m.neumann@dipf.de](mailto:m.neumann@dipf.de)

Prof. Dr. Michael Becker  
Institut für Schulentwicklungsforschung (IFS), Technische Universität Dortmund,  
Vogelpothsweg 78, 44227 Dortmund, Deutschland  
E-Mail: [michael.becker@tu-dortmund.de](mailto:michael.becker@tu-dortmund.de)

auf der Grundlage einer für Deutschland repräsentativen Stichprobe von 762 Lehrpersonen der Sekundarstufe I näher beleuchtet. Die Ergebnisse legen nahe, dass Lehrkräfte insbesondere Qualitätskriterien bei der Bildung mündlicher Noten berücksichtigen, aber auch Aspekte des Arbeits- und Sozialverhaltens relevant sind. In latenten Klassenanalysen ließen sich drei Lehrkraft-Typen identifizieren: Eine Gruppe (All In-Gruppe; 49 %), die nahezu allen Kriterien bei der Leistungsbeurteilung eine hohe Bedeutung zuschrieb, eine zweite Gruppe (Quali-Gruppe; 18 %), die angab, primär Qualitätskriterien mündlicher Leistungen zu berücksichtigen, und eine dritte Gruppe, die neben den Qualitätskriterien auch Kriterien des Arbeits- und Sozialverhaltens (Quali-Plus; 33 %) bei der Bildung mündlicher Noten einbezog, für die aber z. B. die Heftführung und Hausaufgaben weniger wichtig waren als für die All In-Gruppe. Die Bedeutung einer vertieften Berücksichtigung und Untersuchung mündlicher Noten in erziehungswissenschaftlichen Studien wird diskutiert.

**Schlüsselwörter** Mündliche Noten · Leistungsbeurteilung · Gewichtung · Benotung · Lehrkraft-Typen

## **The forgotten ones: Insights into criteria preferences and the weighting of teacher-assigned oral grades**

**Abstract** The genesis, meaning, and significance of school grades have been subjects of socio-political and scientific debates for many decades. Beyond research specifically focused on school grades, they serve an important role as a control variable in many studies. Thus, a substantial amount of educational research and knowledge is based on grades, primarily report card grades. However, remarkably few studies focus on oral grades, even though they hold significant legal weight in the formation of report card grades. So far, it remains largely unclear which criteria teachers use when assigning oral grades, whether there are specific preference groups among teachers in selecting criteria, and to what extent teachers consider oral grades when determining report card evaluations. This study investigates these aspects in greater detail and is based on a representative sample of 762 lower secondary school teachers from Germany. The results suggest that teachers particularly consider quality criteria when assigning oral grades but that aspects of work and social behavior were also relevant. Three groups of teachers were identified in latent class analyses: Teachers in the first group (All-in group; 49%) assigned high importance to nearly all criteria. Teachers in the second group (Quali group; 18%) focused mainly on the quality criteria of oral performance. Teachers in the third group (Quali plus group; 33%) included criteria related to work and social behavior in addition to the quality criteria when evaluating oral performance, but for whom, for example, notebook organization and homework were less important than for the All-in group. The significance of explicitly considering and investigating oral grades in educational research will be discussed.

**Keywords** Oral grades · Performance Feedback · Weighting · Development · Teacher Types

## 1 Einleitung

Bereits seit langer Zeit werden Schulnoten kontrovers diskutiert (Ingenkamp 1971; Klieme 2022). Die Diskussionen resultieren häufig aus dem wahrgenommenen Widerspruch zwischen ihrer großen Relevanz für schulische Laufbahnentscheidungen und erfolgreiche Bewerbungen um Ausbildungs- oder Studienplätze einerseits und ihrer mangelhaften Vergleichbarkeit und unklaren Aussagekraft andererseits (vgl. Bohl 2019; Hübner et al. 2020). Diese Problematik verdeutlicht auch der inzwischen vielfach replizierte Befund, dass SchülerInnen mit identischen Testleistungen und Kompetenzen häufig substantiell unterschiedliche Schulnoten erhalten und identische Noten somit typischerweise ganz unterschiedliche Testleistungen und Kompetenzstände von SchülerInnen abbilden (Hübner et al. 2020, 2024; Neumann et al. 2011). Derartige Befunde werden auf unterschiedliche Faktoren zurückgeführt. So haben verschiedene Studien darauf hingewiesen, dass Schulnoten nicht allein die fachlichen Kompetenzen von SchülerInnen abbilden, sondern in unterschiedlichem Ausmaß von weiteren SchülerInnenmerkmalen beeinflusst werden, beispielsweise von ihrer Leistungsbereitschaft, Gewissenhaftigkeit und weiteren sozioemotionalen Merkmalen (Becker 2024; Meyer et al. 2023). Darüber hinaus weisen Beutel und Pant (2020) auf verschiedene Beobachtungs- und Beurteilungsfehler von Lehrkräften hin, die einen Einfluss auf die Benotung von SchülerInnen haben können (siehe hierzu auch Bonefeld und Dickhäuser 2018; Olczyk et al. 2023). Auch die Bezugsnormorientierung von Lehrkräften wurde in verschiedenen Beiträgen als bedeutsam für Schulnoten hervorgehoben (Bohl 2019; Hübner et al. 2020; Neumann et al. 2011).

Einen „blinden Fleck“ in den immer wieder aufkeimenden wissenschaftlichen und gesellschaftspolitischen Debatten über Noten bilden mündliche Schulnoten, obwohl diesen rechtlich in allen Bundesländern ein substantielles Gewicht bei der Bildung von Zeugnisnoten eingeräumt wird und sie somit von großer Bedeutung für die Bildungsbiographien und Lebenschancen junger Menschen sein können (Kohler et al. 2025). Auch vor dem Hintergrund der zunehmenden Digitalität von Schule und Unterricht und auf Grund neuartiger Möglichkeiten digitaler Tools, die mit künstlicher Intelligenz (KI) arbeiten, wurde vermehrt auf die Notwendigkeit einer Renaissance des Mündlichen in Lehr-, Lern- und Prüfungssituationen in der Schule hingewiesen (KMK 2021). Für derartige Entwicklungen ist ein besseres Verständnis der Beurteilung mündlicher Leistungen durch Lehrpersonen wichtig.

Bisher ist in weiten Teilen unklar, welche Informationen Lehrkräfte bei der Bildung mündlicher Noten berücksichtigen und welche Lehrpersonenmerkmale mit der Wahl bestimmter Kriterien assoziiert sind. Darüber hinaus liegen keine Befunde zu der Frage vor, ob sich auf Basis der Präferenzen für einzelne Kriterien bestimmte „Benotungstypen“ von Lehrkräften bei der Vergabe mündlicher Noten identifizieren lassen, wie mündliche Schulnoten im Zuge der Bildung von Gesamtnoten gewichtet werden und mit welchen Faktoren diese Gewichtung in Zusammenhang steht. Der vorliegende Beitrag adressiert diese Fragestellung unter Verwendung einer bundeslandübergreifenden, repräsentativen LehrerInnenstichprobe der Sekundarstufe I.

## 2 Mündliche Noten

### 2.1 Forschungsstand zu mündlichen Leistungen und Kriterien mündlicher Noten

Die Studienlage zu mündlichen Schulnoten bzw. zur Beurteilung mündlicher Leistungen ist sehr überschaubar: Seit Jahrzehnten wird immer wieder aufs Neue festgestellt, dass mündliche Noten praktisch keine Berücksichtigung in der Literatur finden (z.B. Kirk 2004; Krüll 2023) oder allenfalls als kurz abgehandelter „Spezialfall“ (Jürgens und Sacher 2008) in den Fokus geraten. Krieger (2003) konnte in einer Erhebung nach erfolgter Kategorienbildung zeigen, dass für 479 retrospektiv befragte Lehramtsstudierende zum „Mündlichen“ folgende Teilaspekte in der angegebenen Reihenfolge gehörten: „Vokabeln aufsagen“, „An der Tafel etwas zeigen, erklären“, „Hausaufgaben vortragen“, „Schriftliche Bearbeitung von Aufgaben (Hausaufgaben, Protokolle, Heftführung)“, „Diskussionsbeiträge“, „Referate und mündliche Stoffwiederholung“ (S. 81). Sofern ihnen von ihren eigenen Lehrkräften in der Schule Kriterien genannt worden waren, handelte es sich hierbei vor allem um regelmäßige Mitarbeit sowie um die Qualität der Beiträge. In der selben Studie von Krieger (2003) nannten 111 Lehrpersonen folgende Kriterien: „Qualität der Beiträge“, „Quantität der Beiträge“, „Hausaufgaben und Heftführung“, „Sprachlicher Ausdruck“, „Referate, Protokolle“, „Leistungen beim Abfragen, Vokabeltests usw.“ (S. 84).

Auch Kirk (2004) stellte Ergebnisse einer Befragung von Lehrpersonen vor. Als ein Ergebnis wurde benannt, dass „sowohl Mitarbeit, Gesprächsbeiträge, Aufmerksamkeit als auch thematisch richtige Unterrichtsbeiträge beurteilt würden“ (Kirk 2004, S. 14) und zudem „Kooperation, Selbstständigkeit, Ausdauer und Konzentration“ bedeutsam für mündliche Noten seien. Diese Nennungen wurden im Text noch weiter fachspezifisch aufgefächert. Hingewiesen wurde außerdem darauf, dass in der Praxis häufig auch schriftliche Leistungen dem Mündlichen zugeordnet würden, so zum Beispiel die „Heftführung, Hausaufgaben, schriftliche Lernkontrollen und Kurztests“ (Kirk 2004, S. 8).

In der Studie von Krüll (2023) wurden nach einer Auswertung von 15 ExpertInneninterviews insgesamt 272 Gymnasiallehrkräfte in Nordrhein-Westfalen gefragt, ob sie 20 in den Interviews identifizierte Leistungsarten der mündlichen Mitarbeit und/oder den sonstigen Leistungen zuordnen würden. Am häufigsten zur mündlichen Mitarbeit gehörend wurden (in dieser Reihenfolge) gewählt: „Mündliche Beiträge im lehrerzentrierten Unterricht“, „mündliche Abfragen“, „Präsentation von Hausaufgaben“, „Podiumsdiskussionen“, „Vorlesen“, „Präsentation von Arbeitsergebnissen“, „Rollenspiele“, „Tafelarbeit“ sowie „mündliche Beiträge in Partner- oder Gruppenarbeiten“ (S. 266). Als besonders relevant für die Bewertung der mündlichen Mitarbeit wurden die „eigene Denkleistung“, eine „kontinuierliche Leistungserbringung“, die „fachliche Richtigkeit“, die „Weiterentwicklung einer Argumentation“ und das „Vorbringen des Unterrichts“ genannt (S. 271); für die „Anzahl an Beiträgen zum Unterrichtsgespräch“ ergab sich hingegen eine etwas weniger hohe Zustimmung. Das Arbeitsverhalten wurde von einer klaren Mehrheit in die mündliche Mitarbeit

einbezogen (S. 306). Darüber hinaus gaben die Lehrkräfte an, dass Bezugsnormen für die Bewertung mündlicher Leistungen im Mittel sehr wichtig seien.

## 2.2 Relevanz mündlicher Noten für Gesamtnoten

Die Frage, welcher Anteil an der Gesamtfachnote dem Mündlichen zukommt, wird zunächst durch rechtliche Regelungen bestimmt, die sich länderspezifisch unterscheiden. Wie Kohler et al. (2025) beispielhaft für Gymnasien zeigen konnten, liegen in einigen Bundesländern keine Vorgaben zur Gewichtung mündlicher Leistungen vor (z.B. Baden-Württemberg), während ihr Anteil in Bayern bei weniger als 50 % liegt und in anderen Bundesländern bei mehr als 50 % (z.B. Schleswig-Holstein). Unterschiede zwischen den Ländern gibt es auch bezüglich der Frage, ob die Mitarbeit als Teil der Fachnote gilt oder, wie z.B. das (Sozial-)Verhalten, als eigener Bereich im Zeugnis ausgewiesen wird (vgl. auch Krüll 2023). In den wenigen vorhandenen Studien zu mündlichen Noten wurde die Frage nach der jeweiligen Gewichtung überwiegend entweder nicht gestellt (z.B. Krieger 2003) oder nicht präzise dokumentiert (z.B. Kirk 2004). Befunde von Krüll (2023) für NRW legen nahe, dass der Anteil der mündlichen Note an der Gesamtnote durchschnittlich bei einem Drittel liegt.

## 2.3 Lehrkräftemerkmale und mündliche Noten

Inwiefern nicht nur Merkmale der Lernenden und Kontextmerkmale, sondern auch Lehrkräftemerkmale relevant für die Bildung mündlicher Noten und ihre Gewichtung sind, ist bisher noch weitgehend unerforscht. Die vorliegende Studie kann hier erste Einblicke liefern, keineswegs jedoch ein vollumfängliches Bündel aller möglicherweise relevanten Merkmale berücksichtigen. Im Fokus liegen daher zunächst Aspekte der Persönlichkeit und Bezugsnormorientierung von Lehrkräften.

