

Samel, Lena; Platz, Friedrich; Hasselhorn, Johannes

Der Einfluss der Informationsvielfalt auf die Urteilsgenauigkeit von angehenden Musiklehrkräften bei der Evaluation und Prognose von Einzelgesangsleistungen

Zeitschrift für Erziehungswissenschaft 29 (2026) 3, S. 871-897



Quellenangabe/ Reference:

Samel, Lena; Platz, Friedrich; Hasselhorn, Johannes: Der Einfluss der Informationsvielfalt auf die Urteilsgenauigkeit von angehenden Musiklehrkräften bei der Evaluation und Prognose von Einzelgesangsleistungen - In: Zeitschrift für Erziehungswissenschaft 29 (2026) 3, S. 871-897 - URN: urn:nbn:de:0111-pedocs-357147 - DOI: 10.25656/01:35714; 10.1007/s11618-026-01382-y

<https://nbn-resolving.org/urn:nbn:de:0111-pedocs-357147>

<https://doi.org/10.25656/01:35714>

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
Internet: www.pedocs.de

Mitglied der


Leibniz-Gemeinschaft

Der Einfluss der Informationsvielfalt auf die Urteilsgenauigkeit von angehenden Musiklehrkräften bei der Evaluation und Prognose von Einzelgesangsleistungen

Lena Samel  · Friedrich Platz  · Johannes Hasselhorn 

Eingegangen: 6. Dezember 2024 / Überarbeitet: 27. November 2025 / Angenommen: 14. Dezember 2025
/ Online publiziert: 20. Februar 2026
© The Author(s) 2026

Zusammenfassung Evaluation und Prognose von Schülerleistungen sind zentrale diagnostische Anforderungen an Lehrkräfte und eine unverzichtbare Voraussetzung für die Individualisierung von Lernprozessen. Hierbei müssen Lehrkräfte gezeigte Leistungen von Schüler:innen bewerten (Evaluation) und zukünftige Leistungen einschätzen können (Prognose). Im Unterrichtsfach Musik gilt dies insbesondere in musikpraktischen Unterrichtssituationen. Die vorliegende Studie geht der Frage nach, wie gut Studierende als angehende Musiklehrkräfte Einzelgesangsleistungen von Schüler:innen einschätzen können und in welchem Maße eine größere Vielfalt an Informationen über deren bisherige Leistung die Genauigkeit der Einschätzungen der Studierenden erhöhen kann. Außerdem wurde untersucht, inwieweit den angehenden Musiklehrkräften die Einschätzung des Anforderungsniveaus ausgewählter Musikstücke in ihrem Prognoseurteil gelingt. Dazu führten wir ein Experiment mit Lehramtsstudierenden durch ($N=98$), bei dem die Niveau-, Differenzierungs- und Rangkomponente der Urteilsgenauigkeit erfasst wurde. Es zeigte sich, dass ihre Urteilsgenauigkeit im Durchschnitt vergleichbar ist mit derjenigen im Beruf stehender Lehrkräfte anderer Fachrichtungen bei allerdings großer Heterogenität ihrer Urteilsgenauigkeit. Die Variation der Vielfalt der Informationen verbesserte die Genauigkeit der Leistungsprognose nur bei der Rangkomponente.

✉ Lena Samel · Prof. Dr. Friedrich Platz
Institut für Ästhetisch-Kulturelle Wissenschaft und Praxis, Europa-Universität Flensburg, Auf dem Campus 1, 24943 Flensburg, Deutschland
E-Mail: lena.samel@studierende.uni-flensburg.de

Prof. Dr. Friedrich Platz
E-Mail: friedrich.platz@uni-flensburg.de

✉ Lena Samel · Prof. Dr. Johannes Hasselhorn
Department Fachdidaktiken, Friedrich-Alexander-Universität Erlangen-Nürnberg, Regensburger Str. 160, 90478 Nürnberg, Deutschland
E-Mail: johannes.hasselhorn@fau.de

Schlüsselwörter Urteilsgenauigkeit · Pädagogische Diagnostik · Musikpraxis · Evaluation · Prognose

The impact of information variety on judgment accuracy of prospective music teachers' evaluation and prediction of pupils' individual singing performances

Abstract For teachers, the evaluation and prediction of pupil performance and achievement are central diagnostic tasks, as well as crucial prerequisites for the individualization of learning processes. In music education, and particularly in music performance situations, diagnostic challenges arise that are demanding in terms of securing diagnostic accuracy. Our study examined 1) the judgment accuracy of pupil music achievement evaluation and prediction among music students playing the role of prospective teachers; 2) the extent to which the provision of various pieces of information about pupil solo singing performances could increase judgment accuracy; 3) to which music students accurately judge the difficulty level of two selected musical pieces as in their prognostic evaluations. An online experiment was conducted with 98 music students. On average, they achieved a level of judgment accuracy comparable to that of teachers of other school subjects, albeit with high between-rater variability. Increasing the variety of information improved the accuracy of performance predictions and evaluations, but only for the rank-order component.

Keywords Judgment accuracy · Pedagogical diagnostics · Singing · Evaluation · Prognosis

Diagnostik gehört zum professionellen Handeln von Lehrkräften, um Unterschiede in den individuellen Lernvoraussetzungen und weitere Gelingensbedingungen erfolgreichen Lernens in ihren Lerngruppen zu erfassen (Hesse und Latzko 2017). Daraus ziehen sie Erkenntnisse für ihr unterrichtliches Handeln. Diagnostik zielt im Unterrichtskontext darauf ab, kontinuierlich didaktische Optimierungspotentiale zu identifizieren. Diese sind Grundlage für die Individualisierung von Lernprozessen, die Planung und Durchführung von adaptivem Unterricht, für nachfolgendes Rückmeldeverhalten sowie für (Förder-)Entscheidungen (Klieme und Warwas 2011). Eine prozessorientierte und änderungssensitive Individualdiagnostik im Fachunterricht umfasst sowohl die Bewertung und Beurteilung aktueller Leistungen (Evaluation) als auch die Einschätzung zukünftiger Leistungen (Prognose) der Schüler:innen. Die Fähigkeit von Lehrkräften, diagnostische Aufgaben erfolgreich zu bewältigen und dabei qualitätsvolle Diagnosen zu erstellen, wird als diagnostische Kompetenz bezeichnet (vgl. Schrader 2013).

Zu den diagnostischen Vorgehensweisen in diesem Sinne zählen im Fach Musik neben formellen, informellen, halb- und vollstandardisierten insbesondere nichtstandardisierte Vorgehensweisen mit unterschiedlicher Evaluationsgüte (vgl. Ehninger et al. 2021; Greuel 2007; Hasselhorn 2015; Imthurn 2020; Knigge et al. 2010;

Wolf und Kopiez 2018). Für Lehrkräfte im Fach Musik bestehen besondere diagnostische Anforderungen, da ein wesentlicher Anteil des Unterrichtsgeschehens zum Erwerb musikpraktischer Kompetenzen auf die Unterrichtsmethode des gemeinsamen Klassenmusizierens entfällt. Sie müssen in der Lage sein, aus diesen komplexen Situationen (musikalische Gruppendarbietung) ausreichende Informationen für eine fundierte Individualdiagnostik zu gewinnen (vgl. Hasselhorn et al. 2022), d. h. auf Grundlage der Leistungsdarbietung einer Lerngruppe auf die Kompetenzen einzelner zu schließen. Die Besonderheit liegt darin, dass die Leistung des gemeinsamen Musizierens auch von Faktoren der wechselseitigen Beeinflussung der musizierenden Personen abhängig ist, wie bspw. das individuelle Timingverhalten im Ensemble verdeutlicht (s. Nowicki et al. 2013). Hinzu kommt, dass im Schulkontext häufig auch die Lehrkraft selbst Teil der musizierenden Gruppe ist, um Musikpraxis überhaupt zu initiieren (Hasselhorn 2015). Hierzu zählt bspw. das instrumentale Begleiten der Klasse, die Unterstützung des Gesangs oder das Anleiten und Interpretieren der Gruppe als Dirigent:in (teilweise auch gleichzeitig), sodass die Performanz der Schüler:innen auch von der Lehrkraft abhängig ist. Um der Forderung nach Individualisierung (Klieme und Warwas 2011) auch im Musikunterricht gerecht zu werden, ist es erforderlich, dass Lehrkräfte trotz dieser komplexen Zusammenhänge in der Lage sind, musikpraktische Kompetenzen von Schüler:innen auf Individualebene korrekt einzuschätzen und zu fördern.

Außerschulische musikalische Aktivitäten tragen darüber hinaus zu einer überdurchschnittlichen Leistungsheterogenität der Schüler:innen in den Bereichen Gesang und Instrumentalspiel bei (Hasselhorn und Lehmann 2015; Lill et al. 2020), weshalb gerade in musikpraktischen Prozessen individuelle Lernprozesse und adaptives Unterrichten von zentraler Bedeutung für individuelle Kompetenzentwicklung im Rahmen des Klassenmusizierens sind (Vogt et al. 2012). Für eine angemessene Leistungsbewertung kommt erschwerend hinzu, dass die Evaluation musikalischer Darbietungen grundsätzlich von verschiedenen Faktoren beeinflusst wird, die über die reine Beurteilung der Frage hinausgehen, wie angemessen musikalische Parameter als Ausdruck erworbener Kompetenzniveaus umgesetzt wurden (vgl. Hasselhorn 2015; Hasselhorn und Wolf 2018; Platz und Kopiez 2022).

Wenngleich die Relevanz musikpädagogischer Diagnostik zur Optimierung musikpädagogischen Handelns bereits dargelegt wurde (Greuel 2007; Greuel und Szczepaniak 2007), fehlt es bisher an Studien, die sich mit der diagnostischen Leistungseinschätzung von Musiklehrkräften beschäftigen. Ziel dieses Beitrags ist es daher, Erkenntnisse über fachspezifische Herausforderungen bei der Diagnostik musikpraktischer Leistungen zu gewinnen. Der Fokus liegt auf der Beurteilung von Einzelgesangsleistungen von Schüler:innen als Ausdruck ihrer musikpraktischen Kompetenz, die im Klassenmusizieren erworben wird. Im Zentrum steht dabei die Frage, wie gut bzw. angemessen bereits Musikstudierende als angehende Musiklehrkräfte in der Lage sind, musikpraktische Individualleistungen von Schüler:innen einzuschätzen.

1 Theoretischer und empirischer Hintergrund

1.1 Diagnostische Kompetenz von Lehrkräften

Das lehrkräfteseitige pädagogische Diagnostizieren lässt sich in zwei Kernbereiche mit unterschiedlichen Anliegen unterteilen: eine retrospektiv-reflektierende sowie eine prospektiv-antizipierende Facette (Ingenkamp und Lissmann 2008). Die retrospektiv-reflektierende Facette umfasst dabei alle Vorgehensweisen zur bilanzierenden Diagnose des bisher gezeigten (Leistungs-)Verhaltens von Schüler:innen nach (Teil-)Abschluss von Programmen und Interventionen (wie dem Unterricht, einer Unterrichtseinheit usw.) im Sinne einer summativen Evaluation. Im Unterschied dazu zielen Evaluationsvorhaben der prospektiv-antizipierenden Facette darauf ab, notwendige Verhaltensmodifikationen zur Steigerung der Lernwirksamkeit noch während der Durchführung von pädagogischen Programmen, Maßnahmen oder Unterrichtsvorhaben festzustellen (Hesse und Latzko 2017). Hierfür werden Vorgehensweisen der prozessbegleitenden Erhebung diagnostischer Informationen wie das formative Assessment eingesetzt (Schütze et al. 2018). Auf Grundlage der dadurch gewonnenen Informationen werden das Leistungspotential über die Einschätzung zukünftiger Leistungen prognostiziert und im Falle einer Differenz zwischen Prognose und Zielvorstellung Gestaltungsmodifikationen zukünftiger Lehr-Lern-Prozesse eruiert (Platz 2018, S. 381). Mit diesen wird versucht, eine ggf. vorliegende Differenz zu minimieren und damit die Lernwirksamkeit des eigenen unterrichtlichen Handelns zu optimieren.

