Z Erziehungswiss (2022) 25:519–549
https://doi.org/10.1007/s11618-022-01105-z

ZfE

SCHWERPUNKT

Interindividuelle Leistungsdifferenzierung von Musiklehrkräften in Lerngruppen als Voraussetzung für Adaptivität im Musikunterricht

Johannes Hasselhorn · Friedrich Platz · Christian Harnischmacher

Eingegangen: 21. September 2021 / Überarbeitet: 7. März 2022 / Angenommen: 17. Mai 2022 / Online publiziert: 28. Juli 2022

Zusammenfassung Adaptiver Fachunterricht gehört zu den besten Möglichkeiten für die Individualisierung von schulischen Lernprozessen. Eine Voraussetzung für dessen Gelingen ist die angemessene Beurteilung der individuellen Leistungsstände der Schüler:innen durch die Lehrkraft und somit ihre pädagogisch-diagnostische Kompetenz. In dieser Studie werden am Beispiel des Klassengesangs zwei Fragen untersucht: zum einen, ob Musiklehrkräfte in der Lage sind, musikpraktischen Gruppenleistungen in unstandardisierten, jedoch für den Musikunterricht typischen Erhebungssituationen ausreichend Informationen für eine angemessene pädagogische Diagnostik zu entnehmen, zum anderen, ob diese Urteile ihre Unterrichtsplanung im Sinne adaptiven Unterrichts beeinflussen. Anhand eines Online-Experiments wurden die Niveau- und Differenzierungskomponente der Urteilsgenauigkeit von $N = 460$ Musiklehrkräften untersucht. Die Ergebnisse zeigen, dass Musiklehrkräfte das Leistungsniveau der musikpraktischen Darbietung einer unbekannten Lerngruppe (Niveaukomponente) anhand ihrer Gruppenmusizierleistung leistungsgerecht einschätzen können, die Leistungsheterogenität (Differenzierungskomponente) hingegen nicht. Dementsprechend fällt der Einfluss der Leistungsstreuung innerhalb der Lerngruppe auf die Unterrichtsplanung im Sinne von Adaptivität unzureichend gering aus.

Dr. Johannes Hasselhorn (✉)
Hochschule für Musik und Darstellende Kunst Frankfurt am Main, Eschersheimer Landstraße 29–39, 60322 Frankfurt am Main, Deutschland
E-Mail: johannes.hasselhorn@hfmdk-frankfurt.de

Prof. Dr. Friedrich Platz (✉)
Hochschule für Musik und Darstellende Kunst Stuttgart, Urbanstr. 25, 70182 Stuttgart, Deutschland
E-Mail: friedrich.platz@hmdk-stuttgart.de

Prof. Dr. Christian Harnischmacher
Universität der Künste Berlin, Lietzenburger Straße 45, 10789 Berlin, Deutschland
E-Mail: harnischmacher@udk-berlin.de

Schlüsselwörter Urteilsgenauigkeit · Diagnostische Kompetenzen · Gruppenleistungen · Adaptiver Musikunterricht

Interindividual performance differentiation of music teachers in learning groups as a prerequisite for adaptivity in music teaching

Abstract Adaptive teaching has long been considered as one of the most effective ways to support individual learning. A prerequisite for successful adaptive teaching is pedagogical diagnosis, which is often operationalized in terms of judgment accuracy with regard to both the mean level of the group performance (level component) and the extent of individual differences within the group (differentiation component). In this study, we use the example of class singing in music lessons to investigate whether music teachers can extract enough information for pedagogical diagnostics even in unstandardized group performance situations and whether these judgments influence their lesson planning in terms of adaptive teaching. Using data from an online experiment with $n = 460$ music teachers, we show that music teachers are able to assess the level component of a class music performance, but are unable to assess the differentiation component. Accordingly, the influence of students' performance heterogeneity within a class turns out to be rather low on teachers' lesson planning in terms of adaptivity.

Keywords Judgement accuracy · Diagnostic competence · Group performance · Adaptive music lessons

1 Einleitung

Die Leistungsbeurteilung von Schüler:innen ist eine zentrale diagnostische Aufgabe von Lehrkräften (Hesse und Latzko 2017), da sie insbesondere mit unterrichtsrelevanten Anforderungen verbunden ist, die an Lehrkräfte gestellt werden (Kunter und Trautwein 2013, S. 150). So gilt sie als unverzichtbare Voraussetzung für die Konzeption und Durchführung eines adaptiven Fachunterrichts, insbesondere für die inhaltliche Gestaltung von Unterricht, die Auswahl von geeigneten Lernaufgaben und die Planung von Hilfestellungen im Lernprozess (Hasselhorn und Gold 2017, S. 244).

Bisherige Forschung zur Leistungsbeurteilung stützt sich in der Regel auf Leistungsbeurteilungen in den Fächern Mathematik und Deutsch (vgl. Südkamp et al. 2012). Eine besondere Rolle wird dabei der Urteilsgenauigkeit zugeschrieben, die in *Niveaukomponente, Differenzierungskomponente* und *Rangkomponente* unterteilt wird (Schrader und Helmke 1987) und ein Maß der Übereinstimmung zwischen der lehrkräfteseitig antizipierten und in standardisierten Einzelerhebungssituationen empirisch beobachtbaren Individualleistung von Schüler:innen ist. Während sich diese auf die individuelle Beurteilung von Schülerleistungen ausgerichtete Verfahrensweise in den zuvor genannten Kernfächern bewährt hat, findet ein derartiges Vorgehen zur Beurteilung schülerseitiger Leistungen in der Musikpraxis, einem Hauptbereich des Schulfachs Musik, keine Anwendung. Gegen eine standardisierte

Erhebungssituation (u. U. vor dem Klassenverband) – wie es bspw. das Einzelvorsingen in Castingshows darstellt – spricht neben leistungsmindernden motivationalen Dispositionen und Ängsten („Lampenfieber“, vgl. Hasselhorn et al. 2012; Moeller und Castringius 2005) vor allem eine zentrale Handlungsmaxime unterrichtender Musiklehrkräfte: Künstlerischer Ausdruck als handlungsleitende Zielvorstellung gemeinsam musizierender Personen soll beim Musizieren, vor allem beim Singen, zu ästhetischen Erfahrungen beitragen (z. B. Zill 2016). Infolgedessen wird zum Erreichen dieses Ziels der Entwicklung der individuellen Gesangsqualität von Schüler:innen einer Klasse während des musikpraktischen Erarbeitungsvorgangs weniger Bedeutung beigemessen. Stattdessen steht die Qualitätsentwicklung des erarbeiteten Gesamtklangs der Lerngruppe im Zentrum der schulbezogenen Unterrichtstätigkeit. Darüber hinaus handelt es sich beim Musizieren allein und beim Musizieren mit anderen in Hinblick auf die musikalische Entwicklung um zwei unterschiedliche Entwicklungsbereiche (vgl. Busch 2010). Beim Musizieren mit anderen ist die eigene Leistung nicht unabhängig von den Leistungen der anderen, es gilt u. a. gemeinsam ein Tempo zu finden, Klangfarben zu gestalten und überhaupt eine gemeinsame musikalische Interpretation zu realisieren.

Trotz vieler gruppenabhängiger Fachleistungen sind Musiklehrkräfte auch aufgefordert, Unterricht nach den Prinzipien der individuellen Förderung zu planen, durchzuführen und zu evaluieren (Greuel 2007; Platz 2018). Die allgemeinpädagogische Forderung nach einer Stärkung der individuellen Förderung im schulischen Unterricht (Klieme und Warwas 2011) wurde in diesem Zusammenhang auch für den Musikunterricht als Zielvorstellung ausgemacht (Platz et al. 2021). Dabei kommt insbesondere dem adaptiven Unterrichten eine Schlüsselrolle zu (Greuel und Szczepaniak 2007). Die für adaptiven Musikunterricht auch im Teilbereich Singen notwendige pädagogische Diagnostik (vgl. Schrader 2013) können Musiklehrkräfte oft nur am Ergebnis ihrer Anleitung zum Gruppenmusizieren im Unterricht betreiben. Im Unterschied zu schriftlichen Leistungen oder leistungsrelevanten mündlichen Beiträgen im Unterrichtsgespräch ist eine Identifikation von Einzelleistungen, aus denen sich die Gruppendarbietungsleistung zusammensetzt, aufgrund der akustisch komplexen Natur des Gruppengesangs erheblich schwieriger zu bewältigen (Stadler-Elmer 2005). Schüler:innen singen nicht nur qualitativ unterschiedlich, sondern auch unterschiedlich laut. Dennoch sollen Musiklehrkräfte insbesondere im Sinne formativer Messung (vgl. Platz 2018) in der Lage sein, derart komplexen Gruppengesangsdarbietungen auch reichhaltige diagnostische Informationen über Individualleistungen entnehmen zu können, um zukünftigen Unterricht im Teilbereich Singen adaptiv gestalten und hierdurch eine individuelle Förderung der Schüler:innen gewährleisten zu können.

Dieser Beitrag beschäftigt sich daher mit der Frage, wie differenziert eine Leistungsdiagnostik von Musiklehrkräften unter derartigen Bedingungen ausfallen kann. Erst mit diesem Wissen kann eine Einschätzung gelingen, ob eine Grundvoraussetzung für adaptiven Musikunterricht im Bereich der Musikpraxis, einem fachrelevanten und zentralen Anforderungsbereich (z. B. Jank 2021; Ott 2018), ausreichend erfüllt ist.

2 Theoretischer und empirischer Hintergrund

Individuelle Förderung von Schüler:innen ist seit vielen Jahren ein bestimmender Grundsatz pädagogischen Arbeitens (Fischer et al. 2014), die in drei Varianten zum Ausdruck kommt: (1) kompensatorische Trainings- und Zusatzangebote, (2) vielfältige Lernwege durch offenen Unterricht sowie (3) Binnendifferenzierung durch adaptiven Unterricht (Klieme und Warwas 2011).