**Persönlichkeitsmerkmale** Unter Lehrkräftepersönlichkeit sollen an dieser Stelle Persönlichkeitsmerkmale von Lehrpersonen verstanden werden, die ein „Ensemble relativ stabiler Dispositionen [darstellen], die für das Handeln, den Erfolg und das Befinden im Lehrerberuf bedeutsam sind“ (Mayr und Neuweg 2006, S. 183). In der vorliegenden Studie wurde eine gängige Taxonomie der Persönlichkeit verwendet (die Big Five), die eine Unterscheidung von fünf Faktoren (Extraversion, Verträglichkeit, Gewissenhaftigkeit, Neurotizismus und Offenheit) vorsieht (John et al. 2008).

Das Modell der Determinanten und Konsequenzen professioneller Kompetenzen von Lehrkräften (Kunter et al. 2011) geht davon aus, dass das Professionswissen, z.B. das pädagogische (psychologische) Wissen (Voss et al. 2011), aber auch professionelle Werte, Überzeugungen, subjektive Theorien und normative Präferenzen (Baumert und Kunter 2013), durch die Persönlichkeit von Lehrpersonen beeinflusst werden können. In verschiedenen Studien fanden sich Zusammenhänge zwischen Persönlichkeitsmerkmalen und weiteren Variablen, die für den Unterricht und die Benotungsprozesse relevant sein können. Beispielsweise zeigte eine Studie von Baier et al. (2019), dass Persönlichkeitsmerkmale mit Dimensionen von Unterrichtsqualität

assoziiert waren. Darüber hinaus verdeutlichen Ergebnisse von Rausch et al. (2016), dass die Persönlichkeit von Lehrpersonen auch ganz konkrete Beurteilungsprozesse beeinflussen kann: Die Ergebnisse der AutorInnen legen nahe, dass Lehrpersonen SchülerInnen mit einer Persönlichkeit, die ihrer Persönlichkeit ähnlicher ist, besser beurteilen, auch nach Kontrolle der Testleistungen. Auch Beutel und Pant (2020) stellen Überlegungen an, dass unterschiedliche Beobachtungs- und Beurteilungsfehler im Zuge schriftlicher oder mündlicher Leistungskontrollen als Effekte der Lehrkräftepersönlichkeit auftreten könnten (vgl. hierzu auch Hübner et al. 2024).

In Bezug auf mündliche Noten erscheint es daher nicht unplausibel, dass gewissenhaftere Lehrkräfte ein breiteres und stärker elaboriertes Spektrum möglicher Kriterien bei ihrer Beurteilung berücksichtigen, um SchülerInnen möglichst genau und umfassend zu bewerten. Derartige Praktiken zur (vermeintlichen) Überwindung der Subjektivität von Leistungsbeurteilungsprozessen durch die Berücksichtigung möglichst vielfältiger Informationen über SchülerInnen wurden in der Literatur bereits verschiedentlich diskutiert (vgl. Bräu und Schlickum 2015; Zaborowski et al. 2011). Möglich erscheint auch, dass verträglichere Lehrpersonen grundsätzlich mehr Kriterien berücksichtigen, beispielsweise, um bei Prozessen der Leistungsbeurteilung über größere Spielräume zu verfügen und somit insgesamt zu wohlwollenderen/verträglicheren Urteilen zu gelangen. Derartige Prozesse wären sehr vergleichbar zu bereits von Baumert et al. (2003) diskutierten Einflüssen von Milde- und Strengentendenzen auf die Notengebung. Nicht zuletzt könnte auf Basis der Befunde von Rausch et al. (2016) vermutet werden, dass extrovertierte Lehrpersonen mündliche Noten stärker gewichten als introvertierte Lehrpersonen, weil sie mehr Wert auf die Häufigkeit der Beteiligung an Unterrichtsgesprächen legen.

**Bezugsnormorientierung** Ganz allgemein wird mit Bezugsnorm eine Art Standard oder Referenz beschrieben, die zur Beurteilung der Leistung von SchülerInnen verwendet wird (Heckhausen 1974; Rheinberg und Fries 2018). In der Literatur wird berichtet, dass sich Lehrkräfte mit unterschiedlichen Bezugsnormorientierungen beispielsweise auch in Bezug auf die Kausalattribution von Leistungsunterschieden, Erwartungen an die Leistungsentwicklung, Sanktionsstrategien und die Individualisierung von Unterrichtsangeboten unterscheiden (Köller 2005). So attribuieren Lehrkräfte mit stärkerer individueller Bezugsnormorientierung SchülerInnenleistungen eher auf zeitvariable Faktoren (z. B. Anstrengung oder Interesse), während Lehrkräfte mit stärkerer sozialer Bezugsnormorientierung hier eher zeitstabile Faktoren (z. B. Intelligenz, Begabung, Arbeitscharakter) für bedeutsam halten (Rheinberg und Fries 2018). Davon ausgehend erscheint es plausibel, dass Lehrkräfte mit einer stärkeren individuellen Bezugsnormorientierung andere Kriterien bei der Bildung mündlicher Noten berücksichtigen als Lehrkräfte, die stärker nach der sozialen Bezugsnorm beurteilen.

## 2.4 Fragestellungen

Im vorliegenden Beitrag werden insgesamt drei Fragekomplexe bearbeitet.

**Fragestellung 1a–c** Welche *Kriterien* berücksichtigen Lehrpersonen bei der Bildung der mündlichen Noten (a), welche zugrundeliegenden latenten Faktoren lassen sich in Bezug auf die Kriterien identifizieren (b) und welche Lehrpersonenmerkmale beeinflussen die Wahl bestimmter Kriterien (c)? Auf der Basis bisheriger Befunde wird erwartet, dass die Qualität mündlicher Beiträge im Unterrichtsgespräch von primärer Bedeutung für die Bildung der mündlichen Note ist. Darüber hinaus erscheint es auf Basis der referierten Literatur nicht unplausibel, dass sich faktorenanalytisch unterschiedliche latente Teilbereiche identifizieren lassen, die sich etwa auf die Qualität mündlicher Leistungen, schriftlicher Leistungen/Hausaufgaben und das Arbeits- und Sozialverhalten beziehen.

**Fragestellung 2a–b** Welche *Benotungs-Typen* lassen sich auf Basis der unterschiedlichen kriterialen Präferenzen identifizieren (a) und welche Lehrpersonenmerkmale sagen die Zugehörigkeit zu einem bestimmten Benotungs-Typ vorher (b)? Trotz des auf Grund der kaum entwickelten Befundlage stark explorativen Charakters der Fragestellungen erscheint es plausibel, dass ein Großteil der Lehrpersonen Kriterien berücksichtigt, die mit der Qualität von mündlichen Beiträgen zusammenhängen. Inwieweit sich darüber hinaus Gruppen von Lehrpersonen identifizieren lassen, die weitere Kriterien für wichtig erachten (z. B. auch das Arbeits- und Sozialverhalten oder schriftliche Leistungen) bleibt eine offene Forschungsfrage. In Bezug auf die Lehrkräftepersönlichkeit erscheint es nicht unplausibel, dass verträglichere Lehrpersonen grundsätzlich mehr Kriterien berücksichtigen, beispielsweise, um bei Prozessen der Leistungsbeurteilung über größere Spielräume zu verfügen und somit insgesamt zu wohlwollenderen/verträglicheren Urteilen zu gelangen. Ebenfalls plausibel erscheint, dass gewissenhaftere Lehrkräfte zur (vermeintlichen) Überwindung der Subjektivität von Leistungsbeurteilungsprozessen mehr Kriterien berücksichtigen (vgl. Abschn. 2.3).

**Fragestellung 3a–b** Wie wird die mündliche Note bei der Bildung der Gesamtnote *gewichtet* (a) und welche LehrerInnenmerkmale sind mit der Gewichtung assoziiert (b)? Auf Basis der Ergebnisse von Kohler et al. (2025) wäre in Bezug auf die Gewichtung ein gemischtes Bild zu erwarten, da in einem substantiellen Anteil der Länder keine einheitlichen Vorgaben zur Gewichtung der mündlichen Note in der Gesamtnote existieren. Denkbare Unterschiede zwischen Fächergruppen zeigen könnten, beispielsweise dahingehend, dass mündliche Beiträge in geistes- oder sozialwissenschaftlichen Fächern anders gewichtet werden als in MINT-Fächern.

### 3 Methode

#### 3.1 Stichprobe und Studiendesign

In der vorliegenden Studie wurden Daten aus einer repräsentativen Stichprobe von 762 LehrerInnen an allgemeinbildenden Schulen ausgewertet. Die Erhebung wurde vom 14. bis 26. September 2023 im Rahmen des repräsentativen Online-Befragungspanel *forsa.omninet* durchgeführt. Um sicherzustellen, dass die Stichprobe den bekannten Populationsparametern von Lehrkräften aus der Schulstatistik 2022/23 hinsichtlich des Alters, des Geschlechts, der Schulform und in Bezug auf das Bundesland entspricht, wurde unter Berücksichtigung dieser Informationen von *forsa* ein Gewichtungsfaktor berechnet, der bei den Analysen (vgl. Abschn. 4.3 Statistische Analyse) verwendet wurde. Die TeilnehmerInnen wurden per E-Mail eingeladen und bei Nichtteilnahme durch zwei weitere E-Mails erinnert. Insgesamt wurden 1662 Personen zur Befragung eingeladen, von denen 1362 (82 %) an der Befragung teilnahmen. Informationen von 247 Personen konnten nicht berücksichtigt werden, da sie aktuell nicht als Lehrperson tätig waren (128), an ihrer Schule keine Noten vergeben wurden (73), sie nicht an einer allgemeinbildenden Schule tätig waren (6) oder weil sie die Befragung abbrachen (40). Berücksichtigt wurden in dieser Studie nur Lehrpersonen aus der Sekundarstufe I, sodass sich die finale Analysestichprobe auf insgesamt 762 Lehrkräfte belief (68 % weiblich). Rund 56 % der Lehrpersonen unterrichteten an einer nichtgymnasialen Schulform. Von den TeilnehmerInnen unterrichteten 60 % Mathematik, Informatik, eine Naturwissenschaft oder Technik (MINT), 59 % unterrichteten eine Sprache, 46 % eine Geistes- oder Sozialwissenschaft und 30 % ein anderes Fach (z. B. Musik, Sport oder Kunst). Im Durchschnitt waren die Lehrkräfte 45 Jahre alt und unterrichteten seit 16 Jahren an einer Schule.

#### 3.2 Instrumente

**Kriterien mündlicher Noten** Von besonderer Bedeutung war in der vorliegenden Studie ein Instrument, bei dem Lehrkräfte die Wichtigkeit unterschiedlicher Kriterien bei der Bildung mündlicher Noten in einem zuvor angegebenen Fach beurteilen sollten. Der Fachbezug wurde erfasst, um mögliche fachspezifische Unterschiede bei der Wahl der Kriterien bei den späteren Analysen berücksichtigen zu können. Hierzu wurde den Lehrpersonen im Onlinefragebogen zufällig ein Fach aus der Liste der von ihnen zuvor angegebenen unterrichteten Fächer vorgegeben, für das sie den Fragebogen ausfüllen sollten. Das Instrument zur Untersuchung der für die mündliche Benotung relevanten Kriterien bestand aus insgesamt 16 Items, für die Lehrpersonen die Wichtigkeit für die von ihnen (im entsprechenden Fach) vergebenen mündlichen Noten auf einer Skala von sehr unwichtig (1) bis sehr wichtig (6) angeben sollten. Die 16 Items resultierten aus einem zunächst deutlich umfangreicheren Katalog möglicherweise relevanter Kriterien, der einerseits auf einer Literaturrecherche im Fachportal Pädagogik, bei Google Scholar und PsycInfo basierte sowie andererseits auf ExpertInnen-Interviews (Meuser und Nagel 2009) mit verschiedenen Lehrpersonen. Der Kriterienkatalog wurde von den AutorInnen der Studie zusammengefasst, in ein Instrument überführt und erneut von fünf Lehrpersonen durch cognitive labs

validiert. Im Rahmen einer Pilotierung mit 50 Lehrpersonen wurde im Anschluss das Verständnis der Items geprüft und andererseits durch offene Antwortmöglichkeiten sichergestellt, dass keine wichtigen Kriterien ausgelassen wurden. Bei den Kriterien handelte es sich beispielsweise um die „Qualität der Antwort bei Abfragen“, die „Qualität von Präsentationen“, die „Aufmerksamkeit der Schülerin/des Schülers“, den „sprachlichen Ausdruck“ oder die „Qualität der Hausaufgaben“. Eine Liste aller Items findet sich in Appendix B. Die Reliabilität des Weighted Likelihood Schätzers (WLE) belief sich im Rahmen einer Schätzung mittels Generalized Partial Credit Model (Muraki 1992) auf  $Rel.=0,86$ . Eine differenzierte Untersuchung der zugrundeliegenden faktoriellen Struktur erfolgt im Rahmen von Fragestellung 1 (vgl. Ergebnisse).