In empirischen Arbeiten wird diagnostische Kompetenz häufig über die Urteilsgenauigkeit erfasst, mit der Lehrkräfte bestimmte Lernenden- oder Aufgabenmerkmale einschätzen können. Sie spiegelt folglich den Übereinstimmungsgrad von Lehrkräfteurteil und tatsächlicher Merkmalsausprägung wider. Eine weit verbreitete und nach wie vor etablierte Operationalisierung geht auf Schrader und Helmke (1987) zurück, die drei Komponenten der Urteilsgenauigkeit unterscheiden: die Niveau-, die Differenzierungs- und die Rangkomponente. Zur Erfassung dieser drei Komponenten forderten sie Lehrkräfte zur Abgabe eines Prognoseurteils über Leistungen ihrer Schüler:innen auf, während diese Testaufgaben bearbeiteten („tatsächliche“ Leistung). Zur Bestimmung der drei Urteilsgenauigkeitskomponenten setzten sie die so erhobenen Prognoseurteile der Lehrkräfte mit den Leistungen ihrer Schüler:innen in Beziehung. Hierbei beschreibt die Niveauebene, wie präzise Lehrkräfte das durchschnittliche Leistungsniveau einer Lerngruppe einschätzen, operationalisiert über die Differenz des mittleren Lehrkräfteurteils und der mittleren, „tatsächlichen“ Leistung der Lerngruppe (vgl. Urhahne und Wijnia 2021). Die Differenzierungskomponente bezieht sich darauf, wie präzise Lehrkräfte die Leistungsbreite und -verteilung einer Lerngruppe einschätzen können (Leistungsheterogenität). Sie wird bestimmt, indem die Streuung der Einzelleistungen um das lehrerseitig eingeschätzte mittlere Leistungsniveau der Lerngruppe ins Verhältnis zu ihrer „tatsächlichen“ Leistungsstreuung gesetzt wird. Die Rangkomponente bezieht sich schließlich darauf, wie gut die Rangreihe der Leistungseinschätzungen mit der „tatsächlichen“ Leistungsangabe übereinstimmt, was über die Korrelation zwischen eingeschätzter und tatsächlicher Leistung erfasst wird.

Schrader (1989) hat eine zusätzliche Differenzierung zwischen aufgabenbezogener und aufgabenspezifischer Urteilsgenauigkeit vorgeschlagen. Während die aufgabenbezogene Urteilsgenauigkeit sich darauf bezieht, wie gut eine Lehrkraft die Schwierigkeit von Aufgaben für eine spezifische Lerngruppe einschätzt, handelt es sich bei der aufgabenspezifischen Urteilsgenauigkeit um eine Art prognostische Fertigkeit der Lehrkraft: Wie gut kann sie einschätzen, ob konkrete Schüler:innen bestimmte Aufgaben korrekt lösen werden.

Eine weitere Ausdifferenzierung der Operationalisierungen der Urteilsgenauigkeit von Südkamp et al. (2012) grenzt informiertes und uninformatiertes Urteilen voneinander ab. Ein Urteil gilt danach erst dann als informiert, wenn einer Lehrkraft das Testinstrument oder der Vergleichsstandard bekannt ist, auf dem das Vergleichsurteil basiert. Sollte die Lehrkraft dagegen in einem nichtstandardisierten Verfahren ein globales Urteil über die grundsätzliche Leistungsfähigkeit der Schüler:innen in einem Fach fällen, handelt es sich um ein uninformatiertes Urteil. Im Zuge einer Metaanalyse auf Grundlage von 75 Studien konnten Südkamp et al. (2012) eine personenbezogene Ranggenauigkeit des Lehrerurteils von beachtlicher Gesamteffektgröße ermitteln ($r = 0,63$, 95 % CI [0,62; 0,64]). Die Heterogenitätsanalyse der Effektgrößenverteilung ergab, dass informierte Urteile ($r = 0,76$, 95 % CI [0,74; 0,78]) höhere Übereinstimmungen aufwiesen als uninformierte Urteile ($r = 0,61$, 95 % CI [0,60; 0,62]). In einer jüngeren Überblicksarbeit über die Urteilsgenauigkeit von Lehrkräften wurde eine Tendenz zur Überschätzung des mittleren Leistungsniveaus für die Niveauebene nachgewiesen (Urhahne und Wijnia 2021). Für die Differenzierungskomponente ergibt sich aus den Befunden vorliegender Studien keine verallgemeinerbare Tendenz.

1.2 Beurteilung musikpraktischer Leistungen

Das Zustandekommen und die Qualität von Leistungsurteilen wurden auch in musikwissenschaftlichen Untersuchungen thematisiert, meist im Zusammenhang mit dem Instrumentalspiel (z.B. Wesolowski 2017). Die Erhebungssituationen ermöglichten Rückschlüsse auf das überwiegend ästhetische Beurteilungsverhalten von Gruppen- oder Einzelleistungen, wie sie bspw. im Rahmen von Wettbewerben (Asmat et al. 2023), Aufnahmeprüfungen an Musikhochschulen oder Orchesterproben (Bellmann 2020) vorkommen. Die ästhetischen Urteile fungieren dabei als Basis für unterschiedliche Zielvariablen: im Falle der Wettbewerbe als Leistungsvariable, im Falle der Orchesterproben als Indikator für Berufseignung. Bei diesen Forschungsansätzen geht es in der Regel um die Fähigkeit zur (summativen) Evaluation einer vorgestellten Leistung. Der Hauptunterschied zur oben skizzierten Operationalisierung der Urteilsgenauigkeit besteht jedoch darin, dass nicht die Abweichung des Leistungsurteils von einem ‚wahren‘, existierenden Leistungswert gemessen wird. Stattdessen wird die Zwischen-Beurteiler-Reliabilität (sogenannte *Inter-Rater-Reliabilität*) fokussiert, verbunden mit der Frage, in welchem Maße Expert:innen in ihrer Evaluation von musikalischen Darbietungen zu vergleichbaren Urteilen kommen (z.B. Thompson et al. 1998) oder in welchem Maße Selbst- und Fremdeinschätzung übereinstimmen (z.B. Morijiri und Welch 2022). In weiteren Studien wurde die Urteilsstabilität bzw. -zuverlässigkeit von Personen untersucht,

die zur mehrmaligen Urteilsfindung derselben Leistungen aufgefordert wurden (so genannte *Intra-Rater-Reliabilität*; z. B. Wesolowski et al. 2017).

Die Variabilität von Expertenurteilen, für die kein ‚wahrer Vergleichswert‘ (*true value*) vorliegt, wurde auch in anderen Domänen untersucht, u. a. bei Richterurteilen oder Versicherungsprognosen (Kahneman et al. 2021). Die Ursachen vorfindbarer Urteilsverzerrungen erklärt Kahneman (2011) mit seinem Modell zum Urteilsprozess, in dem er zwischen zwei Denkprozessen unterscheidet, die beim Urteilen und Treffen von Entscheidungen ablaufen: Denkprozesse bei Urteilen, die schnell, intuitiv und oft automatisiert erfolgen, und solche, die langsam ablaufen, weil logisches Denken und mehr kognitive Anstrengung erforderlich sind. Diesem Modell zufolge sind Urteilsverzerrungen durch den Rückgriff auf mindestens eine von drei zentralen Heuristiken – Repräsentativität, Verfügbarkeits- und Ankerheuristik – zurückzuführen. Diese treten vor allem bei Urteilen unter Unsicherheit (*judgment under uncertainty*, Tversky und Kahneman 1974) oder bei schnell zu fällenden Urteilen auf (Kahneman 2011).

Die Variabilität von (Experten-)Urteilen in Kontexten, in denen es kein objektiv bestimmbares ‚richtiges‘ oder ‚falsches‘ Urteil gibt, lässt sich Kahneman et al. (2021) zufolge darüber hinaus zwischen der systematischen, teils vorhersehbaren Urteilsverzerrung (*bias*) und einer Zufallsstreuung (*noise*) unterscheiden. In seinen Arbeiten konnte er zeigen, dass *bias* aus dem Rückgriff auf Heuristiken und dem schnellen Denkprozess resultiert (s. oben), während hingegen *noise* durch situative Einflüsse (z. B. das individuelle Strengemaß, die Stimmung oder Müdigkeit einer Lehrkraft bei der Leistungsbewertung) und andere unvorhersehbare Faktoren entstehen kann.

Passarotto et al. (2023) stellten u. a. heraus, dass die Zufallsstreuung einen erheblichen Anteil an der Variabilität von Expertenurteilen bei Instrumentaldarbietungen ausmacht. Im Zusammenhang mit dem Mehrkomponentenmodell zur Beurteilung musikalischer Darbietungen (McPherson und Schubert 2022) speist sich diese sogenannte Noise-Komponente vermutlich aus situativ unterschiedlichen Quellen, etwa dem Auftreten oder der Bewegung der Interpret:innen (vgl. Platz und Kopiez 2022), oder nichtmusikalischen Kontextfaktoren wie der Raumakustik oder Unterstützung des Publikums. Selbst wenn ausschließlich musikalisch-interpretative Beschreibungsgrößen als Evaluationskriterien vorgegeben sind, zeigen die Arbeiten von Wesolowski (2017), dass die Urteilsvarianz zwischen den beurteilenden Personen trotz kriterialer Standardisierung auf individuell gestaltete Bewertungsprozesse zurückzuführen ist. Ihre Individualität spiegelt sich in einer unterschiedlich ausgeprägten „Urteilsstrenge“ zwischen den beurteilenden Personen wider und resultiert vor allem aus der individuell gestalteten Gewichtung von Beurteilungskriterien. Auch Juslin (2023) nimmt an, dass die Evaluation musikalischer Darbietungen durch mehrere subjektiv-gewichtete Kriterien bestimmt wird, weshalb er Leistungsurteile als Form des ästhetischen Urteilsprozesses ansieht.

Das Verhältnis zwischen ästhetischem Urteil, das sich bereits im einfachen Gefallensurteil widerspiegeln kann, und dem diagnostischen Urteil im pädagogischen Kontext ist bislang unklar. Auch wenn sowohl das ästhetische als auch das diagnostische Urteil eine Form der Evaluation sind, bezieht sich ersteres im Unterschied zur pädagogischen Diagnostik lediglich auf die Beurteilung, wie überzeugend das

Verhalten von Interpret:innen auf die Rezipient:innen wirkt. Die Vermutung, dass ein Zusammenhang zwischen ästhetischem und diagnostischem Urteilsverhalten bestehen könnte, lässt sich durch das dreidimensionale Handlungsmodell der musikalischen Interpretation (Lehmann 2011) sowie dessen vierdimensionaler Erweiterung (Kopiez 2018, S. 188) stützen, da die darin beschriebenen Dimensionen jeweils auf Fertigkeiten zielen, die sich instrumentalpädagogisch und – in gewissen Grenzen – im schulischen Musikunterricht vermitteln lassen. Die pädagogische Diagnostik wirft jedoch darüber hinaus weitere Begleitfragen auf: Welche Fertigkeiten- und Kompetenzniveaus lassen sich aus der musikalischen Darbietung ableiten? Ist die musikalische Darbietungsleistung alters- und entwicklungsgerecht? Welche Referenz liegt dieser Einschätzung zugrunde und in welchem Zusammenhang stehen (frühere oder zukünftige) pädagogische Maßnahmen zu anvisierten Entwicklungsmöglichkeiten (in Bezug auf Fertigkeiten, Kompetenzen oder Einstellungen)? Auch wenn es in der musikpädagogischen Forschung bereits erste Ansätze gibt, die Möglichkeiten aufzeigen, wie solche pädagogisch-diagnostischen Informationen gewonnen werden können (Greuel und Horst 2011; Imthurn 2020), bleibt bislang offen, wie gut Musiklehrkräften diese diagnostischen Anforderungen gelingen und wie der diagnostische Urteilsprozess abläuft.

Für den schulischen Musikunterricht gelten musikpraktische Prozesse in weit verbreiteten didaktischen Ansätzen des Musikunterrichts, bspw. des Aufbauenden Musikunterrichts, als ein erstes, zentrales Anliegen (Jank 2021), da die didaktisch begründeten und methodisch initiierten Erfahrungen und Erlebnisse der Schüler:innen aus diesen Prozessen die Grundlage und den Beginn des weiteren musikalischen Lernens im Zuge musikalischer Bildung (Kaiser 2010) darstellen (Bähr 2001; Gruhn 2018; Jank 2021).

Mit einem experimentellen Vorgehen untersuchten Hasselhorn et al. (2022), wie hoch die Urteilsgenauigkeit von Musiklehrkräften bei der Evaluation von für den schulischen Musikunterricht typischen Gruppengesangsleistungen ausfiel und welchen Einfluss die Einschätzung des mittleren Leistungsniveaus (Niveauebene) und der Heterogenität einer Lerngruppe (Differenzierungskomponente) auf die Unterrichtsplanung der Lehrkräfte hatte. Die Befunde legten nahe, dass den Lehrkräften zwar die Einschätzung der Niveauebene gelingt, jedoch nicht die der Differenzierungskomponente. Die Varianz der Unterrichtsplanung scheint somit ausschließlich durch die Niveauebene systematisch beeinflusst, nicht jedoch durch die Differenzierungskomponente, die insbesondere für die Individualisierung von Lernprozessen hochrelevant ist. Dabei konnte kein Effekt zunehmender Expertise durch Berufserfahrung gefunden werden. Da Hasselhorn et al. (2022) ausschließlich die Einschätzung von Gruppenleistungen untersucht haben, bleibt unklar, wie genau Musiklehrkräfte Schülerleistungen auf Individualebene diagnostizieren können. Vor diesem Hintergrund kann vermutet werden, dass Musiklehrkräfte im Unterrichtsalltag aufgrund der überwiegenden Gruppensituationen und der damit verbundenen Herausforderungen nur in geringem Umfang Informationen über die Leistungsfähigkeit der Schüler:innen auf Individualebene erhalten. Hinzu kommt, dass in der Unterrichtspraxis eher flüchtige Höreindrücke die Regel sind, die von der Lehrkraft schnell und intuitiv verarbeitet werden müssen.