Hierbei basieren kompensatorische Trainings- und Zusatzangebote üblicherweise auf einer eher defizitorientierten Sichtweise, derzufolge Schüler:innen, die in einem spezifischen schulischen Leistungsbereich Minderleistungen zeigen, bspw. zusätzliche Übungsgruppen, Förder- oder Nachhilfeunterricht in Anspruch nehmen. Dabei können die Zusatzangebote fakultativ oder obligatorisch sein. Die bisherige Forschung zu diesem Bereich bezieht sich – sofern es um fachliche Leistungen geht – jedoch in aller Regel auf die schulischen Hauptfächer (Arnold 2008). Im Nebenfach Musik existieren derartige kompensatorische Angebote eher nicht: Trainings- und Zusatzangebote sprechen dagegen eher musikalisch leistungsstarke Schüler:innen an (Hasselhorn und McElvany 2016; Jordan 2014), reichen sie doch von Musik-AGs in der Schule (z. B. Schulchöre, -orchester oder -bands) bis hin zu außerschulischen musikalischen Aktivitäten wie dem Instrumentalunterricht (Dartsch und Heß 2018). Insbesondere leistungsschwächere Schüler:innen werden durch diese Variante individueller Förderung für das Fach Musik nicht erreicht. Im Gegenteil, mit dem spezifischen Blick auf das Musizieren verstärken solche Angebote die bereits ohnehin erhebliche Leistungsheterogenität innerhalb einer Lerngruppe (Hasselhorn und Lehmann 2015).

Offene Unterrichtskonzepte werden bereits seit mehreren Jahrzehnten in den schulischen Unterricht integriert. Sie umfassen einen breiten Methodenkatalog von Projekt- über Stationen- bis hin zu Wochen- oder Monatsplanarbeit. So kann vor allem die Projektarbeit im Musikunterricht auf eine lange Tradition verweisen (Legrand 2012), ebenso die Stationenarbeit im Bereich der Musikpraxis (Brunner 2020). Während offene Unterrichtskonzepte vor allem positive Auswirkungen auf die Selbstwahrnehmung der Schüler:innen zu haben scheinen (Hartinger 2005), scheinen ihre Auswirkungen auf die schulischen Leistungen hingehen nicht eindeutig positiv auszufallen. Vielmehr scheinen auch von offenen Unterrichtsformen vor allem leistungsstärkere Schüler:innen zu profitieren, während die Leistungsschwächeren mehr vorgegebene Struktur und Anleitung benötigen (Messner und Blum 2019). Der Erfolg offener Konzepte hinsichtlich der fachbezogenen Leistungen im Fach Musik ist bislang allerdings nicht empirisch untersucht, sodass vermutet werden darf, dass auch hier eher leistungsstärkere Schüler:innen in dem Sinne profitieren, dass bei selbstständiger Erarbeitung komplexerer Aufgaben, wie es bei Stationenarbeiten üblich ist, diejenigen mit einem strukturierteren, erfahrungsbasierten Überblick über den Unterrichtsinhalt zügiger und zielgerichteter zufriedenstellende Aufgabenlösungen erarbeiten können. Dass solche teilweise extremen Leistungsheterogenitäten auch im Schulfach Musik vorliegen, konnte bereits empirisch gezeigt werden (Lill et al. 2020).

Adaptiver Unterricht erreicht seine binnendifferenzierende Wirkung dadurch, dass der Unterricht sowohl in der Planungs- (Makro-Ebene) als auch Durchführungspha-

se (Mikro-Ebene) „an die lernrelevanten Unterschiede zwischen den Schülerinnen und Schülern“ (Weinert 1997, S. 51; vgl. auch Helmke 2017, S. 251) der Lerngruppe angepasst wird, indem Leistungsanforderungen ins Verhältnis zu den individuellen Voraussetzungen der Schüler:innen gesetzt werden. Dabei wird adaptiver Unterricht in der Regel als ein Ideal verstanden, dem sich Lehrkräfte so weit wie möglich annähern sollen. In letzter Konsequenz müsste jede Schülerin und jeder Schüler einer Lerngruppe eine individuelle Aufgabe mit Bezug zum gleichen Unterrichtsinhalt erhalten, die so ausgewählt wird, dass sie optimal an ihrer oder seiner aktuellen Leistungs- bzw. Kompetenzschwelle lernen können (Häcker 2017). Im täglichen Unterricht werden in Bezug auf die Makro-Ebene häufig Strategien angewendet, bei denen entweder verschiedene Aufgaben unterschiedlicher Schwierigkeit vorbereitet werden, oder Aufgabensequenzen mit aufeinander aufbauenden, im Schwierigkeitsgrad ansteigenden Teilaufgaben, die dann in individuellem Tempo bearbeitet werden können. Auf der Mikro-Ebene wird adaptiver Unterricht häufig mit der zielorientierten Anwendung von Methoden des Scaffolding oder der konstruktiven Unterstützung gleichgesetzt (vgl. Hardy et al. 2019). Dabei wird in anderen Ansätzen die lernunterstützende Aktivität durch Lehrkräfte in der Unterrichtssituation als Mikro-Scaffolding bezeichnet, wohingegen Zielanalyse, Lernstandsanalyse und die Unterrichtsplanung unter dem Begriff Makro-Scaffolding gefasst werden (vgl. Gibbons 2015). Die Bedeutung der Planungsaktivitäten für die konkreten adaptiven Aktivitäten im Unterricht konnte mindestens für den Mathematikunterricht bereits festgestellt werden (Prediger und Pöhler 2015). Empirische Forschung zu adaptiver Planungskompetenz von Lehrkräften analysiert diese in der Regel anhand schriftlicher Unterrichtsentwürfe. Insbesondere der Einbezug diagnostischer Informationen und differenzierender Betrachtungen kognitiver Unterschiede der Lernenden werden hier herangezogen (Rey et al. 2018). Im Zuge von Unterrichtsplanung setzen sich Lehrkräfte in der Regel vorwiegend mit den Unterrichtsinhalten und zugleich mit den ins Zentrum des geplanten Unterrichts gesetzten Aufgaben auseinander (König et al. 2015). Gerade in kompetenzorientiertem Unterricht haben die konkreten Aufgaben eine zentrale Funktion und sind daher auch zentraler Punkt in der Planung (vgl. Müller et al. 2013).

Auch für den Musikunterricht spielen Strategien des adaptiven Unterrichts eine wichtige Rolle (Harnischmacher et al. 2021). Auf adaptiven Musikunterricht ausgerichtete musikpädagogische Forschung befasst sich dabei vorwiegend mit der Beschreibung von unterrichtlichen Handlungen, die der Mikro-Ebene zuzuordnen sind (z. B. Göllner und Niessen 2016; Kranefeld und Heberle 2016; Kranefeld et al. 2015). Die Makro-Ebene, als das Anlegen adaptiven Musikunterrichts bereits in der Planungsphase, wird dagegen in der musikpädagogischen Literatur erst in den letzten Jahren verstärkt thematisiert (vgl. Harnischmacher 2012; Jank 2018; Jank und Meyer 2017).

Grundvoraussetzung für die beschriebene Form von adaptivem (Musik)Unterricht ist es, dass die unterrichtenden Lehrkräfte die lernrelevanten Unterschiede zwischen den Schüler:innen kennen, die bei der Planung und Durchführung des Unterrichts berücksichtigt werden sollen. Andernfalls ist eine entsprechende Anpassung ausgeschlossen. Es muss also sowohl vor der Unterrichtsplanung als auch während des Unterrichts durch die Lehrkräfte dauerhaft Diagnostik betrieben werden (vgl.

Hasselhorn und Gold 2017, S. 228). Unter pädagogischer Diagnostik lassen sich Tätigkeiten von Lehrkräften zusammenfassen, die zu differenzierenden Urteilen von Schüler:innen verschiedener lernrelevanter Merkmale führen können. Die Fähigkeit, diagnostische Aufgaben im Schulalltag unter Anwendung dieser Tätigkeiten erfolgreich bewältigen zu können, wird als diagnostische Kompetenz von Lehrkräften verstanden (Schrader 2013). Mit Blick auf die Planung adaptiven Unterrichts ist zudem die Eingrenzung der diagnostischen Kompetenz auf die Fähigkeit sinnvoll, Informationen für unterrichtsrelevante Entscheidungen zu sammeln (Praetorius et al. 2017). Dabei ist zu berücksichtigen, dass diagnostische Kompetenz in der empirischen Forschung häufig mit Urteilsgenauigkeit gleichgesetzt wird (Karst und Förster 2017; Praetorius und Südkamp 2017). Aus guten Gründen wird jüngst jedoch gefordert, die Urteilsgenauigkeit als Subdomäne diagnostischer Kompetenz aufzufassen (vgl. Urhane und Wijnia 2021). Die Genauigkeit, mit der Lehrkräfte die fachbezogenen Leistungen ihrer Schüler:innen einschätzen können, wird in der Meta-Analyse von Südkamp et al. (2012) lediglich mit $r = 0{,}63$ quantifiziert. Da sich dieser Schätzer auf Ergebnisse bezieht, die unter optimierten und standardisierten Bedingungen gefunden wurden, ist anzunehmen, dass die Urteilsgenauigkeit von Lehrkräften in komplexeren, unstandardisierten Unterrichtssituationen, wie sie im Musikunterricht der Regelfall sind, noch geringer ausfällt. Urhane und Wijnia (2021) stellen darüber hinaus fest, dass Lehrkräfte die Leistungen ihrer Schüler:innen tendenziell überschätzen. Die Urteilsgenauigkeit scheint insbesondere von der Berufserfahrung der Lehrkräfte und der Schulform abhängig zu sein. Für beide Moderatoren sind die empirischen Befunde allerdings uneinheitlich.

Aus dem Review von Urhane und Wijnia (2021) geht auch hervor, dass sich Erkenntnisse zur Urteilsgenauigkeit von Lehrkräften nahezu ausschließlich auf Ergebnisse aus den Fächern Mathematik und Sprachen beziehen. Inwieweit diese Ergebnisse auf andere Fächer übertragbar ist, ist unklar. Auch die Frage, ob es einen positiven Einfluss auf die Urteilsgenauigkeit hat, wenn Lehrkräfte die zu beurteilenden Schüler:innen besser kennen – z. B. im Verlaufe eines Schuljahres – ist demnach bislang nicht empirisch befriedigend beantwortet.