**Gewichtung der mündlichen Note** Die Gewichtung der mündlichen Note bei der Bildung einer Gesamtnote/Zeugnisnote wurde durch ein Instrument erfasst, bei dem Lehrpersonen einen Prozentwert angeben konnten. Für die Fragekomplexe 1 und 2 fungierte die Gewichtung der mündlichen Note als Kovariate (siehe nächster Abschnitt), für Fragestellung 3 als abhängige Variable.

**Persönlichkeit** Zusätzlich wurde die Persönlichkeit der Lehrpersonen mit dem Big Five Inventory-2 XS (Rammstedt et al. 2020; Soto und John 2017) erfasst, das eine Kurzversion des Big Five Inventory-2 darstellt und aus insgesamt 15 Items besteht. Die Validität des Instruments wurde im Rahmen einer deutschen Validierungsstudie bestätigt. Das Instrument erfasst die Big Five (Gewissenhaftigkeit, Neurotizismus, Verträglichkeit, Extraversion und Offenheit) mit jeweils 3 Items. Die WLE-Reliabilität für Gewissenhaftigkeit belief sich auf  $Rel.=0,75$ , für Neurotizismus auf  $Rel.=0,70$ , für Verträglichkeit auf  $Rel.=0,59$ , für Extraversion auf  $Rel.=0,62$  und für Offenheit auf  $Rel.=0,41$ . Zur Kontrolle des Messfehlers der verschiedenen Schätzer wurde in allen nachfolgenden Modellen ein in der Literatur häufig verwendeter Single-Indicator-Approach verwendet. Hierbei wird die Residualvarianz der Persönlichkeitsindikatoren auf den Wert  $(1-Reliabilität)*Stichprobenvarianz$  fixiert (Hoyle 2012; Hübner et al. 2022a).

**Bezugsnormorientierung** Die Wichtigkeit der Berücksichtigung der sozialen, kriterialen und individuellen Bezugsnorm bei der Vergabe mündlicher Noten wurde auf einer Skala von 1 (sehr unwichtig) bis 6 (sehr wichtig) gemessen. Die drei Bezugsnormen wurden jeweils mit einem Item erfasst (z. B. „Wie wichtig ist das Erreichen von zuvor festgelegten Kriterien bzw. eines zuvor festgelegten Erwartungshorizontes bei den von Ihnen vergebenen mündlichen Noten“).

**Kovariaten** Zur Beantwortung der Fragestellungen und zur Sicherstellung einer möglichst umfassenden Exploration des Forschungsgegenstandes wurde ein umfangreiches Bündel an Kontrollvariablen berücksichtigt. Hierbei handelt es sich beispielsweise um demographische Angaben zum Geschlecht, Dienstalter und Unterrichtsfach. Ausführliche Informationen zu den Kovariaten finden sich in Appendix A0.

## 4 Statistische Analyse

Vor der Beantwortung der einzelnen Fragestellungen erfolgte zunächst eine umfassende Datenaufbereitung in R (R Core Team 2024). Hierbei wurden alle Variablen inspiziert und ggf. rekodiert, sodass höhere numerische Werte auf den einzelnen Items größeren Ausprägungen des gemessenen Konstrukts entsprachen. Anschließend wurden die Daten mit dem R Package MplusAutomation (Hallquist und Wiley 2018) für die weiteren Analysen in Mplus 8.10 (Muthén und Muthén 2017) vorbereitet. Beispielcodes für zentrale Modelle zur Beantwortung der Fragekomplexe 1–3 befinden sich in Appendix C.

Im ersten Schritt zur Beantwortung von *Fragestellung 1a* erfolgte eine deskriptive Betrachtung der 16 Items zu den Kriterien der mündlichen Leistung. Hierbei wurden die Zustimmung zu den einzelnen Kategorien auf einer Skala von 1 (sehr unwichtig) bis 6 (sehr wichtig) sowie die Mittelwerte und Standardabweichungen je Item betrachtet.

Da es sich bei dem Instrument zur Erfassung der für die Bildung mündlicher Noten relevanten Kriterien um ein neuartiges Instrument handelt, für das zunächst keine klaren Hinweise zur Spezifikation des Messmodells vorlagen, wurde im nächsten Schritt zur Beantwortung von *Fragestellung 1b* die zugrundeliegende faktorielle Struktur der 16 Kriterien mit Exploratorischen Strukturgleichungsmodellen (ESEM) bestimmt. ESEMs ermöglichen eine Kombination von Eigenschaften exploratorischer und konfirmatorischer Faktorenanalysen (Hübner et al. 2022b; Marsh et al. 2014). Zur Identifikation einer passenden Lösung wurden ESEMs mit einer ansteigenden Zahl von Faktoren (3 bis 6 Faktoren) und geomin-rotierten Faktorladungen spezifiziert. Zur Beurteilung der Modellgüte wurden verschiedene Fitindikatoren berücksichtigt. Hierbei handelte es sich um den Root Mean Square Error of Approximation (RMSEA), den Standardized Root Mean Squared Residual (SRMR), den Tucker-Lewis Index (TLI) sowie den Comparative Fit Index (CFI). Schließlich wurden zur Beantwortung von *Fragestellung 1c* Multiple Indicator Multiple Causes (MIMIC)-ESEM-Modelle spezifiziert. Diese Modelle lassen sich als Erweiterung zu den für Fragestellung 1b spezifizierten ESEM-Modellen verstehen, in denen die resultierenden latenten Beurteilungs-Faktoren (Messmodell) durch die oben genannten Kovariaten (vgl. Abschn. 5.2) im Rahmen von multiplen Regressionsmodellen vorhergesagt werden (Strukturmodell).

Im Rahmen von *Fragestellung 2a* sollten mögliche Benotungstypen von Lehrpersonen identifiziert werden. Hierzu wurden latente Klassenanalysen unter Verwendung der 16 Kriterien zur mündlichen Notenbildung mit der Mplus Option TYPE = MIXTURE durchgeführt. Die Entscheidung für die Verwendung von LCAs anstelle möglicher Alternativen (z. B. Mixed-Rasch-Modellen) resultierte dabei insbesondere aus Abwägungen hinsichtlich der Angemessenheit der Annahmen (Rost 1990), der Sparsamkeit und der Verbreitung des verwendeten Verfahrens in der relevanten Fachliteratur. Es wurden insgesamt drei Lösungen mit 3–5 Gruppen spezifiziert und miteinander verglichen. Hierbei wurden insbesondere informatorische Maße wie das Akaike's Information Criterion (AIC), das Bayesian Information Criterion (BIC) und das für die Stichprobengröße korrigierte BIC (aBIC) Kriterium betrachtet. Darüber hinaus wurden die mittleren Klassenzuordnungswahrscheinlichkeiten,

das Entropy-Maß und die Interpretierbarkeit der Klassenlösung berücksichtigt. Auf Grundlage der Empfehlungen von Nylund-Gibson und Masyn (2016) wurden in den Klassenanalysen zunächst keine weiteren Kovariaten berücksichtigt.

Zur Beantwortung von *Fragestellung 2b* wurden die im Rahmen von Fragestellung 2a generierten individuellen Zuordnungswahrscheinlichkeiten für die gewählte Klassenlösung extrahiert. Anschließend wurden die individuellen Zuordnungswahrscheinlichkeiten im Rahmen von multiplen Regressionsmodellen vorhergesagt. Hierzu fungierten sie als abhängige Variable und die Kovariaten (vgl. Abschn. 5.2) als unabhängige Variablen.

Zur Beantwortung von *Fragestellung 3a* erfolgte zunächst eine Betrachtung der Gewichtung der mündlichen Note bei der Bildung einer Gesamtnote sowohl für alle Lehrpersonen als auch differenziert aufgeschlüsselt für das Geschlecht der Lehrkraft, die Schulform und das Fach. Zur Beantwortung von *Fragestellung 3b* wurde anschließend die Gewichtung im Rahmen eines multiplen Regressionsmodells durch die Kovariaten, äquivalent zu Fragestellung 1c und 2b, vorhergesagt. Im Gegensatz zu den Fragestellungen 1c und 2b fungierte hier die Gewichtung als abhängige und nicht mehr als unabhängige Variable.

Alle Analysen wurden mit Stichprobengewichten und als Full Information Maximum Likelihood (FIML)-basierte Schätzungen durchgeführt (Graham 2009). Die Nestung der Daten (LehrerInnen in Bundesländern) wurde in den jeweiligen Prädiktionsmodellen (Fragestellung 1c, 2b und 3b) durch die Schätzung cluster-robuster Standardfehler berücksichtigt (McNeish 2023).

## 5 Ergebnisse

### 5.1 Fragekomplex 1: Kriterienwahl, faktorielle Struktur und Determinanten der Kriterienwahl

Die Ergebnisse von Fragestellung 1a zur Wichtigkeit der unterschiedlichen Kriterien bei der Bildung mündlicher Noten sind in Abb. 1 näher dargestellt. Die Wichtigkeit der 16 Kriterien variierte von  $M=2,97$  ( $SD=1,44$ ) für eine „sorgfältige Heftführung“ bis zu  $M=5,22$  ( $SD=1,05$ ) für die „Qualität von Beiträgen in Unterrichtsgesprächen“. Insgesamt beurteilte ein Großteil der befragten Lehrpersonen die Kriterien als wichtig. Lediglich die Items „Qualität von Hausaufgaben“, „kein störendes Verhalten im Unterricht“ und „sorgfältige Heftführung“ wurden von 50 % oder mehr Lehrkräften als weniger wichtig beurteilt. Besonders wichtig erschienen hingegen, neben der „Qualität von Beiträgen in Unterrichtsgesprächen“, die „Leistungsbereitschaft“, die Qualität von „Präsentationen“, von „Antworten bei Abfragen“ und von „Beiträgen in Gruppenarbeiten“.

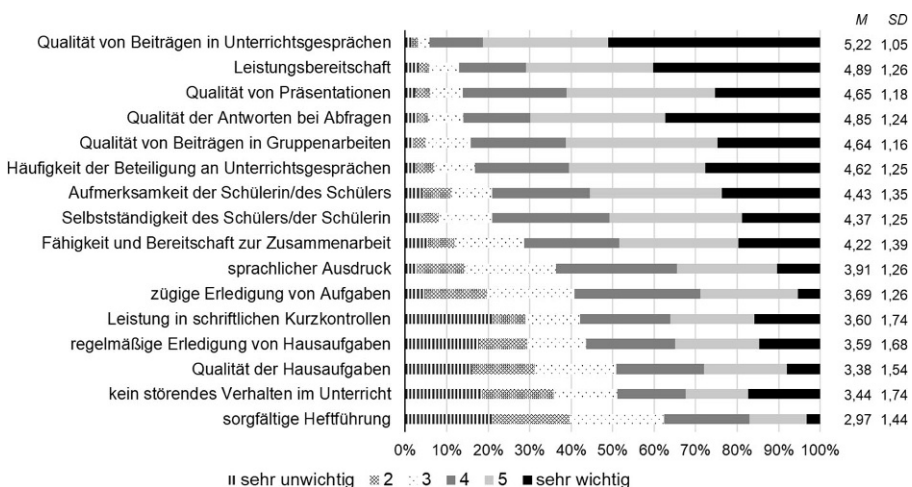
Fragestellung 1b beschäftigte sich mit möglichen übergeordneten latenten Dimensionen der Kriterien aus Fragestellung 1a. Hierzu wurden verschiedene ESEM-Modelle spezifiziert. Die Modellfits für die verschiedenen Lösungen (von 3 bis 6 Faktoren) sind in Appendix A1 dargestellt. Wie ersichtlich ergab sich bei der Spezifikation einer Lösung mit 4 Faktoren ein zufriedenstellender Modellfit  $\chi^2=143,900$ ,  $df=60$ ,  $CFI=0,97$ ,  $TLI=0,93$ ,  $RMSEA=0,04$ ,  $SRMR=0,02$ , der sich durch die

Hinzunahme weiterer Faktoren kaum verbesserte. Inhaltlich ergaben sich aus der 4-Faktor-Lösung die Faktoren: (1) „Qualität mündlicher Beiträge“, (2) „Arbeits- und Sozialverhalten“, (3) „Qualität schriftlicher Leistungen und Hausaufgaben“ sowie (4) „Gruppenarbeit“. Die Faktorladungen finden sich in Tab. A2 im Anhang.

Bei der Beantwortung von Fragestellung 1c nach den Bezügen der identifizierten Faktoren zu LehrerInnenmerkmalen (vgl. Tab. 1) zeigte sich, dass die Qualität mündlicher Beiträge für Lehrerinnen wichtiger war als für Lehrer und von Lehrkräften an Gymnasien als bedeutsamer beurteilt wurde als von Lehrpersonen an nicht-gymnasialen Schulformen. Darüber hinaus zeigte sich, dass der Faktor Qualität mündlicher Beiträge nach Kontrolle der übrigen Variablen im Modell für Lehrpersonen mit Schwierigkeiten bei der Beurteilung mündlicher Leistungen weniger bedeutsam war und in einem positiven Zusammenhang mit der sozialen und kriterialen Bezugsnormorientierung von Lehrpersonen stand.