Aufgrund der situativen Bedingungen des musikpraktischen Unterrichts im Kontext Schule sowie der im Vergleich zu anderen Fächern geringen Stundenzahl sind die leistungsbezogene Informationsqualität und -quantität, die Lehrkräfte von einzelnen Schüler:innen erhalten, eher gering. Die Frage, wie sich die Vielfältigkeit der Informationen, die Lehrkräfte von ihren Schüler:innen in verschiedenen Situationen erhalten, auf die prognostischen Urteile auswirken, ist für den Musikunterricht bislang noch nicht untersucht worden. Generell zeigt sich, dass mehr bzw. vielfältigere Informationen aus unterschiedlichen Situationen zu akkurateren personenbezogenen Urteilen führen können (Kahneman 2011; Letzring et al. 2006). Aus diesem Grund ist ein Ziel dieser Studie zu untersuchen, ob ein solcher Effekt auch im Musikunterricht auftritt, indem insbesondere geprüft wird, ob die Informationsvielfalt, die angehende Musiklehrkräfte von den musikpraktischen Leistungen der Schüler:innen auf Individualebene erhalten, zu genaueren Urteilen führt. Auf Grundlage dieses Wissens wäre es für Lehrkräfte möglich, im Unterricht gezielt Situationen zu schaffen, die im Zuge ihrer pädagogischen Diagnostik zur Erhöhung der Urteilsgenauigkeit führen.

2 Fragestellung und Hypothesen

Die vorliegende Studie untersucht die Komponenten der Urteilsgenauigkeit von Studierenden bei der Beurteilung von Einzelgesangsleistungen in Abhängigkeit von der Vielfalt zur Verfügung stehender Informationen. Im Einzelnen sollen dabei die folgenden Fragestellungen beantwortet werden:

Fragestellung 1 Wie genau können Studierende als angehende Musiklehrkräfte Einzelgesangsleistungen von Schüler:innen evaluieren und prognostizieren?

Durch die Verwendung von Einzelgesangsleistungen ist es möglich, die Komponenten der Urteilsgenauigkeit zu operationalisieren, sodass ein Vergleich mit den Ergebnissen anderer Fachdidaktiken möglich ist. Wie in anderen Fächern üblich, wird in der vorliegenden Studie untersucht, wie gut Studierende in der Lage sind, Leistungen von Schüler:innen zu prognostizieren. Angesichts der aufgezeigten Herausforderungen, die mit der musikpraktischen Leistungsevaluation verbunden sind, wird zusätzlich die Urteilsgenauigkeit für die Leistungsevaluation bestimmt, indem das Maß der Übereinstimmung zwischen den Urteilen von Studierenden über eine gehörte Gesangsdarbietung und den Urteilen geschulter Urteiler:innen betrachtet wird.

Durch die Operationalisierung nach Schrader und Helmke (1987) lässt sich zum ersten Mal für das Fach Musik eine Rangkomponente berechnen, die von Schrader (1989) als eigentliches Maß der Urteilsgenauigkeit bezeichnet wird, da sie nicht von Urteilstendenzen beeinflusst wird. Südkamp et al. (2012) konnten zeigen, dass das Unterrichtsfach keinen signifikanten Einfluss auf die Urteilsgenauigkeit der Lehrkräfte hat. Aus diesem Grund wird angenommen, dass die angehenden Musiklehrkräfte die Rangordnung der Schülerleistungen trotz der fachspezifischen Herausforderungen bei der Diagnostik musikpraktischer Leistungen grundsätzlich in gleicher Weise differenziert einschätzen können (H_{1a} und H_{1b} , s. Tab. 1). Vor dem

Tab. 1 Hypothesenüberblick

H	Formalisierte Hypothese	Hypothese
<i>Fragestellung 1</i>		
H _{1a}	Die Rangkomponente der Prognose wird differenziert eingeschätzt.	$\rho_{StU_Prog;SL} > 0$
H _{1b}	Die Rangkomponente der Evaluation wird differenziert eingeschätzt.	$\rho_{StU_Eva;SL} > 0$
H _{1c}	Die Prognosen angehender Musiklehrkräfte weichen vom theoretischen Idealwert der Niveauekomponente ab.	$M_{StU_Prog} - M_{SL} \neq 0$
H _{1d}	Die Prognosen angehender Musiklehrkräfte weichen vom theoretischen Idealwert der Differenzierungskomponente ab.	$\frac{\sigma_{StU_Prog}}{\sigma_{SL}} \neq 1$
H _{1e}	Die Evaluationen angehender Musiklehrkräfte weichen vom theoretischen Idealwert der Niveauekomponente ab.	$M_{StU_Eva} - M_{SL} \neq 0$
H _{1f}	Die Evaluationen angehender Musiklehrkräfte weichen vom theoretischen Idealwert der Differenzierungskomponente ab.	$\frac{\sigma_{StU_Eva}}{\sigma_{SL}} \neq 1$
<i>Fragestellung 2</i>		
H _{1g}	Die Niveauekomponente ist bei zwei Prognosestücken unterschiedlich hoch.	$(M_{StU_Stück1} - M_{SL_Stück1}) - (M_{StU_Stück2} - M_{SL_Stück2}) \neq 0$
H _{1h}	Die Differenzierungskomponente ist bei zwei Prognosestücken unterschiedlich hoch.	$\frac{\sigma_{StU_Stück1}}{\sigma_{SL_Stück1}} - \frac{\sigma_{StU_Stück2}}{\sigma_{SL_Stück2}} \neq 0$
H _{1i}	Die Rangkomponente ist bei zwei Prognosestücken unterschiedlich hoch.	$\rho_{StU_Stück1;SL_Stück1} - \rho_{StU_Stück2;SL_Stück2} \neq 0$
<i>Fragestellung 3</i>		
H _{1j}	Die Abweichung vom theoretischen Idealwert der Niveauekomponente der Prognose ist bei hoher Informationsvielfalt geringer.	$\left (M_{StU_Prog} - M_{SL})_{hoch} \right - \left (M_{StU_Prog} - M_{SL})_{gering} \right < 0$
H _{1k}	Die Abweichung vom theoretischen Idealwert der Differenzierungskomponente der Prognose ist bei hoher Informationsvielfalt niedriger.	$\left \left(\frac{\sigma_{StU_Prog}}{\sigma_{SL}} \right)_{hoch} \right - \left \left(\frac{\sigma_{StU_Prog}}{\sigma_{SL}} \right)_{gering} \right > 0$
H _{1l}	Die Abweichung vom theoretischen Idealwert der Rangkomponente der Prognose ist bei hoher Informationsvielfalt niedriger.	$(\rho_{StU_Prog;SL})_{hoch} - (\rho_{StU_Prog;SL})_{gering} > 0$
H _{1m}	Die Abweichung vom theoretischen Idealwert der Niveauekomponente der Evaluation ist bei hoher und geringer Informationsvielfalt vergleichbar.	$\left (M_{StU_Eva} - M_{SL})_{hoch} \right - \left (M_{StU_Eva} - M_{SL})_{gering} \right = 0$
H _{1n}	Die Abweichung vom theoretischen Idealwert der Differenzierungskomponente der Evaluation ist bei hoher und geringer Informationsvielfalt vergleichbar.	$\left \left(\frac{\sigma_{StU_Eva}}{\sigma_{SL}} \right)_{hoch} \right - \left \left(\frac{\sigma_{StU_Eva}}{\sigma_{SL}} \right)_{gering} \right = 0$

Tab. 1 (Fortsetzung)

H	Formalisierte Hypothese	Hypothese
H ₁₀	Die Abweichung vom theoretischen Idealwert der Rangkomponente der Evaluation ist bei hoher und geringer Informationsvielfalt vergleichbar.	$\left(\rho_{StU_Eva;SL}\right)_{hoch} - \left(\rho_{StU_Eva;SL}\right)_{gering} = 0$

Bedeutung der Indizes: *StU_Eva*: Evaluationsurteil (AV₁); *StU_Prog*: Prognoseurteil (AV₂); *SL*: „tatsächliche“ Schülerleistung; Stück1 & Stück2 beziehen sich auf die unterschiedlichen Prognosestücke (UV₁); hoch & gering beziehen sich auf die Ausprägungen der Informationsvielfalt als weitere, unabhängige Variable (UV₂).

Hintergrund der musikwissenschaftlichen Forschung, die die Variabilität von Expertenurteilen vielfach nachweisen konnte, gehen wir davon aus, dass den angehenden Musiklehrkräften die Leistungsevaluation und -prognose zwar grundsätzlich gelingt, die theoretischen Idealwerte jedoch nicht erreicht werden (H_{1c} bis H_{1f}, s. Tab. 1).

Fragestellung 2 Inwiefern sind angehende Musiklehrkräfte in der Lage, unterschiedliche Anforderungsniveaus einzelner Gesangsstücke bei der Leistungsprognose schülerseitiger Einzelgesangsleistungen zu berücksichtigen?

Damit angehende Lehrkräfte in der Lage sind, einen adaptiven Musikunterricht umzusetzen, ist es notwendig, die Aufgabenschwierigkeit von Gesangsstücken für die einzelnen Schüler:innen auf Individualebene einzuschätzen. Aus diesem Grund wird untersucht, ob den angehenden Musiklehrkräften die aufgabenspezifische Urteilsgenauigkeit (Aufgaben-Personen-Wechselwirkungskomponente) gelingt. Hierfür müssen die angehenden Musiklehrkräfte die Schwierigkeiten eines Stücks erkennen und einschätzen können, wie gut ein:e bestimmte:r Schüler:in diese Schwierigkeiten bewältigen wird. Da diese Facette der Urteilsgenauigkeit bisher kaum untersucht wurde (Südkamp et al. 2012; Urhahne und Wijnia 2021), erlaubt die bisherige Forschung hierzu keine Vorhersagen. Unsere Erwartung ist, dass den angehenden Musiklehrkräften die Einschätzung der stückspezifischen Anforderungsniveaus je nach Stück unterschiedlich gut gelingt (H_{1g} bis H_{1i}, s. Tab. 1).

Fragestellung 3 Beeinflusst die Vielfalt an Informationen, die angehende Musiklehrkräfte aus den Leistungen der Schüler:innen gewinnen, die Genauigkeit ihrer Urteile?

Urteilsverzerrungen bei Vorhersagen lassen sich durch eine höhere Informationsvielfalt über eine Person verringern. Daher erwarten wir, dass Studienteilnehmer:innen, denen vielfältigere Informationen über die Leistungsfähigkeit der Schüler:innen aus unterschiedlichen Situationen vorliegen, zu genaueren Prognoseurteilen gelangen (H_{1j} bis H_{1l}, s. Tab. 1). In einer summativen, kriterial-gestützten Evaluation soll ausschließlich die jeweils dargebotene Leistung als Indikator für die ihr zugrunde liegende (musikpraktische) Kompetenz berücksichtigt werden. Daher wird angenommen, dass eine durch mehrfaches Vorsingen von Musikstücken mit unterschiedlichen Schwierigkeitsgraden entstehende Informationsvielfalt keinen zusätzlichen Einfluss auf das Evaluationsurteil angehender Musiklehrkräfte haben sollte. Diese Annahme beruht auf der testtheoretischen Überlegung, dass die musikalische Darbietung („response“) Ausdruck der ihr zugrunde liegenden musikpraktischen Kompetenz ist, die

unabhängig von den jeweils ausgewählten Musikstücken („Items“) als eher konstant gilt. Eine Varianz in der Darbietungsleistung derselben Schülerin über verschiedene Musikstücke hinweg ist demnach kein Indikator für eine fluktuierende Kompetenz, sondern ist vielmehr auf Unterschiede im Anforderungsniveau der Musikstücke (und eine Zufallskomponente) zurückzuführen. Daher wird angenommen, dass die aus dem Hören unterschiedlicher Musikstücke resultierende Informationsvielfalt keinen Einfluss auf die Evaluationskomponenten hat (H_{1m} bis H_{1o} , s. Tab. 1).