Urteilsgenauigkeit wird überwiegend so operationalisiert, dass Schüler:innen fachbezogene Testaufgaben bearbeiten, und die Lehrkräfte anhand des Aufgabenmaterials die individuellen Leistungen der Schüler:innen a priori einschätzen. Aus den Angaben werden die *Niveaukomponente* (Einschätzung der mittleren Leistung der Lerngruppe), die *Differenzierungskomponente* (Einschätzung der Leistungsstreuung innerhalb der Lerngruppe) und die *Rangkomponente* (Leistungsbezogene Reihenfolge der Schüler:innen der Lerngruppe) berechnet (vgl. Schrader und Helmke 1987). Dieses Vorgehen hat den Vorteil, dass im Anschluss an die Leistungsanforderungs- und -fähigkeitseinschätzung durch die Lehrkräfte die von den Schüler:innen individuell und unabhängig voneinander erbrachte Leistungen als Messlatte für die Qualität der Einschätzungen herangezogen werden können. Dies entspricht auch einem möglichen Vorgehen im realen Unterricht. Ohne diese individuellen, auch im Nachhinein prüfbaren Leistungen kann die Angemessenheit und Qualität der Leistungseinschätzung durch die Lehrkräfte nicht überprüft werden.

Auf den Bereich der Musikpraxis trifft dies im Alltag leider nicht zu. Die Leistungen der Schüler:innen können nicht so einfach für spätere Überprüfungen fest-

gehalten werden wie z. B. durch Aufschreiben wie in Deutsch oder Mathematik. Es wäre zwar denkbar, die Schüler:innen im Rahmen eines Tests individuelle Aufnahmen anfertigen zu lassen, die später beurteilt werden könnten (z. B. im Rahmen von Kompetenztests, vgl. Hasselhorn 2015), aber dieses Vorgehen entspricht nicht den gegenwärtigen musikpädagogischen Zielen, die das gemeinsame Musizieren, das aufeinander hören, reagieren und abstimmen in den Vordergrund stellen. In diesem Zusammenhang blickt das Fach Musik auf eine lange Tradition zurück, in der das gemeinsame Musizieren und das solistische Musizieren auch bezüglich der dahinterstehenden Fähigkeiten und Fertigkeiten als unterschiedlich bekannt sind (vgl. Hasselhorn 2015; Venus 1969). Dieses gemeinsame Musizieren kann in standardisierten Einzeltestungen bislang nicht abgebildet werden, bzw. es ist immer abhängig von der konkreten Lerngruppe, was wiederum dem Anspruch der Objektivität durch Standardisierung widerspricht. Dies ist auch ein Problem für Musiklehrkräfte, denn sie sollen einerseits die musikpraktischen Leistungsstände der Schüler:innen diagnostizieren, haben aber als Überprüfungsmöglichkeit nur ihre eigene Erinnerung an die bereits vergangenen Leistungen der Gruppe.

Die Güte derartiger Beurteilungen über künstlerische Darbietungen hängt im hohen Maße vom Zusammenspiel verschiedener Einflussgrößen ab, die zum Teil auch außerhalb des Verantwortungsraums der interagierenden Personen liegt (im Überblick s. Hasselhorn und Wolf 2018; Platz und Kopiez 2022) und nicht als bloßes Ergebnis von auditiven Bewertungsprozessen subsumiert werden können. So zeigen die Ergebnisse der Meta-Analyse von Platz und Kopiez (2012), dass visuelle Informationen im Vergleich zu auditiven Informationen eine Beurteilungsverschiebung im Umfang einer halben Standardabweichung zur Folge haben können ($d = 0{,}51$, 95 % CI [0,42; 0,59]). Hinzu kommt, dass im Musikunterricht nahezu ausschließlich in der Gruppe musiziert wird. Individuelle Leistungen aus einer leistungsheterogenen Gruppe von Schüler:innen beim Singen in der Klasse treffend einschätzen zu können, stellt eine erhebliche Herausforderung für Musiklehrkräfte dar, wenn diese ihre Lerngruppe zum Singen von Musikstücken anleiten soll.

Dabei kommt erschwerend hinzu, dass die gezeigten Leistungen der Schüler:innen nicht nur von der eigenen Kompetenz abhängen, sondern beim gemeinsamen Musizieren auch immer von den konkreten Leistungen der anderen abhängig sind und dass bereits wenige (Gruppen-)Leistungen für eine diagnostische Einschätzung durch die Lehrkräfte ausreichen müssen, da der Musikunterricht nicht nur aus Musizieren besteht und auch beim Musizieren noch in verschiedene Kompetenzbereiche zu unterteilen ist (Hasselhorn 2015).

3 Fragestellungen

Vor dem Hintergrund des skizzierten Forschungsstands geht dieser Beitrag der Frage nach, ob Musiklehrkräfte in der Lage sind, aus gesanglichen Gruppenleistungen hinreichend genau diagnostisch relevante Informationen herauszufiltern, um ihren Musikunterricht angemessen adaptiv planen zu können. Im Einzelnen sollen folgende Fragen beantwortet werden:

1. Können Musiklehrkräfte das mittlere Leistungsniveau (*Niveaukomponente*) einer Lerngruppe und ihre Leistungsbreite (*Differenzierungskomponente*) anhand gesanglicher Gruppendarbietungsleistungen differenzieren?
2. Haben die tatsächlichen Leistungen der Schüler:innen über die Einschätzung der beiden Komponenten der Urteilsgenauigkeit der Musiklehrkräfte einen Einfluss auf ihre adaptive Unterrichtsplanung?

4 Methode

Zur Klärung dieser Fragen wurde eine Onlinestudie mit experimentellen Bedingungsvariationen des gesanglichen Leistungsniveaus (hoch vs. mittel vs. niedrig) und der Leistungsbreite (niedrig vs. hoch) in der Lerngruppe realisiert.

4.1 Stichprobe

An der Onlinestudie nahmen insgesamt 528 Personen teil. Ihre Rekrutierung erfolgte über Fachverbände, persönliche Kontaktaufnahme mit örtlichen Seminaren sowie der Weitergabe der Einladung unter den Fachkräften selbst („Schneeballsystem“). Für die Datenanalyse wurden Personen ausgeschlossen, die angaben, aktuell nicht als Lehrkraft zu arbeiten ($n=5$), Musik nicht zu unterrichten ($n=6$), als Förderschullehrkraft tätig zu sein ($n=14$), oder aber kein Lehramtsstudium erfolgreich abgeschlossen zu haben ($n=22$). Als weiteres Ausschlusskriterium galt die Verletzung der Mindestverweildauer von 10 s für das Lesen der Instruktionsseite ($n=6$) sowie für das Hören des jeweiligen Hörbeispiels ($n=4$).

Daraus resultierte eine Analysestichprobe von insgesamt $N=471$ Personen ($n=329$ weiblich, $n=140$ männlich, $n=2$ ohne Nennung) im Altersbereich von 23 bis 64 Jahren ($M=43{,}43$; $SD=9{,}77$; $Mdn=44$). Alle Personen unserer Stichprobe

Tab. 1 Zusammenfassung der Stichprobenmerkmale

					Häufigkeit	
	Min	*Max*	*M*	*SD*	Absolut	Relativ
Alter	23	64	43,43	9,77	–	–
Geschlecht						
Weiblich	–	–	–	–	329	69,9
Männlich	–	–	–	–	140	29,7
Keine Angabe	–	–	–	–	2	0,4
Studium und Berufsabschluss						
Musikstudium	–	–	–	–	396	84,1
Referendariat	–	–	–	–	458	97,2
Schulform						
Gymnasium	–	–	–	–	192	40,8
Grundschule	–	–	–	–	164	34,8
Andere Schulform	–	–	–	–	115	24,4

Der Gesamtumfang der Stichprobe betrug $N=471$

verfügten über ein erstes Staatsexamen oder einen Master of Education als Studienabschluss und $n = 458$ (97,2 %) von ihnen zusätzlich über ein abgeschossenes Referendariat als Berufsabschluss. Die Mehrheit der Stichprobe ($n = 456$) gab an, ein Musikstudium absolviert zu haben (vgl. Tab. 1), wobei von ihnen 384 Personen (81,5 % der Gesamtstichprobe) angaben, sowohl einen musikbezogenen Studienabschluss als auch ein abgeschlossenes Referendariat vorweisen zu können. In 40,8 % aller Fälle unserer Stichprobe ($n = 192$) lag eine Berufsausübung im Gymnasium sowie in 34,8 % im Grundschulbereich vor ($n = 164$). Alle übrigen Fälle verteilten sich auf sonstige, teilweise bundesländerspezifische weiterführende Schulformen.

4.2 Fallvignetten mit experimenteller Bedingungsvariation

Die an der Online-Studie teilnehmenden Personen sollten für standardisierte gesangliche Gruppenleistungen die Niveaukomponente und die Differenzierungskomponente der Leistungen in der Gruppe einschätzen. Dazu wurden sechs Fallvignetten konstruiert, die die 3 (Leistungsniveaus) × 2 (Leistungsbreiten) experimentellen Bedingungsvariationen repräsentierten. Die Fallvignetten bestanden aus einer narrativen Text- und einer auditiven Beurteilungskomponente. Die narrative Textkomponente diente zur Induktion einer fiktionalen Szene, in der ein Fachkollege um Rat für seine Unterrichtplanung bittet und in die sich die teilnehmenden Personen hineinversetzen sollten:

> Ein junger Kollege kommt auf Sie zu und bittet Sie um Rat bei der Vorbereitung seines Musikunterrichts in der Klasse 9E. Sein Ziel sei es – so Ihr Kollege –, dass die Klasse in naher Zukunft einfache, mehrstimmige Lieder im Musikunterricht singen könne. Um dieses Ziel zu erreichen, habe er einen Kanon aus dem Musikbuch ausgewählt, zu dem verschiedene Aufgaben und Herangehensweisen im Begleitband für Lehrerinnen und Lehrer vorgeschlagen werden. Sein Problem sei, dass er sich unsicher sei, welche Aufgaben er auswählen und wie viel Zeit er für diese einplanen soll. Bisher hat die Klasse 9E mit 15 Schülerinnen und 11 Schülern im Musikunterricht überwiegend einstimmige Lieder gesungen, so wie das Lied „Guten Abend, gute Nacht", das die Klasse am Ende der vergangenen Stunde mit Keyboardunterstützung und Playback aufgenommen hat. Diese Aufnahme hat Ihr Kollege mitgebracht, damit Sie ihm bei der Planung der nächsten Stunde anhand des aktuellen Leistungsstandes der Klasse beraten können.