Mit dem zweiten Faktor, dem Arbeits- und Sozialverhalten (vgl. auch Tab. A2), waren das Bundesland (Ost/West) sowie die soziale und individuelle Bezugsnormorientierung der Lehrpersonen positiv assoziiert: Lehrpersonen in Ostdeutschland gaben an, diesen Faktor bei der Bildung mündlicher Noten im Mittel höher zu gewichten als Lehrpersonen in Westdeutschland. Gleichzeitig wiesen Lehrpersonen mit höheren Werten auf dem Faktor Arbeits- und Sozialverhalten auch höhere Werte bei der sozialen und individuellen Bezugsnormorientierung auf.

Die Qualität schriftlicher Leistungen und von Hausaufgaben war insbesondere für Lehrpersonen in einem höheren Dienstalter und gewissenhaftere Lehrpersonen von Bedeutung. Darüber hinaus zeigte sich, dass Lehrkräfte, die auf diesem Faktor höhere Werte aufwiesen, die mündliche Note insgesamt höher gewichteten, sich häufiger Notizen zur mündlichen Leistung von SchülerInnen machten und die Beurteilung nach der kriterialen und sozialen Bezugsnorm für wichtiger erachteten.



**Abb. 1** Häufigkeiten, Mittelwerte und Standardabweichungen der 16 Kriterien zur Beurteilung mündlicher Leistungen. Häufigkeiten, Mittelwerte und Standardabweichungen wurden mit FIML und Gewichten berechnet.  $N=759$

**Tab. 1** Ergebnisse der MIMIC-ESEM-Modelle zur Vorhersage der vier Faktoren der mündlichen Notenvergabe

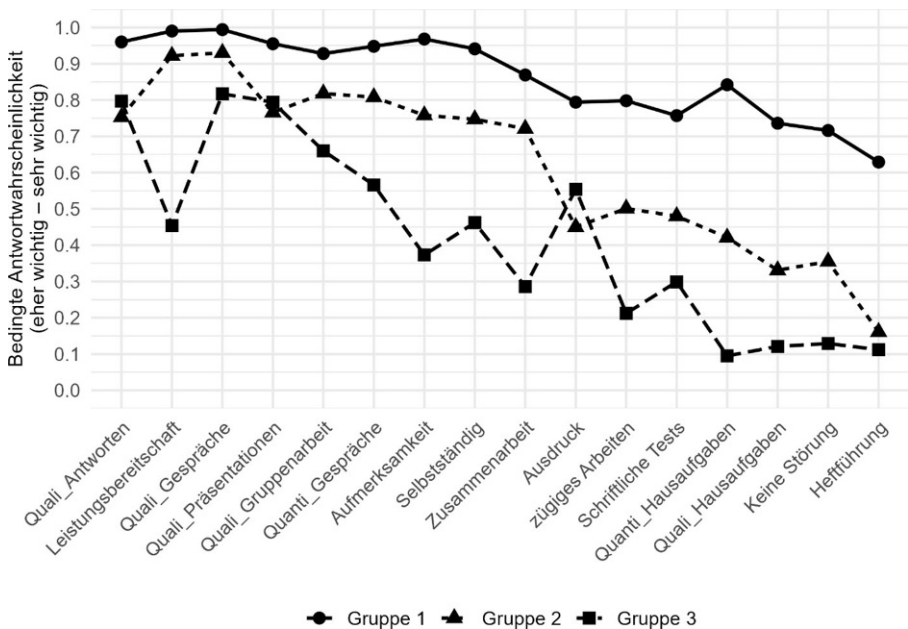
| Abhängige Variable                      | Qualität mündlicher Beiträge |      |        | Arbeits- und Sozialverhalten |      |        | Qualität schriftlicher Leistungen und Hausaufgaben |             |        | Gruppenarbeit |      |       |
|-----------------------------------------|------------------------------|------|--------|------------------------------|------|--------|----------------------------------------------------|-------------|--------|---------------|------|-------|
|                                         | B                            | SE   | p      | B                            | SE   | p      | B                                                  | SE          | p      | B             | SE   | p     |
| Prädiktor                               |                              |      |        |                              |      |        |                                                    |             |        |               |      |       |
| Geschlecht (1 = weiblich)               | <b>0,32</b>                  | 0,06 | <0,001 | 0,00                         | 0,14 | 0,994  | -0,02                                              | 0,13        | 0,896  | 0,02          | 0,18 | 0,908 |
| Dienstalter                             | -0,10                        | 0,06 | 0,093  | -0,05                        | 0,06 | 0,468  | <b>0,14</b>                                        | <b>0,06</b> | 0,032  | 0,04          | 0,06 | 0,575 |
| Fach: MINT                              | 0,01                         | 0,09 | 0,897  | -0,19                        | 0,11 | 0,078  | 0,21                                               | 0,16        | 0,185  | <b>-0,36</b>  | 0,13 | 0,007 |
| Fach: Sprache                           | 0,04                         | 0,11 | 0,694  | -0,17                        | 0,23 | 0,451  | 0,41                                               | 0,25        | 0,104  | -0,31         | 0,23 | 0,172 |
| Schulform (1 = GYM)                     | <b>0,39</b>                  | 0,12 | 0,001  | 0,01                         | 0,14 | 0,939  | -0,15                                              | 0,16        | 0,348  | -0,14         | 0,12 | 0,248 |
| Bundesländer (1 = Ostdeutschland)       | -0,18                        | 0,17 | 0,272  | <b>0,36</b>                  | 0,17 | 0,035  | 0,16                                               | 0,28        | 0,564  | 0,22          | 0,17 | 0,189 |
| Extraversion                            | 0,03                         | 0,15 | 0,830  | -0,05                        | 0,14 | 0,743  | -0,19                                              | 0,16        | 0,236  | -0,35         | 0,25 | 0,154 |
| Verträglichkeit                         | 0,21                         | 0,16 | 0,194  | 0,10                         | 0,09 | 0,284  | 0,07                                               | 0,13        | 0,608  | -0,03         | 0,09 | 0,750 |
| Gewissenhaftigkeit                      | 0,05                         | 0,07 | 0,499  | 0,15                         | 0,08 | 0,050  | <b>0,19</b>                                        | 0,09        | 0,029  | -0,06         | 0,11 | 0,610 |
| Neurotizismus                           | -0,02                        | 0,07 | 0,767  | 0,06                         | 0,10 | 0,525  | 0,15                                               | 0,08        | 0,055  | -0,13         | 0,16 | 0,413 |
| Offenheit                               | -0,03                        | 0,24 | 0,897  | 0,06                         | 0,13 | 0,666  | 0,11                                               | 0,22        | 0,605  | <b>0,54</b>   | 0,27 | 0,044 |
| Gewichtung mündliche Note in %          | 0,08                         | 0,07 | 0,286  | 0,11                         | 0,07 | 0,120  | <b>0,18</b>                                        | 0,05        | 0,001  | 0,00          | 0,07 | 0,973 |
| Häufigkeit Notizen zu SuS               | -0,07                        | 0,08 | 0,351  | 0,10                         | 0,08 | 0,203  | <b>0,18</b>                                        | 0,04        | <0,001 | <b>0,21</b>   | 0,10 | 0,033 |
| Beurteilung mündl. Leistungen schwierig | <b>-0,08</b>                 | 0,04 | 0,038  | -0,02                        | 0,04 | 0,717  | 0,02                                               | 0,04        | 0,549  | -0,07         | 0,07 | 0,310 |
| Bezugsnorm: Kriterium                   | <b>0,28</b>                  | 0,05 | <0,001 | -0,05                        | 0,05 | 0,344  | <b>0,18</b>                                        | 0,05        | 0,001  | 0,09          | 0,08 | 0,278 |
| Bezugsnorm: Sozial                      | <b>0,20</b>                  | 0,07 | 0,008  | <b>0,27</b>                  | 0,06 | <0,001 | <b>0,17</b>                                        | 0,05        | <0,001 | -0,10         | 0,11 | 0,351 |
| Bezugsnorm: Individuell                 | 0,08                         | 0,09 | 0,417  | <b>0,32</b>                  | 0,06 | <0,001 | 0,06                                               | 0,07        | 0,370  | <b>0,37</b>   | 0,13 | 0,005 |
| R <sup>2</sup>                          | 0,22                         |      |        | 0,23                         |      |        | 0,20                                               |             |        | 0,28          |      |       |

Alle kontinuierlichen Variablen wurden z-standardisiert. Alle Items wurden so rekodiert, dass höhere Ausprägungen größeren Werten entsprechen. Referenzgruppe bei den Fach-Indikatoren (MINT und Sprache) bildete die Kategorie Geistes-, Sozialwissenschaft oder Anderes Fach. Werte  $p < 0,05$  wurden fett markiert.

Für den Faktor Gruppenarbeit zeigte sich schließlich ein negativer Zusammenhang mit dem MINT-Indikator, der nahelegt, dass Lehrpersonen in MINT-Fächern Gruppenarbeiten im Zusammenhang mit der Bildung der mündlichen Note als weniger wichtig einschätzen als die Referenzgruppe (Lehrkräfte in geistes- und sozialwissenschaftlichen oder anderen Fächern). Darüber hinaus fand sich ein positiver Zusammenhang mit der Offenheit von Lehrkräften, der Häufigkeit der Notizen von Lehrpersonen und der individuellen Bezugsnormorientierung. Unterschiede zwischen Lehrpersonen hinsichtlich der verschiedenen Faktoren konnten zu 20 %–28 % durch die berücksichtigten Variablen erklärt werden.

## 5.2 Fragekomplex 2: Kriteriale Präferenzen von Lehrpersonen und Benotungstypen

Zur Untersuchung von Fragestellung 2 nach Benotungstypen von Lehrpersonen wurden zunächst unterschiedliche Modelle mit 3 bis 5 latenten Klassen spezifiziert. Die Fits für diese Modelle sind in Appendix A3 dargestellt. Wie ersichtlich zeigten sich zwischen den Fits der Modelle nur geringfügige Unterschiede. Betrachtet man die verschiedenen Lösungen genauer, so ließ sich in der 3-Klassenlösung zunächst eine Gruppe von Lehrpersonen identifizieren (Gruppe 1: 48,6 %), die nahezu alle Kriterien als wichtig oder sehr wichtig beurteilte. Bezüglich des absoluten Niveaus, zeigten sich bei dieser Gruppe für alle Kriterien bedingte Antwortwahrscheinlichkeiten  $\geq 0,63$  (vgl. Abb. 2). Der Mittelwert der bedingten Antwortwahrscheinlichkeiten lag bei 0,86. Diese Gruppe wird im Folgenden als „All In-Gruppe“ bezeichnet.



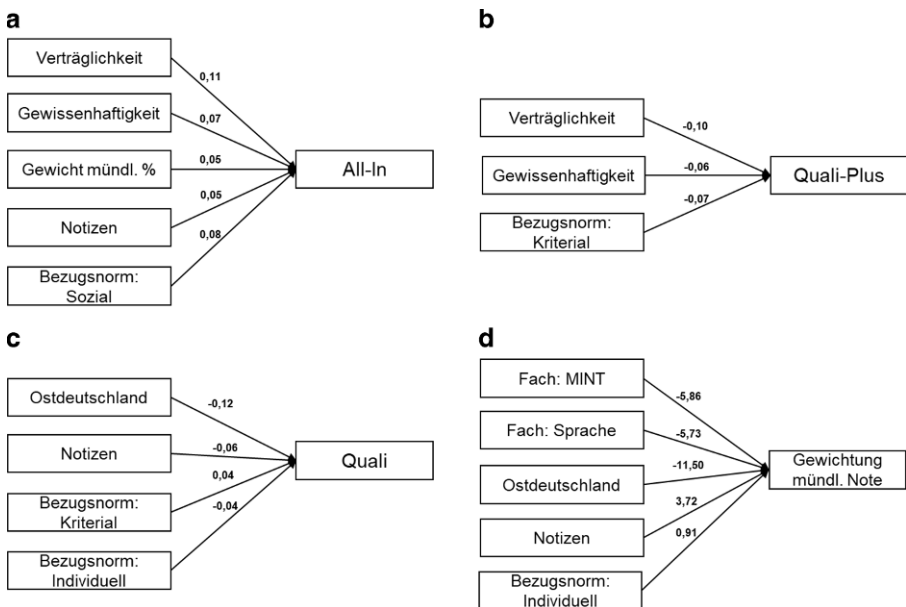
**Abb. 2** Ergebnisse der latenten Klassenanalyse mit drei Gruppen. Die Reihung der Kriterien erfolgte absteigend gemäß der mittleren Gewichtung durch die Lehrkräfte (vgl. Abb. 1)