3 Methode

3.1 Stichprobe

Insgesamt nahmen an der Onlinestudie $n=190$ Personen teil, wobei $n=88$ Personen die Studie in Präsenz unter Aufsicht und $n=102$ ortsungebunden durchgeführt haben. Von der weiteren Datenanalyse wurden die Antwortdatensätze derjenigen Teilnehmenden ausgeschlossen, die der Datenverarbeitung nicht zugestimmt haben ($n=2$), das Testvideo nicht abspielen konnten ($n=12$) oder die im Testvideo dargebotenen Ziffern falsch wiedergaben ($n=13$). Auf der Fragebogenseite des Testvideos konnte zudem eine hohe Anzahl an Teilnehmenden festgestellt werden, die die Studie abgebrochen haben ($n=34$). Dies ist vermutlich darauf zurückzuführen, dass an dieser Stelle darauf hingewiesen wurde, dass die Studie Videos beinhaltet und somit eine ruhige Umgebung und geeignete Lautsprecher oder Kopfhörer benötigt werden. Des Weiteren wurden die Teilnehmenden ausgeschlossen, die angaben, keinen musikbezogenen Studiengang zu belegen ($n=1$) und die alle Schüler:innen in der Prognose oder Evaluation immer gleich bewerteten ($n=2$). Zudem wurde eine Kohorte ($n=11$) ausgeschlossen, in deren Erhebungssituation keine vollständige Standardisierung sichergestellt werden konnte. Darüber hinaus haben $n=17$ Teilnehmende die Datenerhebung zwischen dem ersten und letzten Testitem abgebrochen. Diese Gruppe wurde ebenfalls für die weiteren Analyse aus dem Datensatz entfernt.

Somit ergab sich eine bereinigte Stichprobe von $N=98$ Teilnehmenden ($n=61$ weiblich, $n=36$ männlich, $n=1$ divers), wobei $n=54$ unter Aufsicht und $n=44$ online teilgenommen haben. Von ihnen gaben $n=60$ Personen an, Lehramt für Musik an Gymnasien zu studieren, $n=1$ Lehramt für Musik an anderen Schulformen und $n=37$ ein anderes musikbezogenes Studienfach (z.B. Instrumental- und Gesangspädagogik). Die Teilnehmenden befanden sich zum Zeitpunkt der Studienteilnahme zwischen dem 1. und 13. Semester ($M=5,89$, $SD=4,46$).

3.2 Material

3.2.1 *Gesangsdarbietungen als Gegenstand der Evaluation*

Ziel war es, für die Evaluationsaufgaben Hörbeispiele von Schüler:innen zu finden, die verschiedene Musikstücke mit unterschiedlichem Anforderungsniveau beinhalten. Zugleich sollten für diese Darbietungsleistungen bereits Bewertungen mit hoher

Tab. 2 Auswahl der Gesangsdarbietungen als Urteilsobjekte

Schülercode	Geschlecht	WLE Gesang	Niveaustufe
1	w	-2,370	0
2	m	-0,415	1
3	w	-0,540	1
4	w	-0,415	1
5	m	1,594	2
6	w	1,435	2
7	m	3,198	3
8	m	4,046	3
9	w	3,500	3

Berechnung des WLE Gesang und der Niveaustufe auf Grundlage von Hasselhorn (2015).

Übereinstimmung zwischen den beurteilenden Personen vorliegen. Darüber hinaus sollte die tatsächliche Leistungsverteilung eine Breite aufweisen, die der Beurteilungsbreite des Evaluationsinstrumentes entsprach (Hasselhorn 2015), unter dessen Anwendung die Studienteilnehmer:innen ihr Urteil finden und fällen sollten.

In einem mehrschrittigen Verfahren konnte diese Zielsetzung dadurch erreicht werden, dass auf Einzelgesangsaufnahmen zurückgegriffen wurde, die im Zuge der Kompetenztestentwicklung zur Erhebung der musikpraktischen Kompetenz von Schüler:innen entstanden sind (Hasselhorn 2015). Da diese aufgefordert wurden, mehrmals die Gesangsstimme verschiedener, instrumental begleiteter Lieder mit variierendem Schwierigkeitsgrad zu singen, lagen somit mehrere Einzelgesangsaufnahmen von Liedern mit unterschiedlichem Schwierigkeitsgrad pro Schüler:in vor. Hasselhorn (2015) konnte bereits während der Testentwicklung zeigen, dass diese Darbietungsleistungen mit Hilfe seiner Skala („HTR-G“ Skala, s. Abb. 3) durch geschulte Musikstudierende als Urteiler:innen äußerst reliabel beurteilt werden, sodass noch vor Beginn der vorliegenden Studie Leistungsbeurteilungen aller eingesetzten Aufnahmen vorlagen („tatsächliche“ Leistung, s. oben). Zudem ist nachgewiesen, dass das skalengestützte Urteilsverhalten geschulter Musikstudierender einen hohen Ähnlichkeitsgrad mit dem Urteilsverhalten von Musiklehrkräften aufweist (Hasselhorn 2015, S. 90 ff.). Für die vorliegende Studie wurden aus dem Korpus vorhandener Einzelgesangsaufnahmen in einem weiteren Schritt nur solche Gesangsaufnahmen ausgewählt, deren Beurteilung eine überaus hohe Übereinstimmung aufwiesen ($ICC(2,2) \geq 0,88$). Neben dem Kriterium der Urteilshomogenität bestimmten zu-

Tab. 3 Übersicht über das Anforderungsniveau der Musikstücke

Stück	Item-Schwierigkeit	Ungewichteter MNSQ	Verwendung
Guten Abend, gut' Nacht	-0,188 (0,058)	1,08	Evaluation
Bruder Jakob	-0,758 (0,069)	1,20	Evaluation
Kalte Sterne	0,763 (0,067)	1,07	Evaluation
Happy Birthday	-0,088 (0,067)	1,09	Prognose
Country Roads	-0,297 (0,077)	0,94	Prognose

Item-Schwierigkeit und ungewichteter MNSQ berechnet auf Grundlage von Hasselhorn (2015).

Guten Abend, gut' Nacht



Abb. 1 Evaluationsstück für die Studienbedingung mit geringer Informationsvielfalt

sätzlich das *Geschlecht* und das gezeigte *Leistungsniveau der Schüler:innen* sowie die *Aufgabenschwierigkeit* (Item-Schwierigkeit: $Min = -0,758$ $Max = 0,763$; s. Tab. 2 und Tab. 3) der jeweiligen Lieder die Auswahl der Darbietungsleistungen.

Alle Stimuli wurden mit der Software Audacity (3.1.3) auf -20 LUFS (Loudness Units Relative to Full Scale, Einheit für die Messung der Lautheit einer Audiodatei) normalisiert (Audacity-Team 2021). Anschließend wurden Störgeräusche in den Einzelgesangsaufnahmen mittels Spektralanalyse (Fast-Fourier-Transformation) identifiziert und die für die Störungen verantwortlichen Frequenzen je nach ihrer Beschaffenheit durch die Anwendung von Rausch- oder Kerbfiltern entfernt. Damit die Teilnehmenden den Notentext zu den Aufnahmen mitlesen konnten, wurden die Aufnahmen zusammen mit dem dazugehörigen Notenbild (Standbild) als Videosequenz finalisiert. Zur Untersuchung der Forschungsfrage 3 wurden zwei Varianten

Abb. 2 Ausschnitte aus exemplarisch ausgewählten Evaluationsstücken zum Erzeugen der Studienbedingung mit hoher Informationsvielfalt



Abb. 3 Adaptierte Fassung der verwendeten HTR-G-Skala (vgl. Hasselhorn 2015, S. 82)

Wie gut, vermuten Sie, wird die Schülerin / der Schüler das Lied singen?

Die Schülerin / der Schüler wird ...

- 1: das Lied nahezu oder völlig korrekt singen.
- 2: das Lied mit einiger Genauigkeit singen und beginnt in der angegebenen Tonart.
- 3: das Lied mit einiger Genauigkeit singen, beginnt aber in einer anderen als der vorgegebenen Tonart oder moduliert innerhalb des Liedes.
- 4: den melodischen Umriss mit deutlich falschen Tönhöhen singen.
- 5: das Lied mit einem anderen melodischen Umriss als das Lied besitzt singen
- 6: nicht singen. Eine sinnvolle Bewertung ist nicht möglich.

Happy Birthday



Abb. 4 Illustrativ ausgewähltes Notenbeispiel, anhand dessen Leistungsprognosen erstellt werden sollten

erzeugt. In der Variante mit geringer Informationsvielfalt enthielt das Video nur Darbietungen des Liedes *Guten Abend, gut' Nacht* (s. Abb. 1) in voller Länge. Für die Variante mit hoher Informationsvielfalt wurde ein Zusammenschnitt mit Ausschnitten aus den drei Liedern *Bruder Jakob*, *Guten Abend, gut' Nacht* und *Kalte Sterne* erstellt, die mit einem Fade-out und Fade-in von je einer Sekunde sowie einer Pause von zwei Sekunden miteinander verbunden wurden. Die Zuteilung der Stücke zur jeweiligen Experimentalbedingung erfolgte anhand deren Aufgabenschwierigkeiten. So wurde *Guten Abend, gut' Nacht* als Kernstück für beide Bedingungen ausgewählt, da es sich um ein Stück mittlerer Schwierigkeit (s. Tab. 3) handelt. In der Variante mit hoher Informationsvielfalt wurden zusätzlich ein vergleichsweise leichteres und schwereres Stück ergänzt.

Um den Teilnehmenden den gehörten Ausschnitt im Notentext direkt sichtbar zu machen, wurde dieser mit einem roten Kasten gekennzeichnet und die nicht gehörten Ausschnitte wurden verschwommen dargestellt (s. Abb. 2). Zur Standardisierung wurden alle Videos auf eine Gesamtlänge von 46 s geschnitten, wobei in allen Bedingungen jeweils 42 s Gesang zu hören war. Als Bewertungsskala diente die HTR-G-Skala des KOPRA-M-Kompetenztests (Hasselhorn 2015, S. 82), wobei für die Prognoseaufgaben eine sprachliche Modifizierung vorgenommen wurde (s. Abb. 3).

3.3 Aufgaben zur prognostischen Leistungsbeurteilung

Zur Erhebung der Prognoseleistung wurden zwei weitere Musikstücke ausgewählt (s. Abb. 4), die den Studienteilnehmer:innen ausschließlich in schriftlich fixierter Form (d. h. als Notentexte) ohne dazugehörige Gesangsdarbietungen vorgelegt wurden. Zu diesen beiden Musikstücken existieren dennoch Gesangsaufnahmen aller ausgewählten Schüler:innen sowie reliable Leistungsbewertungen (s. Tab. 3).

3.4 Design

Die Studie wurde als Onlinestudie über www.soscisurvey.de (Leiner 2022) durchgeführt. Zu Beginn der Onlinestudie wurde überprüft, ob die technischen Voraussetzungen zur erfolgreichen Teilnahme erfüllt waren. Hierfür mussten zwei Ziffern korrekt wiedergeben werden, die in einem Video auditiv und visuell dargeboten wurden. Anschließend erhielten die Teilnehmenden Informationen über den Ablauf der Studie. Im nächsten Schritt erfolgte eine randomisierte Zuteilung der Teilnehmenden zu einer der beiden Experimentalbedingungen (hohe/geringe Informationsvielfalt). Die Zuteilung zu einer Bedingung bestimmte für den weiteren Verlauf, ob die Teilnehmenden nur Gesangsaufnahmen des Lieds *Guten Abend, gut' Nacht* präsentiert bekamen (Bedingung 1: geringe Informationsvielfalt) oder den Zusammenschnitt

dreier Liedausschnitte aus *Bruder Jakob*, *Guten Abend, gut' Nacht* und *Kalte Sterne* (Bedingung 2: hohe Informationsvielfalt). Es folgte ein Angebot, sich mit dem musikalischen Material der jeweils zugeteilten Bedingung und der Bewertungsskala vertraut zu machen, indem der Notentext des im Folgenden gehörten Stücks bzw. der gehörten Stücke ohne auditive Information sowie die HTR-G-Skala vorgelegt wurden.