Alle sechs Fallvignetten hatten dieselbe narrative Textkomponente. Sie unterschieden sich jedoch in ihren nachfolgenden Klassengesangsdarbietungen, so dass erst in der auditiven Beurteilungskomponente der Fallvignetten die experimentellen Bedingungsvariationen aus der Kombination von drei unterschiedlichen Klassendurchschnittsleistungen im Gesang als Operationalisierung für die Niveaukomponente (UV_1: hohes vs. mittleres vs. niedriges Leistungsniveau) und zwei Ausprägungsformen in der Leistungsbreite der Gesangsdarbietungsleistungen als Operationalisierung für die Differenzierungskomponente (UV_2: homogene vs. heterogene Leistungsverteilung) realisiert wurden.

Für die konkrete Operationalisierung dienten Gesangsdarbietungen von Schüler:innen der 9. Jahrgangsstufe, die als Einzelaufnahmen vorlagen und durch Expertenurteile annotiert waren (Hasselhorn 2015, S. 122). Aus dem Korpus von insgesamt 6230 Einzelgesangsdarbietungen von neun unterschiedlichen Stücken, die von 445 Schüler:innen im Zuge der Testentwicklung eingesungen worden waren, wurden Einzeldarbietungen verwendet, deren zugrunde liegenden Vokalstücke (= Items) nach Expertenurteil auf einer sechsstufigen Skala entsprechend der Schulnotenskala hohe Intraklassenkorrelationen ($ICC(2, n) \geq 0{,}92$; Hasselhorn 2015, S. 93 ff.) und ein Abdecken des kompletten Leistungsspektrums in Gesang abbildeten (Bereich der Schwellenparameter: $-4{,}30 \leq \tau \leq 2{,}74$). 890 Gesangsdarbietungen des Wiegenlieds von Johannes Brahms (op. 49, Nr. 4) erfüllten diese Vorgaben. Aus ihnen wurden schließlich zufällig 13 Einzeldarbietungen je Schulnotenstufe, d. h. $13 \times 6 = 78$ individuelle Gesangsdarbietungen ausgewählt.

Die Leistungszusammensetzung wurde für jede Experimentalbedingung so bestimmt, dass zum einen drei Leistungsniveaustufen ($UV_{1.1}$: $1{,}50 \leq M_{HLN} \leq 2{,}58$ [hohes Leistungsniveau]; $UV_{1.2}$: $M_{MLN} = 3{,}50$ [mittleres Leistungsniveau]; $UV_{1.3}$: $4{,}42 \leq M_{NLN} \leq 5{,}50$ [niedriges Leistungsniveau]) als Ausprägungsformen der Niveaukomponente (UV 1) sowie zwei Streuungsbreiten der zusammengeführten, individuellen Darbietungsleistungen ($UV_{2.1}$: $SD_{Ho} = 0{,}51$ [homogene normalverteilte Leistungsverteilung]; $UV_{2.2}$: $1{,}60 \leq SD_{HE} \leq 1{,}68$ [heterogene normalverteilte Leistungsverteilung]) als Differenzierungskomponente (UV 2) realisiert werden konnten (vgl. Tab. 2). Die auditiven Komponenten der Fallvignetten mit Operationalisierungen homogener Leistungsverteilungen zeichneten sich weiterhin dadurch aus, dass sie ausschließlich aus individuellen Darbietungsleistungen zusammengesetzt waren, deren zugrundeliegenden Expertenurteile und -annotationen (Hasselhorn 2015) innerhalb einer Leistungsspanne zweier benachbarter Schulnoten lagen ($\Delta = 1 = \text{const.}$) – im Unterschied zu denen mit heterogenen Leistungsverteilungen ($\Delta > 1$).

Unter Berücksichtigung der auf diese Weise zuvor ermittelten bedingungsspezifischen Leistungszusammensetzungen sowie eines festen Geschlechterverhältnisses

Tab. 2 Übersicht über Ausprägungsformen der unabhängigen Variablen innerhalb der Experimentalbedingungen

	Experimentalbedingungen						Leistungszusammensetzung[a]					
n	Leistungsniveau	Leistungsstreuung	*Min*	*Max*	*M*	*SD*	1	2	3	4	5	6
76	Hoch	Homogen	1	2	1,50	0,51	$^{+}$13(8)	13(8)	–	–	–	–
78		Heterogen	1	6	2,58	1,60	8(5)	8(5)	3(2)	3(2)	2(1)	2(1)
81	Mittel	Homogen	3	4	3,50	0,51	–	–	13(8)	13(8)	–	–
80		Heterogen	1	6	3,50	1,68	4(3)	4(3)	5(3)	5(3)	4(2)	4(2)
76	Niedrig	Homogen	5	6	5,50	0,51	–	–	–	–	13(8)	13(8)
80		Heterogen	1	6	4,42	1,60	2(1)	2(1)	3(2)	3(2)	8(5)	8(5)

n = Gruppengrößen innerhalb der Experimentalbedingungen; $^{+} = n$ (n_w), d. h. Umfang ausgewählter Aufnahmen unter zusätzlicher Angabe der absoluten Häufigkeit von Aufnahmen von Schülerinnen

[a] Die Leistungszusammensetzung erfolgte auf Grundlage der in Hasselhorn (2015) dargestellten Expertenbeurteilungen auf einer sechsstufigen Skala, die der [Schul-]Notenskala entspricht

von 16 Schülerinnen und 10 Schülern wurden für jede Versuchsbedingung 26 Einzelgesangsdarbietungen zufällig bestimmt, zusammengeführt und mit der Software Audacity (V. 2.1.2) abgemischt.

4.3 Niveau- und Differenzierungskomponente der diagnostischen Kompetenz von Musiklehrkräften

Die Erhebung der diagnostischen Kompetenz der teilnehmenden Musiklehrkräfte wurde auf die Operationalisierung von Niveau- und Differenzierungskomponente begrenzt. Während die Niveaukomponente ihrer diagnostischen Kompetenz durch ihre Bewertung des durchschnittlichen Leistungsniveaus der Klasse mit Hilfe des sechsstufigen Notensystems für schulische Leistungen operationalisiert wurde („*Wie bewerten Sie das durchschnittliche Leistungsniveau der Klasse 9E im Singen?*“ Bewertungsraum: sehr gut (Note 1) bis ungenügend (Note 6)), spiegelte sich die Differenzierungskomponente in ihrem Urteil zur Leistungsverteilung der Klasse wider („*Die Leistungsverteilung der Klasse 9E ist …*“), das ebenfalls auf einer sechsstufigen Antwortskala erhoben wurde („deutlich homogen“ bis „deutlich heterogen“).

Diese methodische Diskrepanz zur bislang etablierten Erhebungsmethodik in diesem Bereich (s. oben) war aufgrund der fachspezifischen Besonderheit der zu beurteilenden Leistung notwendig. Obwohl Gruppengesang aus den konkreten Einzelleistungen der Schüler:innen einer Lerngruppe besteht, sind diese Einzelleistungen nicht unabhängig von der Gruppenleistung beobachtbar. Ein Schüler wird nicht genauso wie in der Gruppe singen, wenn er eine Stimme allein singen muss. Diese Operationalisierung von Niveau- und Differenzierungskomponente ist daher für diesen fachspezifischen Leistungs- bzw. Kompetenzbereich ökologisch valide.

4.4 Unterrichtsplanungsverhalten

Die Grundlage zur Operationalisierung des Unterrichtsplanungsverhaltens stellte ein Repertoire von vier vorgeschlagenen Herangehensweisen unterschiedlichen Anforderungsniveaus (vgl. Tab. 3) zur Erarbeitung des vierstimmigen Kanons „Im Nebel“ (vgl. Abb. 1) dar. Die Studienteilnehmenden wurden aufgefordert, diejenigen Herangehensweisen auszuwählen, die sie zur Erarbeitung des Kanons für die in der Fallvignette zuvor vorgestellte fiktive Klasse für empfehlenswert hielten. Dabei war die Anzahl der ausgewählten Herangehensweisen den Proband:innen freigestellt, es sollten insgesamt 30 min auf die ausgewählten Herangehensweisen verteilt werden (vgl. Abb. 2).

Eine Einschätzung durch die teilnehmenden Lehrkräfte ergab, dass die kompositorische Anlage des Kanons ein mittleres bis hohes musikpraktisches Leistungsanforderungsniveau für die neunte Klassenstufe darstellte und infolgedessen von einer ausreichenden Messdifferenzierung durch das Notenmaterial ausgegangen werden konnte. So schätzten die Lehrkräfte unserer Stichprobe, dass durchschnittlich (a) 28,48 % ($SD = 21{,}02$) aller Schüler:innen der neunten Jahrgangsstufe die Melodiestimme des Kanons intonationssicher allein (entsprechend der Niveaubeschreibung 3 nach Hasselhorn 2015, S. 193), (b) 29,70 % ($SD = 21{,}08$) die Melodiestimme des Kanons intonationssicher im Kanongesang und (c) 48,63 % ($SD = 28{,}07$)

Tab. 3 Überblick über die vorgeschlagenen Herangehensweisen zur Erarbeitung des Kanons und ihre schwierigkeitsbestimmenden Merkmale

	Herangehensweise	Anforderungsniveau	Schwierigkeitsbestimmende Merkmale	
A	Die Schülerinnen und Schüler lernen die Begleitstimme und singen diese im Klassenverband einstimmig	Elementar	*Mono-motivisches Begleit-Ostinato*	
			Ambitus:	Reine Quarte
			Maximalintervall:	Reine Quarte
			Notenwerte:	Halbe Note, Viertelnote (und -pause)
			Umfang:	Ein Takt, jedoch zweimal wiederholt und einmal variiert (mit retrograder Tonfolge, sog. „Krebs“)
			Darbietung:	Einstimmig
B	Die Schülerinnen und Schüler lernen die ersten beiden Kanoneinsätze (System 1 und 2) und singen diesen Teil im Klassenverband einstimmig	Grundlegend	*Kanonphrase*	
			Ambitus:	Reine Oktave
			Maximalintervall:	Reine Quinte
			Notenwerte:	Halbe Note, Viertelnote, Achtelnote
			Umfang:	8 Takte
			Darbietung:	Einstimmig
C	Die Schülerinnen und Schüler werden in zwei Gruppen geteilt. Die erste Gruppe singt die ersten zwei Systeme des Kanons, die zweite Gruppe singt dazu die Begleitstimme. Anschließend wird getauscht	Gehoben	*Kanonphrase mit Begleit-Ostinato*	
			Ambitus:	Reine Oktave
			Maximalintervall:	Reine Quinte
			Notenwerte:	Halbe Note, Viertelnote (und -pause), Achtelnote
			Umfang:	8 Takte
			Darbietung:	Zweistimmig
D	Die Schülerinnen und Schüler lernen den ganzen Kanon gemeinsam. Anschließend werden sie in vier Gruppen geteilt und der Kanon wird vierstimmig gesungen	Hoch	*Vollständiger Kanon*	
			Ambitus:	Reine Undezime
			Maximalintervall:	Große Sexte
			Notenwerte:	Halbe Note, Viertelnote, Achtelnote
			Umfang:	16 Takte
			Darbietung:	Vierstimmig