**Tab. 2** Vorhersage der individuellen Klassenwahrscheinlichkeiten durch eine Multiple Regression

| Abhängige Variable                      | Gruppe 1: „All In“ |      |       |              | Gruppe 2: „Quali-Plus“ |         |              |      | Gruppe 3: „Quali“ |      |       |   |
|-----------------------------------------|--------------------|------|-------|--------------|------------------------|---------|--------------|------|-------------------|------|-------|---|
|                                         | B                  | SE   | p     | B            | SE                     | p       | B            | SE   | B                 | SE   | p     | p |
| Prädiktor                               |                    |      |       |              |                        |         |              |      |                   |      |       |   |
| Geschlecht (1 = weiblich)               | 0,04               | 0,06 | 0,527 | -0,05        | 0,06                   | 0,346   | 0,02         | 0,02 | 0,02              | 0,02 | 0,475 |   |
| Dienstalter                             | 0,00               | 0,02 | 0,851 | -0,02        | 0,02                   | 0,259   | 0,02         | 0,02 | 0,02              | 0,02 | 0,411 |   |
| Fach: MINT                              | -0,02              | 0,04 | 0,685 | 0,04         | 0,05                   | 0,454   | -0,02        | 0,04 | -0,02             | 0,04 | 0,577 |   |
| Fach: Sprache                           | 0,02               | 0,04 | 0,592 | 0,04         | 0,04                   | 0,385   | -0,06        | 0,05 | -0,06             | 0,05 | 0,279 |   |
| Schulform (1 = GYM)                     | 0,02               | 0,03 | 0,574 | -0,06        | 0,04                   | 0,107   | 0,04         | 0,04 | 0,04              | 0,04 | 0,265 |   |
| Bundesländer (1 = Ostdeutschland)       | 0,11               | 0,10 | 0,311 | 0,01         | 0,09                   | 0,885   | -0,12        | 0,06 | 0,05              | 0,06 | 0,036 |   |
| Extraversion                            | -0,05              | 0,06 | 0,335 | 0,00         | 0,03                   | 0,968   | 0,05         | 0,03 | 0,05              | 0,03 | 0,119 |   |
| Verträglichkeit                         | <b>0,11</b>        | 0,03 | 0,001 | <b>-0,10</b> | 0,04                   | 0,003   | -0,01        | 0,03 | -0,01             | 0,03 | 0,749 |   |
| Gewissenhaftigkeit                      | <b>0,07</b>        | 0,03 | 0,004 | <b>-0,06</b> | 0,03                   | 0,027   | -0,01        | 0,03 | -0,01             | 0,03 | 0,629 |   |
| Neurotizismus                           | 0,04               | 0,05 | 0,389 | -0,02        | 0,04                   | 0,714   | -0,03        | 0,03 | -0,03             | 0,03 | 0,302 |   |
| Offenheit                               | 0,05               | 0,04 | 0,247 | 0,01         | 0,07                   | 0,904   | -0,05        | 0,06 | -0,05             | 0,06 | 0,339 |   |
| Gewichtung mündliche Note in %          | <b>0,05</b>        | 0,02 | 0,013 | 0,00         | 0,02                   | 0,834   | -0,05        | 0,03 | -0,05             | 0,03 | 0,084 |   |
| Häufigkeit Notizen zu SuS               | <b>0,05</b>        | 0,02 | 0,012 | 0,01         | 0,03                   | 0,612   | <b>-0,06</b> | 0,03 | <b>-0,06</b>      | 0,03 | 0,034 |   |
| Beurteilung mündl. Leistungen schwierig | 0,00               | 0,02 | 0,938 | -0,02        | 0,02                   | 0,365   | 0,02         | 0,02 | 0,02              | 0,02 | 0,174 |   |
| Bezugsnorm: Kriterial                   | 0,03               | 0,02 | 0,193 | <b>-0,07</b> | 0,02                   | < 0,001 | 0,04         | 0,01 | 0,04              | 0,01 | 0,001 |   |
| Bezugsnorm: Sozial                      | <b>0,08</b>        | 0,03 | 0,012 | -0,05        | 0,03                   | 0,123   | -0,03        | 0,01 | -0,03             | 0,01 | 0,059 |   |
| Bezugsnorm: Individuell                 | 0,05               | 0,03 | 0,162 | 0,00         | 0,03                   | 0,928   | <b>-0,04</b> | 0,02 | <b>-0,04</b>      | 0,02 | 0,003 |   |
| R <sup>2</sup>                          | 0,16               |      |       | 0,11         |                        |         | 0,13         |      |                   |      |       |   |

*Anmerkung.* Die abhängige Variable wurde zur besseren Interpretierbarkeit nicht standardisiert, alle übrigen kontinuierlichen Variablen wurden z-standardisiert. Alle Items wurden so rekodiert, dass höhere Ausprägungen größeren Werten entsprechen. Referenzgruppe bei den Fach-Indikatoren (MINT und Sprache) bildete die Kategorie Geistes-, Sozialwissenschaft oder Anderes Fach. Werte  $p < 0,05$  wurden fett markiert.

Weniger relevant sind für diese Lehrpersonen lediglich die Heftführung, keine Störungen und die Leistung in schriftlichen Tests. Darüber hinaus fand sich eine weitere Gruppe von Lehrkräften (Gruppe 3: 18,3 %), für die überwiegend Qualitätsmerkmale mündlicher Leistungen relevant waren. Hierbei handelte es sich insbesondere um die Qualität von Antworten, die Qualität von Präsentationen und die Qualität von Unterrichtsgesprächen. Der Mittelwert der bedingten Antwortwahrscheinlichkeiten dieser Gruppe lag bei 0,42. Diese Gruppe wird im Folgenden als „Quali-Gruppe“ bezeichnet. Sie unterschied sich in Bezug auf die bedingten Antwortwahrscheinlichkeiten am deutlichsten von Gruppe 1. Schließlich ließ sich mit Gruppe 2 (33,1 %) eine weitere Gruppe identifizieren, für die einerseits Qualitätskriterien mündlicher Leistungen besonders wichtig waren (vergleichbar zu Gruppe 3), gleichzeitig aber auch Aspekte des Arbeits- und Sozialverhaltens, insbesondere die Selbstständigkeit, Aufmerksamkeit und Leistungsbereitschaft. Der Mittelwert der bedingten Antwortwahrscheinlichkeiten belief sich in dieser Gruppe auf 0,62. Diese Gruppe wird im Folgenden als „Quali-Plus“-Gruppe bezeichnet. Bei einer näheren Betrachtung der Ergebnisse aus Modellen mit vier oder fünf Klassen zeigte sich, dass hier keine neuen, klar interpretierbaren Lösungen resultierten, obwohl die Modellfits vergleichbar oder nur minimal besser ausfielen als bei der 3-Klassen-Lösung.



**Abb. 3** Grafische Darstellung der statistisch signifikanten Variablen aus den Regressionsmodellen zur Vorhersage der individuellen Klassenwahrscheinlichkeiten (a–c; Frage 2b) sowie zur prozentualen Gewichtung der mündlichen Note (d; Frage 3b). Die Teilgrafiken a–c resultierten aus Analysen zur Beantwortung von Fragestellung 2b (vgl. Tab. 2) zur Vorhersage individueller Klassenwahrscheinlichkeiten. Teilgrafik d visualisiert die Ergebnisse der Vorhersage der prozentualen Gewichtung der mündlichen Note an der Gesamtnote zur Beantwortung von Fragestellung 3b aus Tab. 4. Es werden jeweils unstandardisierte und nur die statistisch signifikanten Koeffizienten  $p < 0,05$  berichtet. Teilgrafik d: Referenzgruppe bei den Fach-Indikatoren (MINT und Sprache) bildete die Kategorie Geistes-, Sozialwissenschaft oder Anderes Fach

Bei der Beantwortung von Fragestellung 2b zu den Zusammenhängen von Lehrpersonenmerkmalen und Benotungs-Typen (vgl. Tab. 2 und Abb. 3) zeigte sich, dass die Wahrscheinlichkeit einer Zuordnung zur All In-Gruppe positiv mit der Verträglichkeit und Gewissenhaftigkeit von Lehrpersonen, der Gewichtung der mündlichen Note, der Häufigkeit der Notizen und der sozialen Bezugsnorm assoziiert war. Dies legt nahe, dass Lehrpersonen mit höheren Werten auf diesen Variablen tendenziell eher viele Kriterien bei der Bildung mündlicher Noten verwenden.

Lehrpersonen, die angaben, in einem östlichen Bundesland zu leben, die sich häufiger Notizen zu SchülerInnen machten und die individuelle Bezugsnorm als wichtiger beurteilten, hatten eine geringere Wahrscheinlichkeit, der Quali-Gruppe zugeordnet zu werden. Lediglich die Wichtigkeit kriterialer Bezugsnormen war positiv mit der Wahrscheinlichkeit assoziiert, zur Quali-Gruppe zugeordnet zu werden.

Mit der höheren Wahrscheinlichkeit einer Zuordnung zur Quali-Plus-Gruppe gingen schließlich eine geringe Verträglichkeit und Gewissenhaftigkeit sowie eine geringere Wichtigkeit kriterialer Bezugsnormen von Lehrpersonen einher. Unterschiede in den Zuordnungswahrscheinlichkeiten zu den drei Gruppen konnten zu 11 %–16 % durch die berücksichtigten Kovariaten erklärt werden (vgl. Tab. 2 und Abb. 3).

### 5.3 Fragekomplex 3: Gewichtung der mündlichen Note

In Tab. 3 finden sich die Ergebnisse zu 3a nach der Gewichtung der mündlichen Note bei der Bildung einer Gesamtnote. Im Mittel lag die prozentuale Gewichtung bei 47,4 %. Zur Beantwortung von Fragestellung 3b zu möglichen Gewichtungsunterschieden zwischen Lehrpersonen wurde die prozentuale Gewichtungsvariable im nächsten Schritt durch die verschiedenen Variablen (vgl. Abschn. 5.2) vorhergesagt. Die Ergebnisse sind in Tab. 4 und Abb. 3 dargestellt. Wie ersichtlich zeigten sich statistisch signifikant negative Assoziationen für MINT-Fächer und Sprachen (im Vergleich zu geistes- und sozialwissenschaftlichen und anderen Fächern), die nahelegen, dass die mündliche Note hier (jeweils) weniger stark gewichtet wurde. Zudem zeigte sich, dass Lehrpersonen aus Ostdeutschland mündliche Noten im Mittel weniger stark gewichteten. Das Gewicht der mündlichen Note war darüber hinaus positiv mit der Häufigkeit der Notizen zu den mündlichen Leistungen der SchülerInnen und der Bedeutung der individuellen Bezugsnorm für Lehrpersonen assoziiert. Das Modell erklärte insgesamt 27 % der Unterschiede in der prozentualen Gewichtung der mündlichen Note zwischen Lehrpersonen.

**Tab. 3** Absolute und prozentuale Gewichtung der mündlichen Note bei der Bildung der Gesamtnote

| Gewicht an der Gesamtnote | Bis 10 % | 11–20 % | 21–30 % | 31–40 % | 41–49 %     | 50 % |
|---------------------------|----------|---------|---------|---------|-------------|------|
| Anteil absolut            | 11       | 25      | 72      | 108     | 51          | 161  |
| Anteil in %               | 2        | 4       | 12      | 18      | 8           | 26   |
| Gewicht an der Gesamtnote | 51–60 %  | 61–70 % | 71–80 % | 81–90 % | Größer 90 % |      |
| Anteil absolut            | 93       | 60      | 15      | 2       | 13          |      |
| Anteil in %               | 15       | 10      | 3       | 0       | 2           |      |

*Anmerkung.* Bei der Berechnung wurden Stichprobengewichte berücksichtigt.

**Tab. 4** Vorhersage der prozentualen Gewichtung der mündlichen Note

| Abhängige Variable                      | Gewichtung der mündlichen Note |               |      |        |
|-----------------------------------------|--------------------------------|---------------|------|--------|
|                                         | $\beta$                        | B             | SE   | p      |
| Prädiktor                               |                                |               |      |        |
| Geschlecht (1 = weiblich)               | -0,17                          | -2,87         | 1,56 | 0,066  |
| Dienstalter                             | 0,00                           | 0,01          | 0,09 | 0,946  |
| Fach: MINT                              | <b>-0,35</b>                   | <b>-5,86</b>  | 0,91 | <0,001 |
| Fach: Sprache                           | <b>-0,34</b>                   | <b>-5,73</b>  | 2,28 | 0,012  |
| Schulform (1 = GYM)                     | 0,13                           | 2,22          | 1,88 | 0,236  |
| Bundesländer (1 = Ostdeutschland)       | <b>-0,67</b>                   | <b>-11,50</b> | 3,51 | 0,001  |
| Extraversion                            | 0,08                           | 0,95          | 0,79 | 0,233  |
| Verträglichkeit                         | 0,07                           | 0,97          | 1,15 | 0,398  |
| Gewissenhaftigkeit                      | 0,00                           | 0,01          | 0,96 | 0,990  |
| Neurotizismus                           | 0,02                           | 0,28          | 0,84 | 0,738  |
| Offenheit                               | -0,10                          | -0,90         | 1,47 | 0,538  |
| Häufigkeit Notizen zu SuS               | <b>0,32</b>                    | <b>3,72</b>   | 0,38 | <0,001 |
| Beurteilung mündl. Leistungen schwierig | -0,04                          | -0,73         | 0,61 | 0,232  |
| Bezugsnorm: Kriterium                   | -0,08                          | -0,99         | 0,70 | 0,156  |
| Bezugsnorm: Sozial                      | 0,06                           | 0,71          | 0,65 | 0,272  |
| Bezugsnorm: Individuell                 | <b>0,07</b>                    | <b>0,91</b>   | 0,44 | 0,040  |
| R <sup>2</sup>                          | 0,27                           |               |      |        |

$\beta$  = Alle kontinuierlichen Variablen wurden z-standardisiert. B = Unstandardisierte Koeffizienten (Angaben in Prozent). Alle Items wurden so rekodiert, dass höhere Ausprägungen größeren Werten entsprechen. Referenzgruppe bei den Fach-Indikatoren (MINT und Sprache) bildete die Kategorie Geistes-, Sozialwissenschaft oder Anderes Fach. Werte  $p < 0,05$  wurden fett markiert.