Daran schloss sich die Leistungsevaluations- und Prognosephase an, bei der die Teilnehmenden zu 10 Schüler:innen Urteile auf Grundlage der HTR-G-Skala fällen mussten. Der Ablauf je Schüler:in erfolgte dabei immer identisch und verlief in drei Phasen: In der ersten Phase (Darbietung/Vorstellung) wurde die zur Experimentalbedingung passende Gesangsdarbietung (ein Stück oder drei Stücke) eines Schülers oder einer Schülerin präsentiert. Die Teilnehmenden konnten dabei jede Gesangsdarbietung nur einmal anhören und mit dem Fragebogen erst fortfahren, nachdem die Aufnahme vollständig abgespielt worden war. Daran schloss sich die zweite Phase (Evaluation) an, in der die dargebotene Leistung evaluiert wurde. Abschließend wurde in der dritten Phase (Prognose) anhand der Notentexte von *Country Roads* und *Happy Birthday* (s. Abb. 4) eine Prognose über die mutmaßliche Leistung des/der in Phase 1 gehörten Schülers oder Schülerin zu diesen beiden Stücken abgegeben. Anhand des Antwortverhaltens der Studienteilnehmer:innen in der Evaluations- und Prognosephase konnten die beiden abhängigen Variablen *Urteilsgenauigkeit der Evaluation* (AV_1) und *Urteilsgenauigkeit der Prognose* (AV_2), die sich jeweils aus den drei Komponenten *Niveau*-, *Differenzierungs*- und *Rangkomponenten* zusammensetzten, nach Schrader und Helmke (1987) berechnet werden. Als unabhängige Variablen wurden die unterschiedlichen Prognosestücke (UV_1) sowie die Experimentalbedingung der Informationsvielfalt (UV_2) verwendet. Zur Vermeidung von Positionseffekten wurde die Reihenfolge der Darbietungen zwischen den Versuchspersonen randomisiert, wobei das Video des Schülers 5 für alle Versuchspersonen wiederholt an den Positionen 2 und 10 präsentiert wurde. Zum Schluss wurde der sozioökonomische Hintergrund der Versuchspersonen erfasst.

3.5 Datenanalyse

Die Datenanalyse erfolgte mit dem Programm RStudio (Version 2024.04.2), dessen R-Umgebung (Version 2024.06.14) durch die Pakete „psych“ (Version 2.4.5.26), „tidyverse“ (Version 2.0.0) und „effsize“ (Version 0.8.1) erweitert wurde. Im ersten Schritt wurden die abhängigen Variablen *Urteilsgenauigkeit der Evaluation* (AV_1) und *Urteilsgenauigkeit der Prognose* (AV_2) bestimmt. Hierfür wurden aus dem Antwortverhalten der Teilnehmenden für jede Versuchsperson einzeln die Niveau-, Differenzierungs- und Rangkomponente nach Schrader und Helmke (1987) sowohl für die Evaluation als auch für die Prognose berechnet (s. Tab. 4). Als Referenzwerte („tatsächliche Leistung“, s. oben) dienten die Bewertungen der geschulten Rater des KOPRA-M-Kompetenztests (Hasselhorn 2015) (s. Tab. 5). Zur Bestimmung der Referenzwerte für die Evaluation der gehörten Leistung der Experimentalbedingung 2 (hohe Informationsvielfalt) wurde der auf eine ganze Zahl gerundete Mittelwert der drei Stücke *Bruder Jakob*, *Guten Abend, gut' Nacht* und *Kalte Sterne* verwendet. Für die Prognose erfolgte die Bildung der Komponenten der Urteilsgenauigkeit für

Tab. 4 Berechnung der Komponenten der Urteilsgenauigkeit nach Schrader und Helmke (1987)

Komponente	Berechnung	Idealwert	Referenzwerte
Niveau	$ML\text{-Urteil} - MS\text{-Leistung}$	0	Tendenz zur Überschätzung
Differenzierung	$\sigma L\text{-Urteil} / \sigma S\text{-Leistung}$	1	Keine einheitliche Tendenz
Rang	$\rho L\text{-Urteil}; S\text{-Leistung}$	1	$r = 0,68^a$ $r = 0,76^b$

Formeln entnommen aus Schrader und Helmke (1987, S. 30f.), Referenzwerte entnommen aus Südkamp et al. (2012) sowie Urhahne und Wijnia (2021)

^a Referenzwert für uninformierte Urteile

^b Referenzwert für informierte Urteile.

die beiden Stücke *Country Roads* und *Happy Birthday* zunächst einzeln, um die Fragestellung 2 untersuchen zu können, welchen Einfluss das Prognosestück (UV_1) auf die Urteilsgenauigkeit der Prognose (AV_2) hat. Zusätzlich wurden Gesamtprognosekennwerte berechnet, bei denen sowohl für die Leistungseinschätzungen der Teilnehmenden als auch für das Referenzurteil (Expertenurteile) jeweils der gerundete Mittelwert der beiden Prognosestücke verwendet wurde.

Zur Beantwortung der ersten Fragestellung, wie genau Studierende als angehende Musiklehrkräfte die Gesangsleistungen der Schüler:innen evaluieren und prognostizieren können, wurden zunächst deskriptive Werte der drei Komponenten für Evaluations- und Prognoseleistung berechnet (H_{1a} und H_{1b} , s. Tab. 1). Anschließend wurden für die Niveau- und Differenzierungskomponente Einstichproben-*t*-Tests gegen die theoretischen Idealwerte $x=0$ (Niveauelemente) und $x=1$ (Differenzierungskomponente) durchgeführt (H_{1c} bis H_{1f} , s. Tab. 1).

Die Überprüfung, ob den Studierenden bei der Leistungsprognose die stückspezifische Einschätzung des Anforderungsniveau gelingt (Fragestellung 2), erfolgte über die aufgabenspezifische Urteilsgenauigkeit (Aufgaben-Personen-Wechselwirkungskomponente) (Schrader 1989). Dafür wurden die Komponenten der Urteilsgenauigkeit der beiden Prognosestücke *Country Roads* und *Happy Birthday* mittels *t*-Test für abhängige Stichproben miteinander verglichen (H_{1g} bis H_{1i} , s. Tab. 1). Da

Tab. 5 Überblick über die Schülerleistungen

Schülercode	Evaluation		Prognose		Gesamtprognose
	Gering	Hoch	Country Roads	Happy Birthday	
1	5	5	5	4	4,5
2	4	4	2	4	3
3	4	4	4	4	4
4	2	3	4	4	4
5	2	2	2	2	2
6	1	2	2	2	2
7	1	1	1	1	1
8	1	1	2	2	2
9	1	1	1	1	1

Schülerleistung ermittelt in Hasselhorn (2015). Bewertet auf Grundlage der HTR-G-Skala (s. Abb. 3). Gesamtprognosewert berechnet aus dem Mittelwert von *Country Roads* und *Happy Birthday*.

die Rangkomponente für alle Teilnehmenden einen nicht äquidistanten Korrelationswert mit begrenztem Wertebereich darstellt, wurden für die weitere Auswertung die Werte dieser Komponente zunächst mittels Fisher-z-Transformation in annähernd intervallskalierte Werte überführt (Rasch et al. 2014, S. 87). Ob die Experimentalbedingung *Informationsvielfalt* (UV_2) einen positiven Einfluss auf die Urteilsgenauigkeit der Teilnehmenden in Evaluation (AV_1) und Prognose (AV_2) hatte (H_{1j} bis H_{1o} , s. Tab. 1), wurde mittels *t*-Tests für unabhängige Stichproben überprüft.

Da im Rahmen der Datenanalyse insgesamt 13 zusammenhängende *t*-Tests durchgeführt werden, wurde aufgrund der Alpha-Fehler-Kumulierung das Signifikanzniveau durch eine Bonferroni-Korrektur auf $\alpha < 0,004$ angepasst (Sedlmeier und Renkewitz 2013, S. 421).

4 Ergebnisse

Wie genau Studierende als angehende Musiklehrkräfte die Gesangsleistungen von Schüler:innen evaluieren und prognostizieren können (Fragestellung 1), zeigt Tab. 6. Die Ergebnisse legen nahe, dass Studierende ein ähnliches Urteilsverhalten wie Fachlehrkräfte zeigen (H_{1a} und H_{1b} , s. Tab. 1). So wiesen die Rangkomponenten des Evaluations- als auch Prognoseverhaltens der Studierenden eine hohe gemeinsame Varianz mit dem Urteilsverhalten der geschulten Rater (Hasselhorn 2015, s. Tab. 5) auf ($r_{\text{Evaluation}} = 0,88$; $SD = 0,02$; $r_{\text{Prognose}} = 0,81$; $SD = 0,04$), deren Urteilsverhalten mit dem von Musiklehrkräften vergleichbar ist (Hasselhorn 2015). Die Werte der Rangkomponente fallen zudem entsprechend früherer Untersuchungen über das Urteilsverhalten von Lehrkräften anderer Fächer aus. So konnten Südkamp et al. (2012) in ihrer Metaanalyse zeigen, dass bei informierten Urteilen eine Rangkomponentenprognose in Höhe von $r = 0,76$ vorliegt.

Für die Niveaukomponente der Prognose (H_{1c} , s. Tab. 1) konnte kein signifikanter Unterschied zum theoretisch begründeten Erwartungswert ermittelt werden ($t = 0,18$; $p = 0,82$; $d = 0,02$) (s. Tab. 7). Obwohl somit die Prognose der untersuchten Stichprobe insgesamt eine hohe Übereinstimmung mit der tatsächlichen (summativ-evaluativen) Leistungsbeurteilung von Musiklehrkräften aufwies, zeigten die Einzel-

Tab. 6 Deskriptive Statistik der Komponenten der Urteilsgenauigkeit

Komponenten	<i>M</i>	<i>SD</i>	<i>Min</i>	<i>Max</i>
<i>Evaluation</i>				
Niveau	−0,37	0,49	−1,89	0,44
Differenzierung	0,89	0,17	0,40	1,30
Rang	0,88	0,02	0,35	1,00
<i>Prognose</i>				
Niveau	0,01	0,50	−1,39	1,00
Differenzierung	0,92	0,20	0,42	1,46
Rang	0,81	0,04	0,33	0,99

Standardabweichung der Rangkomponente berechnet nach Borenstein und Hedges (2019, S. 221); Mittelwerte der Rangkomponenten sind in ihrer Originalmetrik (zuvor über Fisher-z-Transformation berechnet und rücktransformiert) dargestellt.

Tab. 7 Vergleich mit theoretischen Idealwerten

Komponenten	<i>M</i>	<i>SD</i>	Idealwert	<i>t</i> (97)	<i>p</i>	Cohen's <i>d</i> (95 % CI)
<i>Evaluation</i>						
Niveau	-0,37	0,49	0	-7,47	<0,001*	-0,75 (-1,17; -0,34)
Differenzierung	0,89	0,17	1	-6,74	<0,001*	-0,68 (-1,09; -0,27)
<i>Prognose</i>						
Niveau	0,01	0,50	0	0,18	0,86	0,02 (-0,38; 0,42)
Differenzierung	0,92	0,20	1	-3,86	<0,001*	-0,39 (-0,79; 0,02)

* $p < 0,004$ nach Bonferroni-Korrektur.

urteile der Studierenden eine auffällig große durchschnittliche Streuung ($SD = 0,50$) sowie eine erheblich große Spannweite ($min = -1,39$; $max = 1,00$).

Die Differenzierungskomponente der Prognose sowie die Niveau- und Differenzierungskomponente der Evaluation wichen von ihren Erwartungswerten ($E_{\text{Niveau}} = 0$, $E_{\text{Differenzierung}} = 1$) ab (s. Tab. 7). Die untersuchte Stichprobe zeigte in ihrer Prognose eine Unterschätzung der antizipierten Leistungsstreuung im Ausmaß einer mittleren Effektstärke ($t = -3,86$; $p < 0,001$; $d = -0,39$). Dieser Befund legt die Schlussfolgerung nahe, dass Studierende mit ihrer Tendenz zur Unterschätzung der Leistungsheterogenität einer Lerngruppe Schwächen in der Differenzierungskomponente ihres prognostischen Urteilsverhaltens aufweisen (H_{1d} , s. Tab. 1). Gleiches gilt für die Niveau- und Differenzierungskomponente ihres Evaluationsverhaltens: Während hier eine deutliche Unterschätzung des mittleren Leistungsniveaus (H_{1e} , s. Tab. 1) zu beobachten war (Niveauelemente: $t = -7,47$; $p < 0,001$; $d = -0,75$), spiegelte sich im Evaluationsverhalten der Stichprobe zusätzlich eine Unterschätzung der Leistungsheterogenität der Schüler:innen (H_{1f} , s. Tab. 1) wider (Differenzierungskomponente: $t = -6,74$; $p < 0,001$; $d = -0,68$). Auffällig ist der starke signifikante Zusammenhang zwischen den jeweiligen Komponentenausprägungen von Evaluations- und Prognoseverhalten (s. Tab. 8). So korrelierten zum einen die Niveauelemente von Evaluation und Prognose stark miteinander ($r = 0,84$; $p < 0,001$), zum anderen zeigten ihre Differenzierungskomponenten auch einen positiven Zusammenhang ($r = 0,66$; $p < 0,001$), während die Rangkomponenten des Evaluations- und Prognoseverhaltens zwar ebenfalls einen positiven Zusammenhang aufwiesen, dieser jedoch mit einer mittleren Effektstärke deutlich geringer ausfiel ($r = 0,42$; $p < 0,001$).