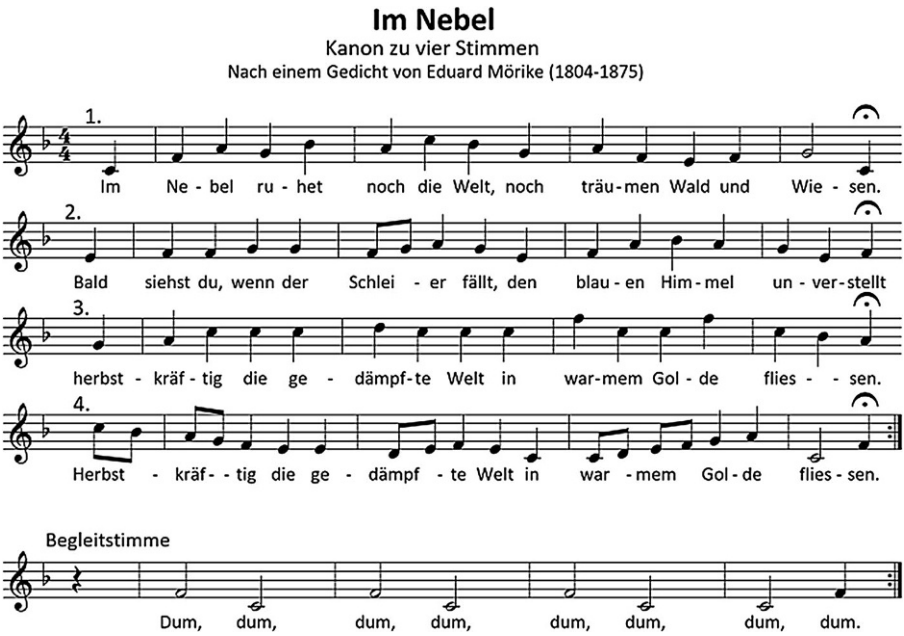

Abb. 1 Vierstimmiger Kanon „Im Nebel" mit Begleit-Ostinato

Welche Aufgaben sollte Ihr Kollege für die nächste Stunde in der Klasse 9E Ihrer Empfehlung nach mit wie vielen Minuten einplanen?
Sie können 30 Minuten verteilen. Wenn Sie eine Aufgabe nicht empfehlen wollen, lassen Sie das Feld einfach frei oder tragen Sie „0" ein.

Aufgabe A: ____Minuten
Aufgabe B: ____Minuten
Aufgabe C: ____Minuten
Aufgabe D: ____Minuten

Bisher verplant: ____Minuten

Abb. 2 Fragebogen-Item zur Erhebung von Aufgabenselektion und zugehörigem Zeitprofil als Operationalisierung der zweiten Zielkriteriumsvariable Unterrichtsplanung

die Begleitstimme zum Kanon intonationssicher allein hätten singen können. Diese Einschätzungen wurden auf Grundlage der Berufserfahrung der teilnehmenden Musiklehrkräfte für Schüler:innen der Klassenstufe 9 im Allgemeinen vorgenommen.

Die Anforderungscharakteristika der vier vorgeschlagenen Herangehensweisen orientierten sich an den Kompetenzniveaubeschreibungen für die Dimension Gesang der Musikpraktischen Kompetenzen (Hasselhorn 2015). Zentrale, schwierigkeitsbestimmende Merkmale für Gesang wie Ambitus („Tonumfang"), Maximalintervall, Darbietung und weitere Parameter wurden im Notentext (vgl. Abb. 1) sowie in den vorgeschlagenen Herangehensweisen so aufeinander abgestimmt, dass eine eindeuti-

ge Rangzuordnung der Herangehensweisen zu ihrem Anforderungsniveaus resultierte (vgl. Shandro 2015; Watkins und Farnum 1954): Die Herangehensweisen reichten vom Einstudieren eines als einfach einzustufenden monomotivischen Begleit-Ostinatos, dessen Anforderungen noch unterhalb des von Hasselhorn (2015) beschriebenen Ausgangsniveaus lagen (Herangehensweise A), bis hin zur Zielsetzung, den gesamten Kanon einzustudieren und vierstimmig darzubieten (Herangehensweise D), was eine Kompetenzleistung der dritten Niveaustufe nach Hasselhorn (2015) erfordert.

Zur Operationalisierung des Unterrichtsplanungsverhaltens wurden daher zwei Variablen herangezogen: *Anzahl Vorgehensweisen* als Summe der eingeplanten Aufgaben (von 1 bis 4) sowie *Anforderungsniveau* als das Niveau der schwersten eingeplanten Aufgabe (von 1 „hohes Niveau“ bis 4 „elementares Niveau“).

4.5 Ablauf

Nach einer schriftlichen Begrüßung erfolgte eine technische Überprüfung der Abspielmöglichkeiten auf den Endgeräten der teilnehmenden Personen. Im Falle einer technischen Fehlfunktion und in der Folge dem Ausbleiben der anwählbaren Musikdarbietung wurden weiterführende Informationen zur Problembehandlung und -lösung zur Verfügung gestellt. Es schloss sich die Erhebung von personen- und berufsbezogenen Hintergrundinformationen an, gefolgt von der Vorstellung der für jede Person zufällig ausgewählten Fallvignette, deren narrative und auditive Komponente auf je eine Bildschirmseite aufgeteilt wurde. Auf eine zeitliche Standardisierung der Anzeigedauer der narrativen Komponente der Fallvignetten wurde verzichtet, um Unterschieden in der Lesegeschwindigkeit gerecht zu werden. In der Folge konnten die Personen individuell entscheiden, wann die Darbietung der jeweiligen auditiven Komponente beginnen sollte. Die Teilnehmenden konnten auch selbst entscheiden, wie häufig diese abgespielt wurde.

Es folgte die Instruktionsphase: Gleichzeitig mit der Präsentation des Notenbildes zum vierstimmigen Kanon „Im Nebel“ wurden die Teilnehmenden aufgefordert, aus den vier vorgeschlagenen Herangehensweisen zur singenden Erarbeitung des Kanons dem fiktiven Kollegen diejenigen vorzuschlagen, die ihrer Ansicht nach für die fiktive Lerngruppe für eine Erarbeitungszeit von maximal 30 min geeignet sind und anzugeben, wieviel Erarbeitungszeit für jede dieser vorgeschlagenen Herangehensweisen eingeplant werden sollte. Zum Abschluss wurden sowohl die Niveau- („Durchschnittliches Leistungsniveau“) als auch die Differenzierungskomponente („Leistungsverteilung“) des diagnostischen Urteils der teilnehmenden Musiklehrkräfte erfasst, wofür ihnen abermals die Abspielmöglichkeit des auditiven Teils der Fallvignette zur Verfügung stand. Zum Abschluss wurde den teilnehmenden Personen nochmals das Notenbild des Kanons gezeigt und ihre Einschätzung erbeten, wie hoch ihres Erachtens der Prozentsatz aller Schüler:innen einer 9. Klasse ist, die (1) die Melodiestimme des Kanons intonationssicher allein singen, (2) die Melodiestimme des Kanons intonationssicher im Kanon mitsingen und (3) die Begleitstimme des Kanons intonationssicher alleine singen können.

4.6 Auswertung

Zur Untersuchung der ersten Fragestellung erfolgte die Datenaufbereitung und -auswertung mit dem Programm SPSS (V. 27.0.0.0), wohingegen die Pfadanalysen zur Auseinandersetzung mit der zweiten Fragestellung mit dem Programm Mplus (V. 8.3) durchgeführt wurden.

Die deskriptive Analyse (vgl. Tab. 4) ergab, dass keine der für das statistische Modell der geplanten Auswertungen relevanten Variablen einen absoluten Schiefe-Wert

Tab. 4 Deskriptive Statistik der in die Modellanalyse aufgenommenen Variablen

Erhobene Variablen	*M*	*SD*	*Min*	*Max*	*Schiefe*	*Kurtosis*
1. Niveaukomponente	3,43	0,99	1	6	0,31	–0,57
2. Differenzierungskomponente	4,35	1,14	1	6	–0,20	–0,63
3. Anforderungsniveau	1,72	0,63	1	4	0,33	–0,39
4. Anzahl Vorgehensweisen	3,06	0,71	1	4	–0,51	0,32
5. Geschlecht[a]	1,71	0,47	1	3	–0,78	–1,04
6. Alter	43,43	9,77	23	64	0,04	–0,96
7. Gymnasiallehrkraft[b]	0,41	0,49	0	1	0,38	–1,87
8. Grundschullehrkraft[b]	0,35	0,48	0	1	0,64	–1,60
9. Leistungsniveau	2,00	0,81	1	3	–0,01	–1,48
10. Leistungsbreite	1,51	0,50	1	2	–0,02	–2,01

Der Gesamtstichprobenumfang vor Analyse auf univariate und multivariate Ausreißer betrug $N = 471$. Der Standardfehler der Schiefe betrug $SE_{\text{Schiefe}} = 0{,}11$, der Standardfehler der Kurtosis: $SE_{\text{Kurtosis}} = 0{,}23$

[a] Das Geschlecht war codiert in 1 (männlich), 2 (weiblich) und 3 (keine Angabe)

[b] Dichotome Dummycodierung mit der Referenzgruppe der Lehrkräfte an nicht-gymnasialen Sekundarschulen

Tab. 5 Deskriptive Statistik der in die MANOVA oder Modellanalyse aufgenommenen Variablen nach Ausschluss aller Personen mit auffälligem Antwortverhalten („outlier")

Erhobene Variablen	*M*	*SD*	*Min*	*Max*	*Schiefe*	*Kurtosis*
1. Niveaukomponente	3,41	0,97	1	6	0,31	–0,53
2. Differenzierungskomponente	4,36	1,12	2	6	–0,13	–0,78
3. Anforderungsniveau	1,72	0,61	1	3	0,24	–0,61
4. Anzahl Vorgehensweisen	3,10	0,66	1	4	–0,25	–0,19
5. Geschlecht*[a]	1,72	0,46	1	3	–0,89	–1,02
6. Alter*	43,50	9,77	23	64	0,04	–0,96
7. Gymnasiallehrkraft[b]	0,41	0,49	0	1	0,37	–1,87
8. Grundschullehrkraft[b]	0,35	0,48	0	1	0,61	–1,63
9. Leistungsniveau	2,00	0,81	1	3	0,00	–1,48
10. Leistungsbreite	1,51	0,50	1	2	–0,03	–2,00

Die mit Sternchen (*) markierten Variablen wurden nicht als Variablen in die Pfadanalysen aufgenommen (vgl. Abschnitt 3.3). Der Standardfehler der Schiefe betrug $SE_{\text{Schiefe}} = 0{,}11$, der Standardfehler der Kurtosis: $SE_{\text{Kurtosis}} = 0{,}23$. Unter Ausschluss univariater und multivariater Ausreißer beträgt der Stichprobenumfang: $N = 460$.