## 6 Diskussion

In Bezug auf die übergeordneten Fragestellungen legen die Analysen ein differenziertes Bild nahe. So zeigte sich beispielsweise im Hinblick auf die verschiedenen Kriterien zur Bildung mündlicher Noten, dass ein Großteil der Lehrkräfte angab, sehr viele verschiedene Kriterien zu berücksichtigen. Ferner legen die Ergebnisse der Studie nahe, dass sich die Kriterien mittels vier übergeordneter Dimensionen zusammenfassen ließen, die sich auf die Qualität mündlicher Beiträge, auf Aspekte des Arbeits- und Sozialverhaltens, auf die Qualität schriftlicher Leistungen und Hausaufgaben und auf die Qualität von Gruppenarbeiten beziehen. Interessanterweise zeigte sich, dass verschiedene LehrerInnenmerkmale (insb. die Bezugsnormorientierung), aber auch demographische Variablen wie die Schulform und das Geschlecht, mit einzelnen der genannten Dimensionen assoziiert waren. Persönlichkeitsmerkmale standen lediglich mit der Qualität schriftlicher Leistungen und Hausaufgaben (Gewissenhaftigkeit) und mit dem Faktor Gruppenarbeit (Offenheit) in einem bedeutsamen Zusammenhang.

Im Rahmen der latenten Klassenanalysen konnten anschließend drei Benotungstypen identifiziert werden: Lehrpersonen, die (nahezu) alle Kriterien (49 %; All In-Gruppe) für wichtig hielten, Lehrpersonen, die primär Qualitätskriterien mündlicher Leistungen (18 %; Quali-Gruppe) berücksichtigten, und Lehrpersonen, die Qualitätskriterien mündlicher Leistungen und Aspekte des Arbeits- und Sozialverhaltens

(33 %; Quali-Plus-Gruppe) als wichtig einschätzten. In Übereinstimmung mit den in der Theorie und Fragestellung postulierten Überlegungen zeigte sich, dass verträglichere und gewissenhaftere Lehrpersonen häufiger der All In-Gruppe und weniger häufig der Quali-Plus-Gruppe zugeordnet wurden. Interessant erscheint, dass die soziale Bezugsnormorientierung positiv mit der Zugehörigkeit zur All In-Gruppe assoziiert war, eine stärkere kriteriale Orientierung hingegen im positiven Zusammenhang mit der Zugehörigkeit zur Quali-Gruppe stand.

Schließlich ergaben die Analysen zur Gewichtung mündlicher Noten, dass lediglich die individuelle Bezugsnormorientierung, nicht jedoch die Persönlichkeitsmerkmale der Lehrkräfte mit einer höheren Gewichtung mündlicher Noten assoziiert waren. Wie im Theorieteil näher beschrieben, könnte dieser Befund darauf zurückzuführen sein, dass Lehrpersonen mit einer stärkeren Individual-Bezugsnormorientierung mündliche Leistungen und Noten stärker gewichten, um ihrem Anspruch einer möglichst umfangreichen Abbildung vielfältiger Entwicklungsprozesse in der Gesamtnote gerecht zu werden. Für Lehrpersonen aus den neuen Bundesländern zeigte sich, bei statistischer Konstanzhaltung der übrigen Variablen im Modell (vgl. Tab. 4), dass diese mündliche Noten geringer gewichteten als Lehrpersonen aus den alten Bundesländern.

## 7 Limitationen

Zunächst ist es wichtig, darauf hinzuweisen, dass die Studienergebnisse auf selbstberichteten Angaben der Lehrkräfte basieren. Inwieweit die Angaben der Lehrkräfte den Prozess der mündlichen Notenbildung in der Praxis tatsächlich adäquat abbilden, lässt sich nicht abschließend beantworten. Es ist beispielsweise nicht auszuschließen, dass die Antworten der Lehrkräfte zu einem gewissen Grad sozial erwünscht erfolgten. Gerade beim Thema der Notenvergabe, das immer einer gewissen Spannung im Kontext Schule unterliegt, erscheint es nicht unplausibel, dass manche Lehrkräfte fälschlicherweise (und unbewusst) mehr Kriterien als wichtig oder eher wichtig einstufen, als sie dies tatsächlich in ihrer beruflichen Praxis tun, die auch von einer Vielzahl notwendiger Bewertungsprozesse geprägt ist (Lüders 2001). Zudem könnte ein Instrument mit vielen Antwortoptionen (insgesamt 16) bei Lehrkräften, die weniger Kriterien berücksichtigen, den Eindruck erweckt haben, dass sie mehr Kriterien berücksichtigen sollten/müssen. Dies könnte letztlich dazu geführt haben, dass der Anteil der befragten Lehrpersonen, die nahezu alle Kriterien als wichtig einstufen (48,6 %), größer ausgefallen ist. Gleichzeitig sollte auch erwähnt werden, dass die Untersuchung individuell (am Computer) und vollständig anonymisiert stattfand, was dem Druck, sozial erwünscht zu antworten, entgegengewirkt haben dürfte, dies jedoch nicht ausschließt.

Im Rahmen der vorliegenden Studie wurde darüber hinaus ein Instrument verwendet, das sich vorab in verschiedenen qualitativen und quantitativen Entwicklungsstadien bewähren musste, um eine inhaltlich adäquate Abdeckung der (für Lehrkräfte) bedeutsamsten Kriterien sicherzustellen. Allerdings haben verschiedene Beiträge darauf hingewiesen, dass (selbst im Falle einer genauen Erfassung der Kriterien) verschiedene Beobachtungs- und Beurteilungsfehler Notenbildungsprozesse beein-

flussen können (Beutel und Pant 2020), die auch die Ergebnisse der vorliegenden Studie beeinflusst haben könnten. Beispielsweise könnten Lehrkräfte einzelnen Faktoren ein unterschiedliches Gewicht gegeben haben (z. B. Quantität und Qualität), die sich in der schulischen Praxis aber oft nur schwer voneinander trennen lassen. Somit verbleibt das Desiderat, Prozesse der Bildung mündlicher Noten und darauf basierender Gesamtnoten in zukünftigen Studien noch genauer zu erfassen, auch mit stärker qualitativen Zugängen und in authentischen Situationen, beispielsweise bei der Festlegung von Zeugnisnoten (z. B. mittels Vignettenstudien).

Eine weitere wichtige Einschränkung besteht darin, dass in der vorliegenden Studie lediglich ein Ausschnitt aller möglicherweise für mündliche Notenbildungsprozesse relevanten Variablen berücksichtigt wurde. Nicht auszuschließen ist, dass weitere Variablen existieren, die kriteriale Präferenzen, die Wahrscheinlichkeit, einem bestimmten Benotungs-Typ zugeordnet zu werden, oder die Gewichtung der mündlichen Note bei der Bildung einer Gesamtnote erklären können. Hierzu könnte beispielsweise das Professionswissen zählen, ebenso wie professionelle Werte und Überzeugungen sowie spezifische diagnostische Kompetenzen von Lehrpersonen. Vor diesem Hintergrund und der insgesamt auf Grund des Forschungsstands zu mündlichen Noten eher exploratorischen Anlage der Studie sind weitere Studien notwendig, die die Bedeutsamkeit dieser weiteren möglicherweise relevanten Konstrukte näher untersuchen.

Wie in nahezu jeder sozialwissenschaftlichen Befragung unterlag auch die in dieser Studie durchgeführte repräsentative LehrerInnenbefragung zeitlichen Restriktionen. So war es im Rahmen der vorgegebenen Befragungszeit (insgesamt circa 20 min) nicht möglich, ein Instrument zu administrieren, das Items zu allen möglichen Kombinationen aus Leistungsformen und Bewertungskriterien beinhaltete. Diese zeitlichen Restriktionen wurden bereits im Rahmen des Entwicklungsprozesses des Instruments berücksichtigt, dessen übergeordnetes Ziel darin bestand, ein Fragebogeninstrument zu entwickeln, das möglichst sparsam ist und gleichzeitig die (für Lehrkräfte) relevantesten Kriterien enthält.

Dies stellt insofern eine mögliche Limitation der vorliegenden Studie dar, als dass die verwendeten Kategorien zwar offensichtlich für die im Zuge der Entwicklung des Instruments befragten Lehrpersonen relevant waren, aber nicht zu allen Leistungsform  $\times$  Bewertungskriterium-Kombinationen Informationen erfasst werden konnten. Es wäre daher perspektivisch sehr wünschenswert, wenn zukünftige Studien die bestehenden Instrumente sichten (vgl. z. B. auch Krüll 2023) und ggf. erweitern. Gleichzeitig sollte in diesen Arbeiten auch näher untersucht werden, ob Lehrkräfte in der Praxis tatsächlich für alle Leistungs- und Sozialformen derart differenziert Bewertungen vornehmen oder sie nicht teilweise eher gröbere Kategorien beurteilen, ohne diese Beurteilungen hinsichtlich qualitativer und quantitativer oder verbaler und schriftlicher Leistungen näher aufzuschlüsseln (z. B. eher die Kategorie „Gruppenarbeit“ anstatt Kategorien wie „Quantität mündlicher Leistungen bei Gruppenarbeiten“ und „Qualität mündlicher Leistungen bei Gruppenarbeiten“, etc.).

Auch methodisch ergeben sich einige wichtige Limitationen, die bei der Interpretation der Ergebnisse berücksichtigt werden sollten. Auf Grund der zeitlichen Restriktionen der Studie konnten verschiedene Konstrukte (z. B. die Bezugsnormorientierung und die Persönlichkeit) nur mit sehr kurzen Skalen, im Falle der Be-

zugsnormorientierung lediglich mit je einem Item, erfasst werden. Auch wenn sich beispielsweise die Kurzskala zur Erfassung der Persönlichkeit in vorherigen Studien bewährt hat (Rammstedt et al. 2020; Soto und John 2017), zeigten sich diesbezüglich zum Teil geringe Reliabilitätswerte (z. B. bei der Offenheit). Um dieser Limitation zu begegnen, wurde zur Kontrolle des Messfehlers in allen Modellen ein Single-Indicator-Approach verwendet. Hierbei wird die Residualvarianz der Persönlichkeitsindikatoren auf den Wert  $(1 - \text{Reliabilität}) \cdot \text{Stichprobenvarianz}$  fixiert (Hoyle 2012; Hübner et al. 2022a). Trotz der Anwendung dieses etablierten Verfahrens ist auf Grund der geringen Itemanzahl pro Persönlichkeitsdimension nicht auszuschließen, dass die genannten Konstrukte nicht in ihrer vollen Breite erfasst werden konnten. Zur genaueren Quantifizierung des Ausmaßes dieser Limitation sind daher weitere Studien notwendig, die auch umfangreichere Instrumentarien bei Lehrpersonen einsetzen und die Ergebnisse entsprechend kontrastieren können.

## 7.1 Einordnungen, Implikationen und Ausblick

Die vorliegende Studie bietet einen umfassenden ersten Einblick in den bisher weitgehend unerforschten Gegenstandsbereich der mündlichen Notenbildung. Vor dem Hintergrund des bisher noch unterentwickelten Forschungsstands sollen einige besonders bedeutsame Ergebnisse abschließend diskutiert werden.