Tab. 8 Zusammenhang zwischen den Komponenten der Urteilsgenauigkeit

Variablen	1.	2.	3.	4.	5.
1. Niveauevaluation	–				
2. Differenzierungsevaluation	-0,15	–			
3. Rangevaluation	0,38***	0,02	–		
4. Niveauprognose	0,84***	-0,12	0,30**	–	
5. Differenzierungsprognose	-0,17	0,66***	-0,04	-0,31**	–
6. Rangprognose	0,21*	0,21*	0,42***	0,12	0,24*

* $p < 0,05$, ** $p < 0,01$, *** $p < 0,001$.

Tab. 9 Urteilsgenauigkeit der einzelnen Prognosestücke

Komponente	Country Roads		Happy Birthday		<i>t</i>	<i>p</i>	Cohen's <i>d</i> (95 % CI)
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>			
Niveau	-0,29	0,53	0,31	0,54	-16,38	<0,001*	-1,13 (-1,30; -0,96)
Differenzierung	0,92	0,18	0,91	0,21	0,52	0,60	0,04 (-0,10; 0,17)
Rang	0,91 ^a	0,29	0,99 ^b	0,35	-2,46	0,02	-0,27 (-0,49; -0,05)

* $p < 0,004$ nach Bonferroni-Korrektur; zweiseitig; $df = 97$ für alle *t*-Tests; für Rangkomponente Fisher-z-transformierte Werte verwendet; Originalmetrik nach Rücktransformation ^a0,72 ^b0,76.

Die Ergebnisse der Fragestellung 2 (s. Tab. 9) zeigten lediglich für die Niveau-komponente (H_{lg} , s. Tab. 1) den erwarteten signifikanten Unterschied im Ausmaß einer großen Effektstärke zwischen den beiden Prognosestücken ($t = -16,38$; $p < 0,001$; $d = -1,13$). Während die Teilnehmenden das mittlere Leistungsniveau der Schüler:innen bei *Country Roads* im Mittel unterschätzten, wurde deren Leistung für das Stück *Happy Birthday* dagegen überschätzt, wobei eine vergleichbare Urteils-heterogenität vorlag ($SD_{Country Roads} = 0,53$; $SD_{Happy Birthday} = 0,54$). Für die Differenzierungskomponente (H_{lh} , s. Tab. 1) und Rangkomponente (H_{li} , s. Tab. 1) konnten hingegen keine signifikanten Unterschiede zwischen beiden Prognosestücken gefunden werden (Differenzierungskomponente: $t = 0,52$; $p = 0,60$; $d = 0,04$; Rangkomponente: $t = -2,46$; $p = 0,02$; $d = -0,27$). Somit zeigten sich keine systematischen Unterschiede in der prognostizierten Leistungsstreuung und Rangreihenfolge der Schülerleistung,

Tab. 10 Einfluss der Informationsvielfalt auf die Komponenten der Urteilsgenauigkeit

Komponente	Informationsvielfalt				<i>t</i>	<i>df</i>	<i>p</i>	Cohen's <i>d</i> (95 % CI)
	Gering		Hoch					
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>				
<i>Evaluation</i>								
Niveau	−0,29	0,46	−0,46	0,51	1,74	93,13	0,09	0,35 (−0,05; 0,76)
Differenzierung	0,86	0,16	0,92	0,17	−1,73	92,99	0,09	−0,35 (−0,76; 0,05)
Rang	1,40 ^a	0,42	1,37 ^b	0,36	0,34	95,67	0,73	0,07 (−0,33; 0,47)
<i>Prognose</i>								
Niveau	0,11	0,48	−0,10	0,51	2,08	94,34	0,04	0,42 (0,02; 0,83)
Differenzierung	0,93	0,19	0,92	0,20	0,12	94,67	0,91	0,02 (−0,38; 0,42)
Rang	1,00 ^c	0,28	1,23 ^d	0,42	−3,17	80,00	0,002*	−0,65 (−1,06; −0,24)

Welch's *t*-Test; * $p < 0,004$ nach Bonferroni-Korrektur; zweiseitig; für Rangkomponente wurden Fisher-z-transformierte Werte verwendet Originalmetrik nach Rücktransformation ^a0,89 ^b0,88 ^c0,76 ^d0,84.

die auf spezifischen Eigenschaften der jeweiligen Musikstücke (*Country Roads* und *Happy Birthday*) beruhen würden.

Eine höhere Informationsvielfalt hat auf die prognostische Urteilsgenauigkeit (Fragestellung 3, s. Tab. 10) nur teilweise einen nachweisbaren positiven Einfluss. So wurde entgegen der Hypothesen H_{1j} und H_{1k} (s. Tab. 1) weder für die Niveau- ($t=2,08$; $p=0,04$; $d=0,42$) noch für die Differenzierungskomponente ($t=0,21$; $p=0,92$; $d=0,02$) ein solcher Einfluss gefunden. Lediglich für die Rangkomponente konnte hypothesenkonform (H_{1l} , s. Tab. 1) bestätigt werden, dass die Prognoseeinschätzung unter der Bedingung einer hohen Informationsvielfalt treffsicherer gelang ($t=-3,17$; $p=0,002$; $d=-0,65$). Hieraus folgt, dass mehr relevante Informationen teilweise mit einem diagnostischen Mehrwert für Studierende einhergehen: Sie ermöglichen zwar eine präzisere Einschätzung der Leistungsanordnungen (Rangkomponente), zeigen jedoch keine weiteren Auswirkungen auf die Prognose des mittleren Leistungsniveaus oder der Leistungsheterogenität. Darüber hinaus bestätigt sich hypothesenkonform (H_{1m} , H_{1n} , H_{1o} , s. Tab. 1), dass die Informationsvielfalt weder einen Einfluss auf die Niveauebene ($t=1,74$; $p=0,09$; $d=0,35$), Differenzierungskomponente ($t=-1,73$; $p=0,09$; $d=-0,35$) noch auf die Rangkomponente der Evaluation hat ($t=0,34$; $p=0,73$; $d=0,07$). Die Urteilsgenauigkeit fällt somit unter Experimental- und Kontrollbedingung vergleichbar aus.

5 Diskussion

Die vorliegende Studie untersuchte die Urteilsgenauigkeit von Studierenden als angehende Musiklehrkräfte bei der Evaluation und Prognose von Schülergesangsleistungen. Dabei interessierten auch die Fragen nach der Güte der aufgabenspezifischen Urteilsgenauigkeit sowie dem Einfluss der Informationsvielfalt auf die Genauigkeit der Urteile. Hierfür sollten die Teilnehmenden auf Grundlage einer Schülerdarbietung zukünftige Leistungen anhand eines Notentextes antizipieren. Die Ergebnisse legen den Schluss nahe, dass den Studierenden die Einschätzung der Leistungsranfolge der Schüler:innen (*Rangkomponente*) grundsätzlich gelingt (vgl. H_{1a} und H_{1b} , s. Tab. 1). In der Prognose entspricht die Niveauebene der Studierenden im Mittel dem Idealwert (vgl. H_{1c} , s. Tab. 1). Die große Streuung der Niveauprogosekomponente sowie die große Bandbreite weisen darauf hin, dass bei Betrachtung der Urteilsgenauigkeit auf Individualebene erhebliche Unterschiede in der Urteilsgüte vorhanden sind. Ähnliche Ergebnisse zeigen sich in der Differenzierungs- (H_{1d} , s. Tab. 1) und Rangkomponente. Während hier Studierende vereinzelt nahezu die Idealwerte erreichten, gab es auch Teilnehmende, deren Werte in der Rangkomponente unterhalb des Rate-Niveaus liegen. Somit zeigt sich, dass den Studierenden die Prognose von Einzelgesangsleistungen im Mittel zwar gelang, jedoch auf Individualebene erhebliche Unterschiede in der Qualität der Urteile vorhanden sind. Die Ursachen für die ausgeprägte Heterogenität im Urteilsverhalten konnten im Rahmen dieser Studie nicht ermittelt werden. Jedoch legen die Ergebnisse nahe, dass für einen Teil der angehenden Musiklehrkräfte ein gezieltes Urteilstraining notwendig ist.

Im Hinblick auf die Evaluation von Schülerleistungen (vgl. H_{1e} und H_{1f} , s. Tab. 1) ließ sich eine Tendenz zur Unterschätzung des mittleren Leistungsniveaus beobachten, was darauf hindeutet, dass die Studierenden die Leistungen der Schüler:innen strenger bewerteten als die geschulten Rater. Eine mögliche Ursache könnte darin liegen, dass den Studierenden eine Referenz fehlt, was als durchschnittliche Leistung von Schüler:innen der 9. Klasse erwartet werden kann. Eine weitere Erklärung könnte eine überwiegende Gewichtung ästhetischer Kriterien anstelle einer pädagogisch orientierten Evaluation sein. Da im Schulmusikstudium an Musikhochschulen der Fokus auf die künstlerische Entwicklung der Studierenden gelegt wird (Hofbauer und Wehen 2024), geben die Ergebnisse einen Hinweis darauf, dass der selbst erfahrene hohe Leistungsanspruch auf die Bewertung von Schülerleistungen übertragen wird. Inwiefern eine Auseinandersetzung mit den Herausforderungen und typischen Schwierigkeiten von Schüler:innen bei Gesangsaufgaben Studierende zu genaueren Urteilen verhilft, müsste durch Folgestudien untersucht werden. Die Ergebnisse dieser Studie weisen jedoch darauf hin, dass hier noch Optimierungspotential in der Lehrkräfteausbildung vorhanden ist. Auffällig ist ein starker Zusammenhang zwischen den Evaluations- und Prognosekomponenten. Dies ist ein Zeichen dafür, dass die Studierenden, denen es gelingt, die gehörten Leistungen von Schüler:innen genau zu evaluieren, eher in der Lage sind, zukünftige Leistungen treffsicherer zu antizipieren.

Darüber hinaus konnte gezeigt werden, dass die Werte der Niveauprognosekomponente für die einzelnen Prognosestücke unterschiedlich hoch sind (vgl. H_{1g} , s. Tab. 1): Während das mittlere Leistungsniveau des Stücks *Happy Birthday* überschätzt wurde, wurde das mittlere Leistungsniveau des Stücks *Country Roads* unterschätzt. Ein solcher Unterschied konnte weder für die Differenzierungs- (H_{1h} , s. Tab. 1) noch für die Rangkomponente (H_{1i} , s. Tab. 1) festgestellt werden. Im Gegensatz zur Einschätzung des Leistungsniveaus scheint das Prognosestück demnach keinen Einfluss auf die Prognose der Leistungsheterogenität und der Rangreihenfolge der Schüler:innen im Vergleich untereinander zu haben. Dieser Befund offenbart, dass den Studierenden die Einschätzung der Aufgabenschwierigkeiten der beiden verwendeten Prognosestücke nur unzureichend gelang, ohne dass dabei die Einschätzung der Leistungsbreite oder -reihenfolge beeinträchtigt wurde. Darüber hinaus ist dieses Ergebnis ein weiterer Hinweis darauf, dass der Vergleich der Niveauprognosekomponente mit dem Idealwert differenziert zu betrachten ist, da diese Komponente aus den beiden unterschiedlich eingeschätzten Prognosestücken zusammengesetzt wird. Da die unterschätzende Abweichung von *Country Roads* und die überschätzende Abweichung von *Happy Birthday* etwa gleich groß ist, besteht die Möglichkeit, dass sich diese beiden Effekte in der Niveauprognosekomponente, die aus den beiden Prognosestücken gebildet wird, gegenseitig aufheben. Somit wird für die aggregierten Werte im Hinblick auf die Niveauprognosekomponente eine hohe Urteilsgenauigkeit suggeriert, die bei differenzierter Betrachtung nicht bestätigt werden kann. Die Ursache für die Unterschätzung der Aufgabenschwierigkeit des Stücks *Happy Birthday*, das für die Schüler:innen der 9. Jahrgangsstufe ein Stück mittlerer Schwierigkeit ist (s. Tab. 3), könnte in der Annahme liegen, dass den Schüler:innen das Stück aufgrund seiner Alltagsrelevanz bekannt ist, sodass die Herausforderungen, die insbesondere in den großen Intervallsprüngen am Ende des Stücks vorhanden sind, nicht berücksichtigt

wurden. Das Stück *Country Roads* ist aufgrund des geringen Ambitus (Tonumfang) und der überwiegend kleinen Intervalle (kleine und große Sekunde) für die Schüler:innen der 9. Jahrgangsstufe ein leichtes Stück (s. Tab. 3). Dass die Studierenden das Stück schwieriger einschätzten, könnte auf die Grenzen der in dieser Studie verwendeten HTR-G-Skala für diagnostische Zwecke hinweisen. Zwar sind Bewertungen, die mit Hilfe dieses Bewertungsinstruments entstehen, reliabler als Bewertungen von Skalen, die mehrere Kriterien umfassen (vgl. Hasselhorn 2015), jedoch beschränkt sich die Beurteilung hier auf ein nicht differenzierendes Gesamturteil. Auch wenn die Ursache der Überschätzung des Leistungsniveaus der Schüler:innen an dieser Stelle nicht zweifelsfrei ermittelt werden kann, können die Ergebnisse ein Hinweis darauf sein, dass die Studierenden in der Einschätzung zusätzliche Kriterien heranziehen, die in der HTR-G-Skala nicht explizit abgebildet sind.