[a] Das Geschlecht war codiert in 1 (männlich), 2 (weiblich) und 3 (keine Angabe).

[b] Dichotome Dummycodierung mit der Referenzgruppe der Lehrkräfte an nicht-gymnasialen Sekundarschulen

Tab. 6 Korrelationen zwischen den erhobenen Variablen

Erhobene Variablen	1	2	3	4	5	6	7	8	9
1. Niveaukomponente	–	–	–	–	–	–	–	–	–
2. Differenzierungskomponente	**0,63**	–	–	–	–	–	–	–	–
3. Anforderungsniveau	**0,25**	**0,22**	–	–	–	–	–	–	–
4. Anzahl Vorgehensweisen	**–0,17**	**–0,18**	**–0,70**	–	–	–	–	–	–
5. Geschlecht	–0,02	0,03	**0,16**	–0,05	–	–	–	–	–
6. Alter	**0,12**	–0,01	–0,08	0,08	**–0,16**	–	–	–	–
7. Gymnasiallehrkraft	0,01	–0,04	**–0,10**	**0,22**	**–0,29**	0,01	–	–	–
8. Grundschullehrkraft	0,05	0,09	**0,11**	**–0,18**	**0,35**	0,06	**–0,62**	–	–
9. Leistungsniveau	**0,60**	**0,48**	**0,20**	**–0,18**	0,02	0,06	–0,05	0,06	–
10. Leistungsbreite	–0,04	0,07	–0,03	0,04	–0,08	0,02	0,01	0,02	0,03

Stichprobenumfang nach Ausschluss von Personen mit auffälligem Antwortverhalten („outlier“, $n = 11$) beträgt nach Bereinigung: $N = 460$
Mit Fettdruck gekennzeichnete Korrelationen fallen auf dem 1 %-Niveau (2-seitig) signifikant aus

außerhalb des Bereiches [–2;2] sowie eine Kurtosis außerhalb des Wertebereichs [–2;2] aufwies. Somit lagen keine Hinweise auf Verletzungen der univariaten Normalitätsannahme gemäß den Empfehlungen von Kim (2013, S. 53) für Stichproben mit einem Umfang $N > 300$ vor. Mit Hilfe einer Analyse auf multivariate Ausreißer (mittels Mahalanobis-Distanz) ließen sich elf Fälle (= 2,3 %) mit extremem Antwortverhalten auf den für die zehn Modellvariablen verwendeten Items identifizieren. Diese wurden von der weiteren Datenauswertung ausgeschlossen (vgl. Weston und Gore 2006), sodass für die nachfolgenden Datenauswertungen ein Datensatz mit vollständigem Antwortverhalten auf den zehn erhobenen Variablen von $n = 460$ Personen zur Verfügung stand (vgl. Tab. 5). Nach der „10 Personen-pro-Parameter“-Daumenregel (vgl. Kline 2015, S. 14 ff.) ließ dieser resultierende Stichprobenumfang eine akzeptable Schätzung eines Umfangs von ca. 46 Modellparametern zu (s. unten).

Eine Korrelationsanalyse (vgl. Tab. 6) ergab keine Hinweise auf Multikollinearität zwischen den Variablen (Kriterium: $r > 0{,}85$, vgl. Weston und Gore 2006, S. 735). Als potenzielle Kovariaten wiesen lediglich zwei Personenvariablen (Tätigkeit als Gymnasiallehrkraft und Tätigkeit als Grundschullehrkraft [dummycodiert mit der Referenzgruppe der Lehrkräfte an nicht-gymnasialen Sekundarstufen]) einen bedeutsamen statistischen Zusammenhang mit mehr als einer der abhängigen Variablen auf ($|r| > .10$), weshalb die anderen potenziellen Kovariaten (Geschlecht und Alter) von der weiteren Modellbildung ausgeschlossen wurden[1].

[1] Die Datensätze und Auswertungssytaxen der im Ergebnisteil beschriebenen Auswertungen sind als ergänzende Materialien in der Online-Fassung dieses Artikels abrufbar.

5 Ergebnisse

Wie präzise die Einschätzung von Niveau- und Differenzierungskomponente bei Gruppengesangsleistungen mit unterschiedlichen durchschnittlichen Leistungsniveaus und Leistungsbreiten durch Musiklehrkräfte erfolgt, zeigen Abb. 3 und 4 (vgl. auch Tab. 7). Eine zweifaktorielle multivariate Varianzanalyse mit den Bedingungsvariablen *Leistungsniveau* (3 Stufen) und *Leistungsbreite* (2 Stufen) sowie den beiden kontinuierlichen Zielvariablen *Niveau-* und *Differenzierungskomponente* (vgl. Tab. 8) ergab einen signifikanten Haupteffekt für das Leistungsniveau auf beiden Zielvariablen (Niveaukomponente: $F(2) = 141{,}88$; $p < 0{,}01$; $\eta^2 = 0{,}39$; Differenzierungskomponente: $F(2) = 72{,}37$; $p < 0{,}01$; $\eta^2 = 0{,}24$), nicht jedoch für die Leistungsbreite (Niveaukomponente: $F(1) = 2{,}70$; $p = 0{,}10$; $\eta^2 = 0{,}01$; Differenzierungskomponente: $F(1) = 2{,}02$; $p = 0{,}16$; $\eta^2 < 0{,}01$). Zusätzlich resultierten Interaktionseffekte aus den beiden UVn *Leistungsniveau* und *Leistungsbreite* für beide AVn (Niveaukomponente: $F(2) = 15{,}539$; $p < 0{,}01$; $\eta^2 = 0{,}06$; Differenzierungskomponente: $F(2) = 9{,}95$; $p < 0{,}01$; $\eta^2 = 0{,}04$). Mit dem Modell können 40,8 % der Varianz der Niveaukomponente und 26,1 % der Varianz der Differenzierungskomponente erklärt werden.

Die deskriptiv gefundenen Mittelwerte (vgl. Tab. 7 und Abb. 3 und 4) zeigen, dass die Einschätzung der mittleren Leistung der Lerngruppe (Niveaukomponente) insbesondere für die homogenen Gruppen besser zu funktionieren scheint als für die heterogenen Gruppen. Die Niveaueinschätzungen für diese Subgruppe folgen erwartungsgemäß nahezu einem linearen Zusammenhang. Gleichzeitig ist in den absoluten Werten im Vergleich zwischen den Niveaueinschätzungen und den berechneten Mittelwerten der Einzelleistungen eine eindeutige Tendenz zur Mitte in den Bewertungen zu beobachten: Die Einschätzung der hochleistenden homogenen Gruppe ist etwa eine Skaleneinheit niedriger als die berechnete Durchschnittsleistung, die der niedrigleistenden homogenen Gruppe etwa eine Skaleneinheit besser. Demgegenüber scheint die heterogenere Leistungsverteilung die Einschätzung des mittleren Leistungsniveaus zu erschweren. Die Werte für diese Subgruppen liegen näher beieinander, die Tendenz zur Mitte ist hier stärker ausgeprägt. Außerdem zeigt

Tab. 7 Mittelwerte und Standardabweichungen der Niveau- und Differenzierungskomponente in Abhängigkeit von Leistungsniveau und -breite der sechs fiktiven Lerngruppen aus den Fallvignetten

Leistungsniveau (UV_1)	Leistungsbreite (UV_2)					
	Homogen			Heterogen		
	M	*SD*	*n*	*M*	*SD*	*n*
	Niveaukomponente (AV_1)					
Hoch	2,57	0,72	76	2,97	0,72	75
Mittel	3,39	0,72	80	3,18	0,73	79
Niedrig	4,48	0,83	71	3,94	0,79	79
	Differenzierungskomponente (AV_2)					
Hoch	3,41	0,87	76	4,05	0,88	75
Mittel	4,45	0,95	80	4,13	0,80	79
Niedrig	5,03	1,21	71	5,09	0,94	79

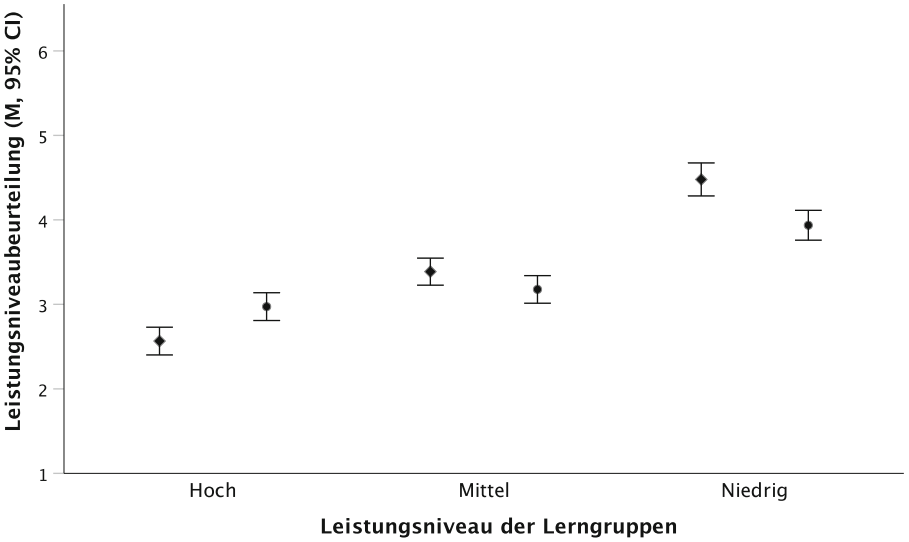

Abb. 3 Leistungsniveaubeurteilung von Lerngruppen (entsprechend der Schulnotengebung) mit unterschiedlichen Leistungsniveaustufen und -breiten (*Rhombus* = homogene Leistungsverteilung, *Kreis* = heterogene Leistungsverteilung)