Erwähnenswert erscheint zunächst ein Befund, der sich mit vorliegenden Ergebnissen aus Krüll (2023) deckt: Ein substantieller Anteil der befragten Lehrpersonen in der vorliegenden Studie (94 %) gab an, dass die Qualität mündlicher Beiträge im Unterrichtsgespräch ein wichtiges Kriterium bei der mündlichen Notenbildung darstellt (vgl. Abb. 1). Die besondere Bedeutung dieses Items für den Faktor Qualität mündlicher Leistungen zeigt sich auch an der vergleichsweise hohen Faktorladung in Tab. A2 im Anhang. Interessant ist in diesem Zusammenhang auch, dass neben der Qualität auch Aspekte der Quantität von Beiträgen eine übergeordnete Rolle zu spielen scheinen. So zeigte sich, dass für einen Großteil der Lehrpersonen (83 %) auch die Häufigkeit der Beteiligung an Unterrichtsgesprächen einen hohen Stellenwert für die mündliche Notengebung besitzt, und zwar als eine Komponente des Faktors Arbeits- und Sozialverhalten. Diese Ergebnisse erweitern somit Befunde von Kohler et al. (2025), die große Unterschiede in der Ausgestaltung und Spezifität rechtlicher Regelungen zur Beurteilung mündlicher Leistungen zwischen Bundesländern nahelegen: Trotz dieser großen Heterogenität in Bezug auf die rechtlichen Vorgaben deutet sich an, dass unter Lehrkräften auch länderübergreifend eine recht große Einigkeit darüber besteht, dass Qualitätskriterien der mündlichen Leistungen und das Arbeits- und Sozialverhalten wichtige Pfeiler der mündlichen Note darstellen. Die Bedeutung dieser beiden Faktoren verdeutlichte sich noch einmal anhand der Ergebnisse der latenten Klassenanalysen.

Diese Befunde sind für sich genommen bereits interessant, könnten darüber hinaus aber auch dabei helfen, Ergebnisse von Studien zum Zusammenhang der Persönlichkeitsmerkmale von SchülerInnen mit ihren Noten noch besser zu verstehen (z. B. Becker 2024; Brandt et al. 2021; Hübner et al. 2022a; Meyer et al. 2023). Neben den direkten positiven Effekten von Persönlichkeitsmerkmalen wie der Gewissenhaftigkeit auf die Leistungen von SchülerInnen, könnten positive Zusammen-

hänge zwischen der Gewissenhaftigkeit und den Zeugnisnoten von SchülerInnen im Wesentlichen auf Effekte der mündlichen Note und hier Facetten des Arbeits- und Sozialverhaltens zurückzuführen sein. Dieser Logik folgend würde sich eine hohe Gewissenhaftigkeit also gleichzeitig über mehrere Kanäle – sowohl über die Leistung selbst als auch über die Leistungsbewertung – positiv auf die Noten von SchülerInnen auswirken.

Schließlich zeigte sich in Bezug auf die Überlegungen zur Bedeutung von Persönlichkeitsmerkmalen und der Bezugsnormorientierung von Lehrpersonen für die unterschiedlichen Kriterien und Gewichtung ein gemischtes Bild. Von Bedeutung zu sein scheinen die beiden Konstrukte insbesondere bei der Vorhersage der übergeordneten Dimensionen der kriterialen Präferenzen, ebenso wie bei der Vorhersage der Benotungs-Typen. Nicht statistisch bedeutsam waren Persönlichkeitsmerkmale bei der Gewichtung mündlicher Noten. Dies könnte unterschiedliche Gründe haben. Einerseits könnten Persönlichkeitsmerkmale von Lehrpersonen tatsächlich ein zu unspezifisches Maß darstellen, um starke direkte Auswirkungen auf die Prozesse der mündlichen Notenbildung zu haben. So könnte eine hoch verträgliche Lehrkraft zwar als Person höher verträglich sein als andere Lehrkräfte, aber im Beurteilungsprozess selbst dennoch zu vergleichbaren Urteilen gelangen, weil sich beide Lehrkräfte hinsichtlich bestimmter professioneller Werte und Überzeugungen (z. B. Vorstellungen über Lern- und Leistungsstandards) ähneln, die relevanter für die Beurteilung von SchülerInnen sind. Darüber hinaus legen verschiedene Arbeiten nahe, dass Lehrkräfte zwar möglicherweise Präferenzen für eine Bezugsnorm haben (Köller 2005; Rheinberg und Fries 2018), jedoch typischerweise verschiedene Bezugsnormen bei der Beurteilung von SchülerInnen simultan verwenden (Hochweber et al. 2014; Hübner et al. 2020). Hierdurch könnte der Einfluss einzelner Bezugsnormen, z. B. bei der Gewichtung der mündlichen Noten, geringer ausfallen. Perspektivisch sind weitere Studien notwendig, um zentrale schülerInnen- und lehrerInnenseitige Einflussfaktoren auf schriftliche und mündliche Notenbildungsprozesse noch genauer herauszuarbeiten.

Auf Basis der Befunde lassen sich darüber hinaus eine Reihe von Implikationen für die Praxis formulieren, die jedoch vor dem Hintergrund des explorativen Charakters der Studie und den nicht als kausal einzustufenden Ergebnissen entsprechend vorsichtig interpretiert werden sollten. Zunächst legen die Befunde nahe, dass alle Lehrkräfte Qualitätskriterien mündlicher Leistungen für wichtig erachten und auch in Bezug auf einzelne Kriterien des Sozialverhaltens (z. B. Leistungsbereitschaft) scheint große Einigkeit in Bezug auf ihre Wichtigkeit zu bestehen. Für schulische Bewertungs- und Notenkulturen deuten diese Befunde an, dass es trotz der in vielen Studien dargestellten großen Variation zwischen Noten bei gleichen Schülerleistungen (z. B. Hübner et al. 2024; Ingenkamp 1971) eine bedeutsame Schnittmenge in Bezug auf die Einschätzung der Wichtigkeit bestimmter Beurteilungskriterien zu geben scheint (insb. in Bezug auf das Kriterium der Qualität). Dies könnte nahelegen, dass zentrale Ursachen für die oft berichteten Unterschiede von Schulnoten trotz eigentlich vergleichbarer SchülerInnenleistungen eher auf eine unterschiedliche Auslegung und Erfassung der betreffenden Kriterien zurückzuführen sind, nicht aber auf fundamental unterschiedliche Vorstellungen bezüglich der Wichtigkeit dieser Beurteilungskriterien. Der für eine stärkere Vergleichbarkeit von

Noten notwendige Aushandlungsprozess in LehrerInnenkollegien und Fachschaften, der älteren Studien folgend schulformspezifisch unterschiedlich stark ausgeprägt ist (Lüders 2001), könnte sich also stärker damit beschäftigen, die genaue Auslegung und Erfassung der jeweiligen Kriterien zu explizieren und zu besprechen. Grundsatzdebatten über die Wichtigkeit aller Kriterien scheinen für einen substanziellen Anteil dieser Kriterien nicht geführt werden zu müssen. Ergebnisse entsprechend kooperativer Abstimmungsprozesse im LehrerInnenkollegium und den Fachschaften gehen zwar möglicherweise mit einer in Bezug auf die Kriterienauswahl und -erfassung individuell wahrgenommenen geringeren Autonomie der Einzellehrkraft einher, könnten aber im Ergebnis dabei helfen, Kriterien und deren Auslegung und Erfassung besser mit SchülerInnen und Eltern zu kommunizieren, und könnten somit für eine größere Transparenz und Legitimation der Benotungskriterien sorgen. Dies könnte auch zu einer Entlastung von Lehrkräften und einer Versachlichung von oft emotional geführten Notendiskussionen mit SchülerInnen führen, weil Entscheidungen über entsprechende Bewertungskriterien und deren Erfassung auf einer anderen (höheren) schulstrukturellen Ebene verortet würden. Entsprechende Aushandlungs- und Kooperationsprozesse und die dafür erforderlichen Voraussetzungen, könnten im Rahmen von Fortbildungen und insbesondere bereits in der ersten und zweiten Phase der LehrerInnenbildung erprobt werden. Auch das Wissen zu unterschiedlichen Benotungstypen könnte in unterschiedlichen Aus- und Fortbildungsphasen systematisch genutzt werden, um diesbezüglich bereits früh Reflexionsprozesse anzustoßen und einen bewussteren Umgang mit den eigenen Vorstellungen und Überzeugungen zur Leistungsbeurteilung zu initiieren und zu fördern.

Perspektivisch wichtig erscheint eine vertiefte Untersuchung der Benotungstypen in zukünftigen Studien. Hierbei sollte die exakte Anzahl der Benotungstypen, auch vor dem Hintergrund des überschaubaren Forschungsstandes und dem daraus resultierenden explorativen Charakter der vorliegenden Studie, nicht als unumstößliche Setzung betrachtet werden, sondern vielmehr als eine Heuristik, die in Abhängigkeit von der jeweils untersuchten Stichprobe und dem eingesetzten Instrumentarium variieren kann. Besonders aufschlussreich erscheint darüber hinaus eine vertiefte Untersuchung der Begründungen der Beurteilung bestimmter Kriterien. Während die Wahl von qualitätsbezogenen Kriterien weniger erklärungsbedürftig scheint, könnten derartige vertiefte (auch qualitative) Untersuchungen insbesondere für Kriterien aufschlussreich sein, bei denen sich größere Unterschiede zeigten (z. B. den Hausaufgaben). Weiterführend wäre es auch wichtig zu untersuchen, ob sich auch bei der Vergabe von Noten für schriftliche Leistungen Unterschiede zwischen den identifizierten Benotungstypen finden lassen. So wäre es beispielsweise nicht unplausibel anzunehmen, dass Lehrpersonen in der All In-Gruppe (auch vor dem Hintergrund der Befunde zu positiven Zusammenhängen mit Gewissenhaftigkeit) auch bei schriftlichen Leistungen mehr Kriterien berücksichtigen, diese intensiver vergleichen und somit letztlich auch zu akkurateren schriftlichen und mündlichen Leistungsbeurteilungen gelangen könnten (vgl. hierzu auch Lohmann et al. 2025). Generell bildet die Untersuchung der Akkuratheit und Qualität der mündlichen Leistungsbeurteilung von Lehrpersonen ein übergeordnetes und wichtiges, wenn auch sehr herausforderndes Forschungsdesiderat, das in zukünftigen Studien angegangen werden sollte.

**Zusatzmaterial online** Zusätzliche Informationen sind in der Online-Version dieses Artikels (<https://doi.org/10.1007/s11618-026-01381-z>) enthalten.

**Förderung** Diese Arbeit wurde durch die Eberhard Karls Universität Tübingen (Projektförderung für NachwuchswissenschaftlerInnen) unterstützt. Nicolas Hübner ist Professor für Schulpädagogik mit den Schwerpunkten Bildung und Erziehung. Anfragen zum Manuskript richten Sie bitte an Nicolas Hübner, am Bonner Zentrum für Lehrer\*innenbildung, Poppelsdorfer Allee 15, 53115 Bonn. E-Mail: [nhuebner@uni-bonn.de](mailto:nhuebner@uni-bonn.de)

**Funding** Open Access funding enabled and organized by Projekt DEAL.

**Interessenkonflikt** N. Hübner, B. Kohler, M. Neumann und M. Becker geben an, dass kein Interessenkonflikt besteht.

**Open Access** Dieser Artikel wird unter der Creative Commons Namensnennung 4.0 International Lizenz veröffentlicht, welche die Nutzung, Vervielfältigung, Bearbeitung, Verbreitung und Wiedergabe in jeglichem Medium und Format erlaubt, sofern Sie den/die ursprünglichen Autor(en) und die Quelle ordnungsgemäß nennen, einen Link zur Creative Commons Lizenz beifügen und angeben, ob Änderungen vorgenommen wurden. Die in diesem Artikel enthaltenen Bilder und sonstiges Drittmaterial unterliegen ebenfalls der genannten Creative Commons Lizenz, sofern sich aus der Abbildungslegende nichts anderes ergibt. Sofern das betreffende Material nicht unter der genannten Creative Commons Lizenz steht und die betreffende Handlung nicht nach gesetzlichen Vorschriften erlaubt ist, ist für die oben aufgeführten Weiterverwendungen des Materials die Einwilligung des jeweiligen Rechteinhabers einzuholen. Weitere Details zur Lizenz entnehmen Sie bitte der Lizenzinformation auf <http://creativecommons.org/licenses/by/4.0/deed.de>.