In den Ergebnissen zum Stückvergleich ist auffällig, dass die Streuung der Niveauekomponente unabhängig von der mittleren Tendenz zur Über- oder Unterschätzung in allen Fällen vergleichbar hoch ist. Das bedeutet, dass hier bei den Ratern ein stabiles Rauschen (*noise*) nach Kahneman et al. (2021) vorhanden ist, das unabhängig von den zentralen Urteilstendenzen (*bias*) existiert. Dieser Befund deckt sich mit den empirischen Ergebnissen im außerschulischen und professionellen Musikbereich, wo ebenfalls ein solches Rauschen beobachtbar ist (Passarotto et al. 2023). Überraschend ist dieses Ergebnis dennoch, da die Uneinigkeit darüber, was eine gute Leistung ist, offenbar auch auf einem niedrigen Leistungsniveau vorhanden ist. Im professionellen Bereich ist die Leistungsbeurteilung weniger von technischen Kriterien, z. B. der Anzahl richtiger Töne, als vielmehr von der persönlichen Interpretation des Werkes abhängig, sodass die Leistungsbeurteilung eng mit der ästhetischen Urteilsbildung und den persönlichen Vorlieben der Rater zusammenhängt. Im Vergleich dazu könnte angenommen werden, dass im schulischen Bereich die Schüler:innen nicht zuletzt aufgrund der großen Leistungsheterogenität als Anfänger:innen betrachtet werden. Dadurch könnten technische Aspekte (wie die Anzahl richtiger Töne, korrekter Rhythmus und Intonation), die objektiv messbar sind, eine höhere Gewichtung erhalten, wohingegen ästhetische Aspekte wie die Interpretation zweitrangig erscheinen. Durch das stabile Rauschen, das sich in den Daten findet, könnte der Schluss gezogen werden, dass auch hier eine enge Verknüpfung mit dem ästhetischen Urteil vorhanden ist und in die Leistungsbeurteilung einfließt.

Für die Experimentalbedingung der Informationsvielfalt konnte erwartungsgemäß ein Einfluss auf die Rangkomponente der Prognose gefunden werden (vgl. H_{11} , s. Tab. 1). Das bedeutet, dass die Teilnehmer:innen, die von den Schüler:innen Ausschnitte aus drei unterschiedlichen Gesangsstücken hörten, die Rangreihenfolge der Schülerleistungen im gegenseitigen Vergleich signifikant besser antizipieren konnten als die Teilnehmer:innen, die von den Schüler:innen ein Stück in voller Länge hörten. Überraschend ist jedoch, dass die Informationsvielfalt keinen Einfluss auf die prognostische Einschätzung des mittleren Leistungsniveaus (*Niveauekomponente*) und die Leistungsstreuung (*Differenzierungskomponente*) hatte (vgl. H_{1j} und H_{1k} , s. Tab. 1). Studierende scheinen zwar sehr gut dazu ausgebildet zu sein, Qualitätsunterschiede zwischen Schüler:innen zu erkennen; konkrete Vorstellungen für erwartbare Niveaus und Leistungsstreuungen sind dagegen (meist) nicht vorhanden. Genau diese stellen jedoch die Grundlage für einen kompetenzorientierten Unter-

richt sowie eine adaptive Unterrichtsplanung dar. Auch wenn die Stichprobe dieser Studie aus angehenden und nicht aus ausgebildeten Lehrkräften bestand, decken sich die Befunde mit den Ergebnissen bei Hasselhorn et al. (2022), die ähnliche Tendenzen bei (erfahrenen) Lehrkräften fanden. Der geringe Einfluss der Informationsvielfalt auf die Prognosegenauigkeit angehender Musiklehrkräfte könnte darüber hinaus ein Hinweis darauf sein, dass die Informationen, die aus unterschiedlichen Situationen gewonnen werden, nicht unabhängig voneinander verarbeitet und in ein Gesamturteil überführt werden. Inwiefern stattdessen andere Kriterien die Prognosegenauigkeit verbessern oder die Ergebnisse auf andere Fächer übertragbar sind, in denen ebenfalls ein großer Gestaltungsspielraum bei der Interpretation von Schülerantworten vorhanden ist oder ästhetische Kriterien angewendet werden könnten, ist zu überprüfen.

Wie erwartet konnte kein Einfluss der Informationsvielfalt auf die Evaluationskomponenten (H_{1m} bis H_{1o} , s. Tab. 1) gefunden werden. Die Qualität der evaluativen Beurteilung ist somit nicht substantiell abhängig von der Anzahl an Stücken, auf denen das Urteil basiert. Stattdessen scheint es, dass in der Evaluation die gehörte Leistung im Gesamten und unabhängig davon, ob es sich um verschiedene Aufgaben handelt, berücksichtigt wird.

Zusammenfassend zeigen angehende Musiklehrkräfte eine Urteilsgenauigkeit auf hohem Niveau, wenngleich die Komponenten der Urteilsgenauigkeit ausschließlich Informationen auf Gruppenebene bieten. Schlussfolgerungen über die individualdiagnostische Urteilsgenauigkeit lassen sich auf ihrer Grundlage dagegen nicht ziehen, da in den Kennwerten zwei aggregierte Werte (Schülergruppe und Stichprobe) konfundiert sind. So kann die Aggregation aller Bewertungsurteile der Studierenden über alle Schülerleistungen trotz vorhandener Über- und Unterschätzung bspw. dazu führen, dass sich diese im Mittel ausgleichen. Auch wenn – wie in unserer Studie – auf eine hohe Urteilsgenauigkeit als Ergebnis einer Verschränkung von Stichprobenverhaltens- und Urteilsobjekteigenschaftsausprägungen geschlossen werden kann, darf diese nicht als Individualmerkmal mit erwartbarer Ausprägungsstärke interpretiert werden. Das Ergebnis ist ein Punktschätzer für die Urteilsgenauigkeit, der für die Individualdiagnostik ungeeignet ist, weil er sich aus dem Zusammenspiel zweier Zufallsfaktoren ergibt (sogenannte cross-random-factor-Struktur), ohne dass diese in den Berechnungen nach Schrader und Helmke (1987) eine Berücksichtigung finden. Dass dieser wenig belastbar scheint, wird an den großen Streuungsmaßen von Niveau- und Differenzierungskomponente deutlich: Die beobachtete große Streuung deutet auf eine große Urteilsheterogenität hin, die auf eine ausgeprägte *Noise*-Komponente des individuellen Prognoseverhaltens von angehenden Musiklehrkräften zurückzuführen ist. Ein solches unpräzises Einschätzungsverhalten – mit einer Ungenauigkeit im Ausmaß zweier Schulnoten – hat unweigerlich zur Folge, dass im Rahmen adaptiven Unterrichts die Wahrscheinlichkeit einer signifikanten Über- oder Unterforderung der einzelnen Schüler:innen steigt. Darüber hinaus legt diese Beobachtung nahe, dass offenbar trotz kriterialer Vorgaben keine übereinstimmende Vorstellung darüber existiert, wie das musikalische Darbietungsverhalten im Sinne einer Leistung von Schüler:innen zu beurteilen ist. Die Ursache für die große Heterogenität in der Urteilsgenauigkeit konnte im Rahmen dieser Studie nicht zweifelsfrei ermittelt werden. Dass einige Teilnehmer:innen in den Komponenten der Urteilsge-

naugigkeit trotzdem Werte erreichten, die sehr nah an den theoretischen Idealwerten liegen, zeigt, dass Studierende vereinzelt bereits sehr gut in der Lage sind, die fachlichen Leistungen von Schüler:innen auf dem Niveau ausgebildeter Lehrkräfte im Fach Musik zu evaluieren und zu prognostizieren. Da jedoch nur ein geringer Anteil der Studierenden bereits in der Phase des Studiums ein vergleichbares Urteilsverhalten zeigte, sollte ein wissens- und kompetenzorientiertes Programm zur Förderung der pädagogischen Diagnostik bereits im Studium in Verbindung mit den einzelnen Orientierungs- und Praxisphasen entwickelt werden. Hierfür ist es zudem notwendig, Folgestudien durchzuführen, die den kognitiven Prozess der Urteilsbildung in den Fokus nehmen.

Auch wenn es sich bei den Komponenten der Urteilsgenauigkeit um eine etablierte Operationalisierung der pädagogischen Diagnostik handelt, die Aufschluss über die Qualität des Ergebnisses eines Urteilsprozesses geben kann, zeigt die Interpretation der Ergebnisse die Grenzen dieses Verfahrens auf. Aus diesem Grund wäre es sinnvoll, zukünftig auch andere statistische Auswertungsverfahren zu nutzen und auf statistische Modelle zurückzugreifen, die eine spezifischere Interpretation der interindividuellen Unterschiede zwischen den Teilnehmenden als Evaluator:innen und zwischen den Evaluationsobjekten ermöglichen und auf die (unberücksichtigte) Verschränkung zuvor aggregierter Werte und der damit einhergehenden Varianzproblematiken verzichten (s. auch Südkamp und Praetorius 2017, S. 151–200).

Bei der Interpretation und Einordnung der Ergebnisse dieser Studie sollte berücksichtigt werden, dass durch das experimentelle Design teilweise Idealbedingungen geschaffen wurden. Eine solche Idealbedingung ergibt sich bereits durch die Zusammensetzung der Schüler-/Leistungsstichprobe: Ihre Auswahl wies hinsichtlich ihrer WLE-Schätzer (s. Tab. 2) entweder sehr ähnliche (z.B. Schüler:in 5 und 6) oder aber stark voneinander abweichende Kompetenzniveaus auf (z.B. Schülerin 1 zu allen anderen). Den Teilnehmenden fiel es deshalb möglicherweise leichter, die Gesangsleistungen im Sinne einer Kontrastbetonung einzuschätzen. Wie sich das diagnostische Verhalten der angehenden Musiklehrkräfte verändert, wenn sich die Schüler:innen in ihren Kompetenzausprägungen graduell feiner unterscheiden, wurde im Rahmen dieser Studie nicht untersucht. Eine weitere Idealbedingung war, dass die Teilnehmenden zu einer systematischen Informationsverarbeitung angeleitet wurden, indem ihnen Urteilskriterien vorgegeben wurden, je eine Gesangsstimme pro Urteilsdurchgang präsentiert und ihnen die Möglichkeit angeboten wurde, während der Audiowiedergabe den entsprechenden Notentext vergleichend mitlesen zu können. Die so angeregte kriterial-gestützte, analytisch-systematische Informationsverarbeitung könnte folglich als Operationalisierung des Systems langsamer Denkprozesse nach Kahneman (2011) aufgefasst werden. Vor dem Hintergrund der vielseitigen situativen Anforderungen an Musiklehrkräfte beim Klassenmusizieren, die besonders dann erschwert sind, wenn die Musiklehrkraft selbst Teil der musizierenden Gruppe ist, darf vermutet werden, dass Urteile im Unterrichtsgeschehen vermutlich intuitiver und auf Grundlage von Heuristiken gefällt werden (schnelle Denkprozesse nach Kahneman). Inwiefern sich das Diagnoseverhalten von (angehenden) Lehrkräften unter den in dieser Studie verwendeten Idealbedingungen von komplexeren, ökologisch valideren Bedingungen unterscheidet, wäre mit Blick auf eine Optimierung von pädagogischer Diagnostik im Musikunterricht eine lohnende

Forschungsfrage. Diese müsste mit Sicherheit neben kognitiven Aspekten auch motivationale Gesichtspunkte sowie Normen und Überzeugungen von Musiklehrkräften berücksichtigen.

Danksagung Wir danken Timo Ruetz und Simone Buckreus vom Wissenschaftslektorat Kelly GmbH und Editage.com für das sorgfältige Lektorat des Manuskripts.