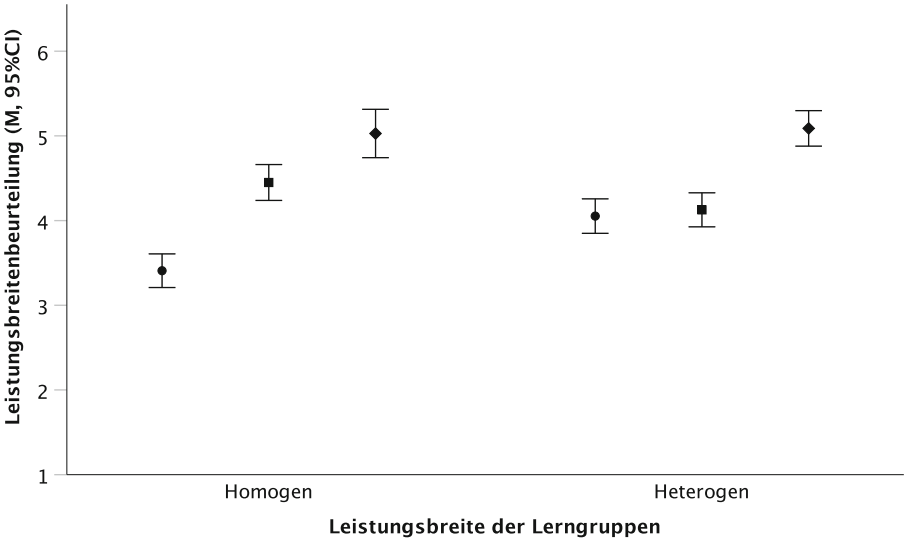

Abb. 4 Leistungsbreitebeurteilung von Lerngruppen (1 = sehr homogen; 6 = sehr heterogen) mit unterschiedlichen Leistungsbreiten und -niveaustufen (*Kreis* = hohes Niveau, *Quadrat* = mittleres Niveau, *Rhombus* = niedriges Niveau)

sich hier keine lineare Veränderung in der Bewertung, die Einschätzungen für die hoch- und die mittelleistenden heterogenen Gruppen liegen näher beieinander als die Einschätzung der niedrigleistenden heterogenen Gruppe.

Tab. 8 Ergebnisse der multivariaten Varianzanalyse mit den AVn *Niveau-* und *Differenzierungskomponente* sowie mit den UVn *Leistungsniveau* und *Leistungsbreite*

		df	*F*	*p*	η^2
Niveaukomponente	Leistungsniveau (LN)	2	141,88	<0,01	0,39
	Leistungsbreite (LB)	1	2,70	0,10	–
	LN * LB	2	15,54	<0,01	0,06
Differenzierungs-komponente	Leistungsniveau (LN)	2	72,37	<0,01	0,24
	Leistungsbreite (LB)	1	2,02	0,16	–
	LN * LB	2	9,95	<0,01	0,04

Die deskriptiven Befunde zur Einschätzung der Leistungsverteilung der Lerngruppen (Differenzierungskomponente) fallen erwartungswidrig aus. Die Verteilung ist dem Verteilungsmuster für die Niveaukomponente erstaunlich ähnlich. Insbesondere für die homogenen Subgruppen scheint zu gelten, dass eine bessere Durchschnittsleistung der Lerngruppe von den Lehrkräften mit dem Prädikat „homogener" in Verbindung gebracht werden. Eine echte Differenzierung im Sinne des in der Differenzierungskomponente anvisierten Konstrukts der Leistungsbreite bzw. -streuung der Lerngruppe gelingt offenbar lediglich für die hochleistenden Subgruppen. Für die anderen leistungsbezogenen Subgruppen unterscheiden sich die Einschätzungen der Differenzierungskomponente kaum. Insbesondere die niedrigleistenden Subgruppen werden als besonders heterogen eingeschätzt.

Um die zweite Fragestellung nach den Zusammenhängen zwischen Gesangsleistungen der Lerngruppe, Urteilsgenauigkeit und Unterrichtsplanung zu untersuchen, wurde eine Auswertungsstrategie im regressionsanalytischen Ansatz verfolgt. Das Pfaddiagramm in Abb. 5 zeigt die Annahmen des für die Beantwortung dieser Fragestellung theoretisch hergeleiteten Vorhersagemodells. In diesem Modell wird angenommen, dass die unabhängigen Variablen (Leistungsniveau und -breite) ausschließlich einen direkten Effekt auf die jeweils korrespondierenden Komponenten der Pädagogischen Diagnostik zeigen sollten, ohne jedoch die andere Komponente zu beeinflussen. Demnach sollte sich die beobachtbare Varianz der Niveaukomponente (NK) ausschließlich durch die Varianz des Leistungsniveaus (LN), nicht jedoch durch die Varianz der Leistungsbreite (LB) als zweite UV erklären lassen ($H_{1a} : (\beta_{LN\rightarrow NK} > 0) \wedge (\beta_{LB\rightarrow NK} = 0)$). In Analogie hierzu steht die Varianzaufklärung der Differenzierungskomponente (DK), die ausschließlich auf die Varianz der experimentell manipulierten Leistungsbreite, nicht jedoch auf die des Leistungsniveaus zurückgeführt werden sollte ($H_{1b} : (\beta_{LB\rightarrow DK} > 0) \wedge (\beta_{LN\rightarrow DK} = 0)$). Weiterhin wurde angenommen, dass keine Kovarianz zwischen Niveau- und Differenzierungskomponente besteht, da nach Leuders, Dörfler, Leuders und Philipp (2018) beide als distinkte Dimensionen der Diagnostik ohne wechselseitigen Bezug zueinander aufzufassen sind ($H_{1c} : \beta_{NK\leftrightarrow DK} = 0$). Obwohl beide Variablen deskriptiv korrelieren, gehen wir zunächst von einem theoriebasierten Modell aus, um prüfen zu können, ob diese bivariate Korrelation auf andere im Modell enthaltene Zusammenhänge zurückzuführen ist. Von beiden Variablen wird jedoch ein direkter Einfluss auf die Variablen der adaptiven Unterrichtsplanung *Anzahl Vorgehensweisen* (UP1) und *Anforderungsniveau* (UP2) angenommen

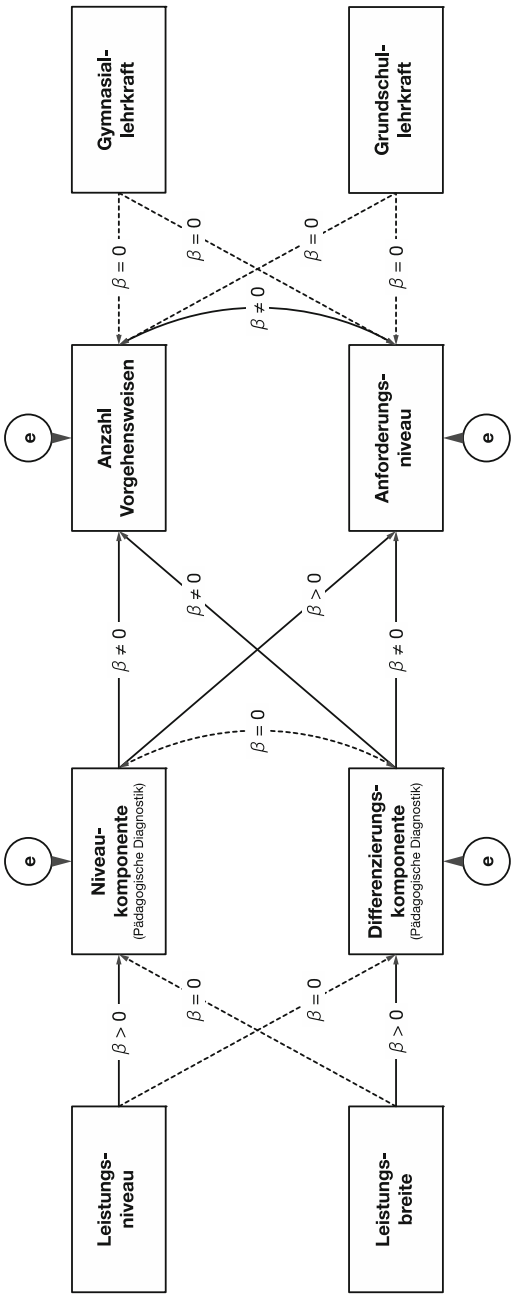

Abb. 5 Pfaddiagramm mit erwarteten Beziehungen zwischen Leistungsniveau und -breite von Lerngruppen sowie Niveau- und Differenzierungskomponente von Musiklehrkräften sowie die erwarteten Beziehungen der diagnostischen Komponenten zur Anzahl und zum Anforderungsniveau der Vorgehensweisen als Merkmale adaptiver Unterrichtsplanung von Musiklehrkräften. Erwartungswerte beziehen sich auf standardisierte Pfadkoeffizienten (β). Die *gestrichelten Verbindungen* verdeutlichen angenommene Beziehungen mit nicht-signifikanter Einflussstärke ($\beta = 0$; $p > 0{,}05$)