## Literatur

- Baier, F., Decker, A., Voss, T., Kleickmann, T., Klusmann, U., & Kunter, M. (2019). What makes a good teacher? The relative importance of mathematics teachers' cognitive ability, personality, knowledge, beliefs, and motivation for instructional quality. *British Journal of Educational Psychology*, 89(4), 767–786. <https://doi.org/10.1111/bjep.12256>.
- Baumert, J., & Kunter, M. (2013). Professionelle Kompetenz von Lehrkräften. In I. Gogolin, H. Kuper, H.-H. Krüger & J. Baumert (Hrsg.), *Stichwort: Zeitschrift für Erziehungswissenschaft* (S. 277–337). Springer. [https://doi.org/10.1007/978-3-658-00908-3\\_13](https://doi.org/10.1007/978-3-658-00908-3_13).
- Baumert, J., Trautwein, U., & Artelt, C. (2003). Schulumwelten – Institutionelle Bedingungen des Lehrens und Lernens. In J. Baumert, C. Artelt, E. Klieme, M. Neubrand, M. Prenzel, U. Schiefele, W. Schneider, K.-J. Tillmann & M. Weiß (Hrsg.), *PISA 2000 – Ein differenzierter Blick auf die Länder der Bundesrepublik Deutschland* (S. 261–331). VS. [https://doi.org/10.1007/978-3-322-97590-4\\_11](https://doi.org/10.1007/978-3-322-97590-4_11).
- Becker, M. (2024). Sozioemotionale Charakteristika und Bildung. *Zeitschrift für Erziehungswissenschaft*. <https://doi.org/10.1007/s11618-024-01217-8>.
- Beutel, S.-I., & Pant, H. A. (2020). *Lernen ohne Noten. Alternative Konzepte der Leistungsbeurteilung*. Kohlhammer.
- Bohl, T. (2019). Leistungsbewertung, Notengebung und Alternativen zur Notengebung. In E. Kiel, B. Herzig, U. Maier & U. Sandfuchs (Hrsg.), *Handbuch Unterrichten an allgemeinbildenden Schulen* (S. 414–425). Klinkhardt.
- Bonefeld, M., & Dickhäuser, O. (2018). (Biased) grading of students' performance: students' names, performance level, and implicit attitudes. *Frontiers in Psychology*, 9, 481. <https://doi.org/10.3389/fpsyg.2018.00481>.
- Brandt, N. D., Becker, M., Tetzner, J., Brunner, M., & Kuhl, P. (2021). What teachers and parents can add to personality ratings of children: Unique associations with academic performance in elementary school. *European Journal of Personality*, 35(6), 814–832. <https://doi.org/10.1177/0890207020988436>.
- Bräu, K., & Schlickum, C. (2015). Die soziale Konstruktion von Leistung und Leistungsbewertung. In K. Bräu & C. Schlickum (Hrsg.), *Soziale Konstruktionen in Schule und Unterricht: Zu den Kategorien Leistung, Migration, Geschlecht, Behinderung, soziale Herkunft und deren Interdependenzen* (S. 49–64). Budrich.

- Graham, J. W. (2009). Missing data analysis: making it work in the real world. *Annual Review of Psychology*, 60, 549–576. <https://doi.org/10.1146/annurev.psych.58.110405.085530>.
- Hallquist, M. N., & Wiley, J. F. (2018). MplusAutomation: An R package for facilitating large-scale latent variable analyses in Mplus. *Structural Equation Modeling: A Multidisciplinary Journal*, 25(4), 621–638. <https://doi.org/10.1080/10705511.2017.1402334>.
- Heckhausen, H. (1974). *Leistung und Chancengleichheit*. Hogrefe.
- Hochweber, J., Hosenfeld, I., & Klieme, E. (2014). Classroom composition, classroom management, and the relationship between student attributes and grades. *Journal of Educational Psychology*, 106(1), 289–300. <https://doi.org/10.1037/a0033829>.
- Hoyle, R. H. (Hrsg.). (2012). *Handbook of structural equation modeling*. Guilford.
- Hübner, N., Wagner, W., Hochweber, J., Neumann, M., & Nagengast, B. (2020). Comparing apples and oranges: Curricular intensification reforms can change the meaning of students' grades! *Journal of Educational Psychology*, 112(1), 204–220. <https://doi.org/10.1037/edu0000351>.
- Hübner, N., Spengler, M., Nagengast, B., Borghans, L., Schils, T., & Trautwein, U. (2022a). When academic achievement (also) reflects personality: Using the personality-achievement saturation hypothesis (PASH) to explain differential associations between achievement measures and personality traits. *Journal of Educational Psychology*, 114(2), 326–345. <https://doi.org/10.1037/edu0000571>.
- Hübner, N., Wagner, W., Meyer, J., & Watt, H. M. G. (2022b). To those who have, more will be given? Effects of an instructional time reform on gender disparities in STEM subjects, stress, and health. *Frontiers in psychology*. <https://doi.org/10.3389/fpsyg.2022.816358>.
- Hübner, N., Jansen, M., Stanat, P., Bohl, T., & Wagner, W. (2024). Alles eine Frage des Bundeslandes? Eine mehr Ebenenanalytische Betrachtung der eingeschränkten Vergleichbarkeit von Schulnoten. *Zeitschrift für Erziehungswissenschaft*, 27(3), 517–549. <https://doi.org/10.1007/s11618-024-01216-9>.
- Ingenkamp, K. (1971). *Die Fragwürdigkeit der Zensurengebung: Texte und Untersuchungsberichte* (9. Aufl.). Beltz.
- John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative big five trait taxonomy: history, measurement, and conceptual issues. In O. P. John, R. W. Robins & L. A. Pervin (Hrsg.), *Handbook of personality. Theory and research* (3. Aufl., S. 114–158). Guilford.
- Jürgens, E., & Sacher, W. (2008). *Leistungserziehung und pädagogische Diagnostik in der Schule: Grundlagen und Anregungen für die Praxis*. Kohlhammer.
- Kirk, S. (2004). *Beurteilung mündlicher Leistungen: Pädagogische, psychologische, didaktische und schulrechtliche Aspekte der mündlichen Leistungsbeurteilung*. Klinkhardt.
- Klieme, E. (2022). Schlusswort. In L. Hoffmann, P. Schröter, A. Groß, S. M. Schmid-Kühn & P. Stanat (Hrsg.), *Das unvergleichliche Abitur. Entwicklungen – Herausforderungen – Empirische Analysen* (S. 319–336). wbv Media.
- KMK (2021). *Lehren und Lernen in der digitalen Welt. Ergänzung zur Strategie der Kultusministerkonferenz „Bildung in der digitalen Welt“*. KMK. [https://www.kmk.org/fileadmin/veroeffentlichungen\\_beschluesse/2021/2021\\_12\\_09-Lehren-und-Lernen-Digi.pdf](https://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2021/2021_12_09-Lehren-und-Lernen-Digi.pdf)
- Kohler, B., Bohl, T., & Hübner, N. (2025). Sechzehn Länder, sechzehn Bestimmungen? Aktuelle rechtliche Regelungen zur Benotung schriftlicher und mündlicher Leistungen an Gymnasien. *Zeitschrift für Pädagogik*, 71(1), 64–86. <https://doi.org/10.3262/ZP250106>.
- Köller, O. (2005). 11: Bezugsnormorientierung von Lehrkräften: Konzeptuelle Grundlagen, empirische Befunde und Ratschläge für praktisches Handeln. In R. Vollmeyer & J. Brunstein (Hrsg.), *Motivationspsychologie und ihre Anwendung* (S. 189–202). Kohlhammer.
- Krieger, R. (2003). Das sogenannte Mündliche: Eine Befragung von Beurteilten und Beurteilern zur Praxis und Problematik der Bewertung mündlicher Leistungen. *Bildung und Erziehung*, 56(1), 75–92.
- Krüll, C. (2023). *Wie erfassen und bewerten Lehrkräfte mündliche Mitarbeit?* Georg Olms. <https://doi.org/10.17879/41009639243>.
- Kunter, M., Kleikmann, T., Klusmann, U., & Richter, D. (2011). Die Entwicklung professioneller Kompetenz von Lehrkräften. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss & M. Neubrand (Hrsg.), *Professionelle Kompetenz von Lehrkräften: Ergebnisse des Forschungsprogramms COACTIV* (S. 55–68). Waxmann.
- Lohmann, J. F., Lötscher, F., Junge, F., Keller, S., Jansen, T., Fleckenstein, J., & Möller, J. (2025). Testing teacher judgments comprehensively: Accuracy, halo, frame of reference, strategy, and personality effects in holistic and analytic assessments of student essays. *Journal of Educational Psychology*. <https://doi.org/10.1037/edu0000969>.
- Lüders, M. (2001). Probleme von Lehrerinnen und Lehrern mit der Beurteilung von Schülerleistungen. *Zeitschrift für Erziehungswissenschaft*, 4(3), 457–474. <https://doi.org/10.1007/s11618-001-0047-6>.

- Marsh, H. W., Morin, A. J. S., Parker, P. D., & Kaur, G. (2014). Exploratory structural equation modeling: An integration of the best features of exploratory and confirmatory factor analysis. *Annual Review of Clinical Psychology*, 10, 85–110. <https://doi.org/10.1146/annurev-clinpsy-032813-153700>.
- Mayr, J., & Neuweg, G. H. (2006). Der Persönlichkeitsansatz in der LehrerInnenforschung. Grundsätzliche Überlegungen, exemplarische Befunde und Implikationen für die LehrerInnenbildung. In M. Heinrich & U. Greiner (Hrsg.), *Schauen, was 'rauskommt: Kompetenzförderung, Evaluation und Systemsteuerung im Bildungswesen* (S. 183–206). LIT.
- McNeish, D. (2023). A practical guide to selecting and blending approaches for clustered data: clustered errors, multilevel models, and fixed-effect models. *Psychological Methods*. <https://doi.org/10.1037/met0000620>.
- Meuser, M., & Nagel, U. (2009). Das Experteninterview – Konzeptionelle Grundlagen und methodische Anlage. In S. Pickel, G. Pickel, H.-J. Lauth & D. Jahn (Hrsg.), *Methoden der vergleichenden Politik- und Sozialwissenschaft* (S. 465–479). VS. [https://doi.org/10.1007/978-3-531-91826-6\\_23](https://doi.org/10.1007/978-3-531-91826-6_23).
- Meyer, J., Jansen, T., Hübner, N., & Lüdtke, O. (2023). Disentangling the association between the big five personality traits and student achievement: Meta-Analytic evidence on the role of domain specificity and achievement measures. *Educational Psychology Review*, 35(12), 1–34. <https://doi.org/10.1007/s10648-023-09736-2>.
- Muraki, E. (1992). A generalized partial credit model: application of an EM algorithm. *Applied Psychological Measurement*, 16(2), 159–176. <https://doi.org/10.1177/014662169201600206>.
- Muthén, L. K., & Muthén, B. O. (2017). *Mplus user's guide* (8. Aufl.). Muthén & Muthén.
- Neumann, M., Trautwein, U., & Nagy, G. (2011). Do central examinations lead to greater grading comparability? A study of frame-of-reference effects on the University entrance qualification in Germany. *Studies in Educational Evaluation*, 37(4), 206–217. <https://doi.org/10.1016/j.stueduc.2012.02.002>.
- Nylund-Gibson, K., & Masyn, K. E. (2016). Covariates and mixture modeling: results of a simulation study exploring the impact of misspecified effects on class enumeration. *Structural Equation Modeling: A Multidisciplinary Journal*, 23(6), 782–797. <https://doi.org/10.1080/10705511.2016.1221313>.
- Olczyk, M., Kwon, S. J., Lorenz, G., Perinetti Casoni, V., Schneider, T., Volodina, A., Waldfoegel, J., & Washbrook, E. (2023). Teacher judgements, student social background, and student progress in primary school: a cross-country perspective. *Zeitschrift Für Erziehungswissenschaft*, 26(2), 443–468. <https://doi.org/10.1007/s11618-022-01119-7>.
- R Core Team (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <http://www.R-project.org>
- Rammstedt, B., Danner, D., Soto, C. J., & John, O. P. (2020). Validation of the short and extra-short forms of the Big Five inventory-2 (BFI-2) and their German adaptations. *European Journal of Psychological Assessment*, 36(1), 149–161. <https://doi.org/10.1027/1015-5759/a000481>.
- Rausch, T., Karing, C., Dörfler, T., & Artelt, C. (2016). Personality similarity between teachers and their students influences teacher judgement of student achievement. *Educational Psychology*, 36(5), 863–878. <https://doi.org/10.1080/01443410.2014.998629>.
- Rheinberg, F., & Fries, S. (2018). Bezugsnormorientierung. In D. H. Rost, J. R. Sparfeldt & S. R. Buch (Hrsg.), *Handwörterbuch Pädagogische Psychologie* (5. Aufl., S. 56–63). Beltz.
- Rost, J. (1990). Rasch models in latent classes: an integration of two approaches to item analysis. *Applied Psychological Measurement*, 14(3), 271–282. <https://doi.org/10.1177/014662169001400305>.
- Soto, C. J., & John, O. P. (2017). Short and extra-short forms of the Big Five Inventory-2: The BFI-2-S and BFI-2-XS. *Journal of Research in Personality*, 68, 69–81. <https://doi.org/10.1016/j.jrp.2017.02.004>.
- Voss, T., Kunter, M., & Baumert, J. (2011). Assessing teacher candidates' general pedagogical/psychological knowledge: test construction and validation. *Journal of Educational Psychology*, 103(4), 952–969. <https://doi.org/10.1037/a0025125>.
- Zaborowski, K. U., Meier, M., & Breidenstein, G. (2011). *Leistungsbewertung und Unterricht*. VS. <https://doi.org/10.1007/978-3-531-93218-7>.

**Hinweis des Verlags** Der Verlag bleibt in Hinblick auf geografische Zuordnungen und Gebietsbezeichnungen in veröffentlichten Karten und Institutsadressen neutral.