Förderung Die Open-Access-Finanzierung des Artikels wurde durch das Projekt DEAL ermöglicht.

Funding Open Access funding enabled and organized by Projekt DEAL.

Interessenkonflikt L. Samel, F. Platz und J. Hasselhorn 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

- Asmat, R., Borowiecki, K.J., & Law, M.T. (2023). Do experts and laypersons differ? Some evidence from international classical music competitions. *Journal of Economic Behavior & Organization*, 214, 270–290. <https://doi.org/10.1016/j.jebo.2023.06.024>.
- Audacity Team, T. (2021). *Audacity (Version 3.1.3)*. Software
- Bähr, J. (2001). Was sollen Kinder können? Musikalische Grundkompetenz und Qualitätsentwicklung im Musikunterricht. *Musik und Bildung*, 33(3), 24–27.
- Bellmann, K. (2020). *Das Probespiel im Orchester als Personaleignungsdiagnostik: Problemstellungen und Lösungsansätze: empirische Studien zur Performance und Bewertung von vorbereitetem Repertoire und Vomblattspiel innerhalb des Auswahlverfahrens professioneller OrchestermusikerInnen*. LIT.
- Borenstein, M., & Hedges, L.V. (2019). Effect sizes for meta-analysis. In H. Cooper, L.V. Hedges & J.C. Valentine (Hrsg.), *The handbook of research synthesis and meta-analysis* (3. Aufl., S. 207–243). Russel Sage Foundation.
- Ehninger, J., Knigge, J., Schurig, M., & Rolle, C. (2021). A new measurement instrument for music-related argumentative competence: the MARKO competency test and competency model. *Frontiers in Education*, 6, 668538. <https://doi.org/10.3389/educ.2021.668538>.
- Greuel, T. (2007). Theorie musikpädagogischer Diagnose. In T. Greuel (Hrsg.), *In Möglichkeiten denken – Qualität verbessern. Auf dem Weg zu einer musikpädagogischen Diagnostik* (S. 25–56). Gustav Bosse.
- Greuel, T., & Horst, U. (2011). Singen zwischen Wunsch und Wirklichkeit – Musikpädagogische Stimm-diagnose in der allgemeinbildenden Schule. In T. Greuel, U. Kranefeld & E. Szczepaniak (Hrsg.), *Singen und Lernen: Perspektiven auf schulische und außerschulische Vokalpraxis* (Bd. 24, S. 69–86). Shaker.
- Greuel, T., & Szczepaniak, E. (2007). Von der musikpädagogischen Diagnose zum musikalischen Arrangement. In T. Greuel (Hrsg.), *In Möglichkeiten denken – Qualität verbessern. Auf dem Weg zu einer musikpädagogischen Diagnostik* (S. 70–77). Gustav Bosse.
- Gruhn, W. (2018). Dimensionen eines musikbezogenen Lernbegriffs. In W. Gruhn & P. Rübke (Hrsg.), *Musiklernen: Bedingungen – Handlungsfelder – Positionen* (S. 16–42). Helbling.

- Hasselhorn, J. (2015). *Messbarkeit musikpraktischer Kompetenzen von Schülerinnen und Schülern: Entwicklung und empirische Validierung eines Kompetenzmodells*. Waxmann.
- Hasselhorn, J., & Lehmann, A. C. (2015). Leistungsheterogenität im Musikunterricht. Eine empirische Untersuchung zu Leistungsunterschieden im Bereich der Musikpraxis in Jahrgangsstufe 9. In A. Niessen & J. Knigge (Hrsg.), *Theoretische Rahmung und Theoriebildung in der musikpädagogischen Forschung* (S. 163–176). Waxmann. <https://doi.org/10.25656/01:12620>.
- Hasselhorn, J., & Wolf, A. (2018). Assessment, Bewertung und Musikkritik. In A. C. Lehmann & R. Koppiez (Hrsg.), *Handbuch Musikpsychologie* (S. 389–409). Hogrefe.
- Hasselhorn, J., Platz, F., & Harnischmacher, C. (2022). Interindividuelle Leistungs differenzierung von Musiklehrkräften in Lerngruppen als Voraussetzung für Adaptivität im Musikunterricht. *Zeitschrift für Erziehungswissenschaft*, 25(3), 519–549. <https://doi.org/10.1007/s11618-022-01105-z>.
- Hesse, I., & Latzko, B. (2017). *Diagnostik für Lehrkräfte* (3. Aufl.). Barbara Budrich.
- Hofbauer, V. C., & Wehen, A. (2024). Das Schulmusikstudium auf dem Prüfstand. Eine Interviewstudie mit Musikhochschulabsolvent*innen. In B. Clausen, J. Ehninger & M. Sachsse (Hrsg.), *45. Jahresband des Arbeitskreises Musikpädagogische Forschung* (S. 165–177). Waxmann. <https://doi.org/10.31244/9783830999188.13>.
- Imthurn, G. (2020). Basic voice check—A diagnostic tool for the singing class. In T. S. Brophy & M. Haning (Hrsg.), *Advancing music education through assessment: honoring culture, diversity, and practice* (S. 121–136). GIA.
- Ingenkamp, K., & Lissmann, U. (2008). *Lehrbuch der pädagogischen Diagnostik* (6. Aufl.). Beltz.
- Jank, W. (2021). *Musikdidaktik. Praxishandbuch* (9. Aufl.). Cornelsen.
- Juslin, P. N. (2023). Musical performance evaluation: ten insights from psychological studies of aesthetic judgment. *Journal of Comparative Literature and Aesthetics*, 46(1), 93–106.
- Kahneman, D. (2011). *Schnelles Denken, langsames Denken*. Siedler.
- Kahneman, D., Sibony, O., Sunstein, C. R., & Schmidt, T. (2021). *Noise: Was unsere Entscheidungen verzerrt – und wie wir sie verbessern können*. Siedler.
- Kaiser, H. J. (2010). Verständige Musikpraxis. Eine Antwort auf Legitimationsdefizite des Klassenmusizieren. *Zeitschrift für Kritische Musikpädagogik*, 10, 47–68.
- Klieme, E., & Warwas, J. (2011). Konzepte der Individuellen Förderung. *Zeitschrift für Pädagogik*, 57(6), 805–818. <https://doi.org/10.25656/01:8782>.
- Knigge, J., Niessen, A., & Jordan, A.-K. (2010). Erfassung der Kompetenz „Musik wahrnehmen und kontextualisieren“ mit Hilfe von Testaufgaben. Aufgabenentwicklung und -analyse im Projekt KoMus. In N. Knolle (Hrsg.), *Evaluationsforschung in der Musikpädagogik* (S. 81–107). Die Blaue Eule. <https://doi.org/10.25656/01:15773>.
- Kopiez, R. (2018). Dimensionen des Künstlerischen aus musikpsychologischer Perspektive. In M. Dartsch, J. Knigge, A. Niessen, F. Platz & C. Stöger (Hrsg.), *Handbuch Musikpädagogik. Grundlagen – Forschung – Diskurse* (S. 187–192). Waxmann.
- Lehmann, A. C. (2011). Expressive variants in the opening of Robert Schumann's Arlequin (from Carnival, op. 59). 54 pianists' interpretations of a metrical ambiguity. In I. Deliege & J. W. Davidson (Hrsg.), *Music and the Mind: Essays in Honour of John Sloboda* (S. 311–324). Oxford University Press.
- Leiner, D. J. (2022). *SoSci Survey (Version 3.3.10)*. Software
- Letzring, T. D., Wells, S. M., & Funder, D. C. (2006). Information quantity and quality affect the realistic accuracy of personality judgment. *Journal of Personality and Social Psychology*, 91(1), 111–123. <https://doi.org/10.1037/0022-3514.91.1.111>.
- Lill, F. S., Hasselhorn, J., & Lehmann, A. C. (2020). Examining heterogeneity in practical music competencies in the music classroom—How does heterogeneity in 5th to 10th grade students compare to a 9th grade sample? In T. S. Brophy & M. Haning (Hrsg.), *Advancing music education through assessment: Honoring culture, diversity, and practice* (S. 137–154). GIA.
- McPherson, G. E., & Schubert, E. (2022). Enhancing music performance appraisal. In G. E. McPherson (Hrsg.), *The Oxford handbook of music performance* (Bd. 2, S. 103–134). Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780190058869.013.7>.
- Morijiri, Y., & Welch, G. F. (2022). Decisions on the quality of piano performance: evaluation of self and others. *Frontiers in Psychology*, 13, 954261. <https://doi.org/10.3389/fpsyg.2022.954261>.
- Nowicki, L., Prinz, W., Grosjean, M., Repp, B. H., & Keller, P. E. (2013). Mutual adaptive timing in interpersonal action coordination. *Psychomusicology: Music, Mind, and Brain*, 23(1), 6–20. <https://doi.org/10.1037/a0032039>.

- Passarotto, E., Altenmüller, E., & Müllensiefen, D. (2023). Music performance assessment: noise in judgments and reliability of measurements. *Psychology of Aesthetics, Creativity, and the Arts*. <https://doi.org/10.1037/aca0000574>.
- Platz, F. (2018). Formen der Leistungserfassung. In M. Dartsch, J. Knigge, A. Niessen, F. Platz & C. Stöger (Hrsg.), *Handbuch Musikpädagogik* (S. 377–384). Waxmann.
- Platz, F., & Kopiez, R. (2022). Stage behavior, impression management, and charisma. In G. E. McPherson (Hrsg.), *The Oxford handbook of music performance 2* (Bd. 2, S. 84–102). Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780190058869.013.6>.
- Rasch, B., Frieze, M., Hofmann, W., & Naumann, E. (2014). *Quantitative Methoden 1: Einführung in die Statistik für Psychologen und Sozialwissenschaftler*. Springer. <https://doi.org/10.1007/978-3-662-43524-3>.
- Schrader, F.-W. (1989). *Diagnostische Kompetenzen von Lehrern und ihre Bedeutung für die Gestaltung von Effektivität des Unterrichts*. P. Lang.
- Schrader, F.-W. (2013). Diagnostische Kompetenz von Lehrpersonen. *Beiträge zur Lehrerbildung*, 31(2), 154–165. <https://doi.org/10.25656/01:13843>.
- Schrader, F.-W., & Helmke, A. (1987). Diagnostische Kompetenz von Lehrern: Komponenten und Wirkungen. *Empirische Pädagogik. Zeitschrift zu Theorie und Praxis erziehungswissenschaftlicher. Forschung*, 1(1), 27–52.
- Schütze, B., Souvignier, E., & Hasselhorn, M. (2018). Stichwort – Formatives Assessment. *Zeitschrift für Erziehungswissenschaft*, 21(4), 697–715. <https://doi.org/10.1007/s11618-018-0838-7>.
- Sedlmeier, P., & Renkewitz, F. (2013). *Forschungsmethoden und Statistik für Psychologen und Sozialwissenschaftler* (2. Aufl.). Pearson.
- Südkamp, A., & Praetorius, A.-K. (Hrsg.). (2017). *Diagnostische Kompetenz von Lehrkräften: Theoretische und methodische Weiterentwicklungen*. Waxmann.
- Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: a meta-analysis. *Journal of Educational Psychology*, 104(3), 743–762. <https://doi.org/10.1037/a0027627>.
- Thompson, W.F., Diamond, C.T.P., & Balkwill, L.-L. (1998). The adjudication of six performances of a chopin etude: a study of expert knowledge. *Psychology of Music*, 26(2), 154–174. <https://doi.org/10.1177/0305735698262004>.
- Tversky, A., & Kahneman, D. (1974). *Judgment under uncertainty: heuristics and biases* (S. 1124–1131).
- Urhahne, D., & Wijnia, L. (2021). A review on the accuracy of teacher judgments. *Educational Research Review*, 32, 100374. <https://doi.org/10.1016/j.edurev.2020.100374>.
- Vogt, J., Heß, F., & Rolle, C. (2012). *Musikpädagogik und Heterogenität. Sitzungsbericht 2011 der Wissenschaftlichen Sozietät Musikpädagogik*. LIT.
- Wesolowski, B. C. (2017). Exploring rater cognition: a typology of raters in the context of music performance assessment. *Psychology of Music*, 45(3), 375–399. <https://doi.org/10.1177/0305735616665004>.
- Wesolowski, B. C., Wind, S. A., & Engelhard, G. (2017). Evaluating differential rater functioning over time in the context of solo music performance assessment. *Bulletin of the Council for Research in Music Education*, 212, 75–98. <https://doi.org/10.5406/bulcouresmusedu.212.0075>.
- Wolf, A., & Kopiez, R. (2018). Development and validation of the musical ear training assessment (META). *Journal of Research in Music Education*, 66(1), 53–70. <https://doi.org/10.1177/0022429418754845>.

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