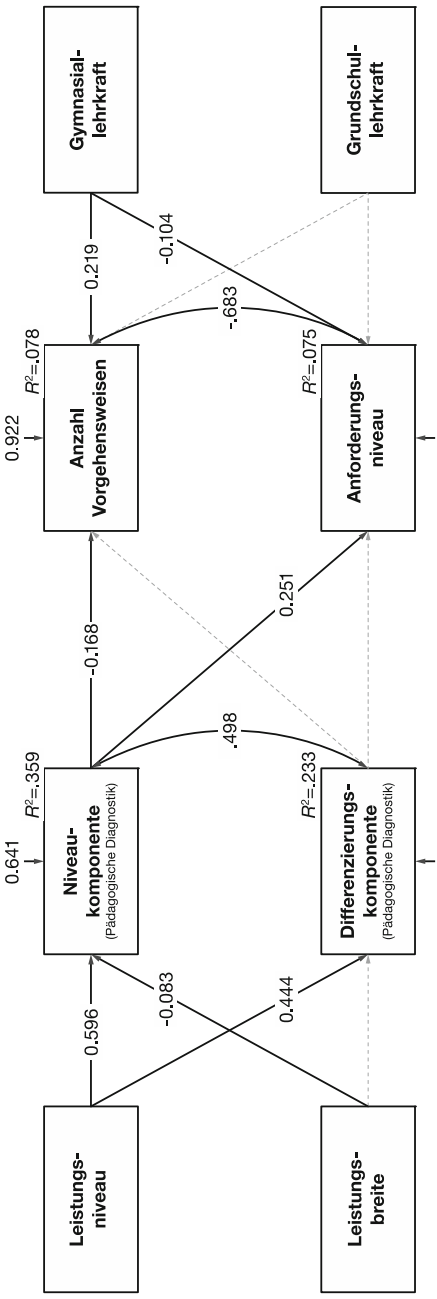

Abb. 6 Ergebnis der Pfadanalyse ($\chi^2 = 15{,}924$; $df = 13$, $p = 0{,}253$; $CFI = 0{,}996$; $TLI = 0{,}994$; $RMSEA = 0{,}02$; 90% CI (0,000; 0,054); $SRMR = 0{,}031$). Die Werte geben die Einflussstärke standardisierter Modellkoeffizienten wieder. Nicht signifikante, auf null restringierte Pfade sind in *grauer Farbe* dargestellt

(H_{1d} : ($\beta_{NK\rightarrow UP1} \neq 0$) $\wedge$ ($\beta_{DK\rightarrow UP1} \neq 0$), H_{1e} : ($\beta_{NK\rightarrow UP2} > 0$) $\wedge$ ($\beta_{DK\rightarrow UP2} \neq 0$)). Zugleich fungieren Niveau- und Differenzierungskomponente als Mediatorvariablen, über die sich Leistungsniveau und -breite indirekt auf die Unterrichtsplanung auswirken (H_{1f} : ($\beta_{LN\rightarrow NK\rightarrow UP1} \neq 0$) $\wedge$ ($\beta_{LB\rightarrow DK\rightarrow UP1} \neq 0$), H_{1g} : ($\beta_{LN\rightarrow NK\rightarrow UP2} \neq 0$) $\wedge$ ($\beta_{LB\rightarrow DK\rightarrow UP2} \neq 0$)). Die Frage, ob sich das Unterrichtsplanungsverhalten von Lehrkräften verschiedener Schulformen voneinander unterscheidet, so wie es die bivariaten Korrelationen (vgl. Tab. 6) vermuten lassen, sollte in diesem Untersuchungsansatz als exploratives Moment der Pfadanalyse zur Kontrolle berücksichtigt werden, ohne dass signifikante Einflüsse angenommen wurden (H_{1h} : ($\beta_{\mathrm{GYL}\rightarrow UP1} = 0$)$\wedge$($\beta_{\mathrm{GYL}\rightarrow UP2} = 0$)$\wedge$($\beta_{\mathrm{GSL}\rightarrow UP1} = 0$)$\wedge$($\beta_{\mathrm{GSL}\rightarrow UP2} = 0$)).

Die Schätzung der Modellparameter des Pfadmodells erfolgte mit Hilfe des in der Auswertungssoftware Mplus (V. 8.3) implementieren ML-Schätzers und Biaskorrigiertem Bootstrapping. In einem iterativen Verfahren erfolgte die Modellgüteoptimierung in drei Schritten (vgl. Abb. 5).

Mit seinen theoretisch abgeleiteten Parameterrestriktionen zeigte das Ausgangsmodell eine unzureichende Anpassung an die Daten (Modell 1: $\chi^2 = 263{,}958$; $df = 11$; $p < 0{,}001$; $CFI = 0{,}688$; $TLI = 0{,}375$; $RMSEA = 0{,}224$; 90% CI (0,201; 0,247); $SRMR = 0{,}153$). Deshalb wurde in einem weiteren Schritt das Ausgangsmodell ohne fixierte Parametervorgaben geschätzt. Dieses zweite Modell mit freier Parameterschätzung zeigte gegenüber dem Ausgangsmodell bereits eine gute Anpassung an die Daten (Modell 2: $\chi^2 = 9{,}170$; $df = 8$, $p = 0{,}328$; $CFI = 0{,}999$; $TLI = 0{,}996$; $RMSEA = 0{,}018$; 90% CI (0,000; 0,059); $SRMR = 0{,}022$). Im nun folgenden dritten Schritt wurden iterativ die nicht signifikanten Pfadschätzungen nacheinander auf 0 fixiert. Dies wurde für insgesamt fünf Pfade durchgeführt, sodass sich ein finales Modell mit guter Passung zu den Daten ohne frei geschätzte nicht signifikante Pfade ergab (finales Modell: $\chi^2 = 15{,}924$; $df = 13 p = 0{,}253$; $CFI = 0{,}996$; $TLI = 0{,}994$; $RMSEA = 0{,}022$; 90% CI (0,000; 0,054); $SRMR = 0{,}026$). Die Parameterschätzungen dieses Modells sind in Abb. 6 als Ergebnisse der Modellevaluation und -optimierung dargestellt.

Direkte Effekte Erwartungsgemäß zeigten sich signifikante Einflüsse der Niveaukomponente auf die beiden Unterrichtsplanungsvariablen *Anzahl Vorgehensweisen* ($b = -0{,}114$; $\beta = -0{,}168$; $z = -3{,}637$; $p < 0{,}001$) und *Anforderungsniveau* ($b = 0{,}159$; $\beta = 0{,}251$; $z = 5{,}481$; $p < 0{,}001$). Die direkten Pfade von der Differenzierungskomponente auf die Unterrichtsplanungsvariablen erreichten entgegen der Erwartungen keine Signifikanz. Insgesamt konnten im finalen Modell 7,8 % Varianz der Variablen *Anzahl Vorgehensweisen* und 7,5 % der Variablen *Anforderungsniveau* erklärt werden (vgl. Tab. 9). Interessanterweise zeigten Musiklehrkräfte am Gymnasium gegenüber Lehrkräften anderer Schulformen ein signifikant abweichendes Unterrichtsplanungsverhalten in beiden Variablen (*Anzahl Vorgehensweisen:* $b = 0{,}293$; $\beta = 0{,}219$; $z = 5{,}265$; $p < 0{,}001$; *Anforderungsniveau:* $b = -0{,}129$; $\beta = -0{,}104$; $z = -2{,}420$; $p = 0{,}016$) in der Form, dass Gymnasiallehrkräfte mehr Vorgehensweisen als ihre Kolleginnen und Kollegen an anderen Schulformen für die zur Verfügung stehenden 30 min auswählten und dabei gleichzeitig ein höheres Anforderungsniveau anlegten. Grundschullehrkräfte unterschieden

Tab. 9 Ergebnisse der Pfadanalyse (finales Modell)

	Unstandardisierte Pfadkoeffizienten		Standardisierte Pfadkoeffizienten						
					95 % *CI*				
Modell	*b*	*SE*	*β*	*SE*	*LL*	*UL*	*p*	R^2	*SE*
Direkte Effekte									
Niveaukomponente								0,359	0,035
(Intercept)	2,225	0,133	2,293	0,170	2,015	2,577	<0,001	–	–
Leistungsniveau	0,715	0,046	0,596	0,029	0,545	0,642	<0.001	–	–
Leistungsbreite	–0,161	0,064	–0,083	0,033	–0,135	–0,029	0,011	–	–
Differenzierungskomponente								0,233	0,037
(Intercept)	2,027	0,120	2,711	0,155	2,457	2,967	<0,001	–	–
Leistungsniveau	0,482	0,039	0,444	0,036	0,415	0,541	<0,001	–	–
Leistungsbreite	0,000	–	0,000	–	–	–	–	–	–
Anforderungsniveau								0,075	0,025
(Intercept)	1,236	0,108	2,019	0,198	1,700	2,346	<0,001	–	–
Gymnasiallehrkraft	–0,129	0,054	–0,104	0,043	–0,174	–0,033	0,016	–	–
Grundschullehrkraft	0,000	–	0,000	–	–	–	–	–	–
Niveaukomponente	0,159	0,031	0,251	0,046	0,175	0,325	<0,001	–	–
Differenzierungskomponente	0,000	–	0,000	–	–	–	–	–	–
Anzahl Vorgehensweisen								0,078	0,024
(Intercept)	3,374	0,116	5,123	0,223	4,759	5,487	<0,001	–	–
Gymnasiallehrkraft	0,293	0,058	0,219	0,042	0,150	0,286	<0,001	–	–
Grundschullehrkraft	0,000	–	0,000	–	–	–	–	–	–
Niveaukomponente	–0,114	0,032	–0,168	0,046	–0,242	–0,089	<0,001	–	–
Differenzierungskomponente	0,000	–	0,000	–	–	–	–	–	–
Niveau- ~ Differenzierungskomponente	0,379	0,039	0,498	0,042	0,423	0,561	<0,001	–	–
Anforderungsniveau ~ Anzahl Vorgehensweisen	–0,254	0,022	–0,683	0,044	–0,752	–0,609	<0,001	–	–

Tab. 9 (Fortsetzung)

Modell	Unstandardisierte Pfadkoeffizienten		Standardisierte Pfadkoeffizienten						
					95 % *CI*				
	b	*SE*	*β*	*SE*	*LL*	*UL*	*p*	R^2	*SE*
Indirekte Effekte									
Höchstes eingeplantes Aufgabenniveau: Leistungsniveau über Niveaukomponente	0,113	0,023	0,150	0,029	0,103	0,198	<0,001	–	–
Höchstes eingeplantes Aufgabenniveau: Leistungsbreite über Niveaukomponente	–0,025	0,011	–0,021	0,009	–0,038	–0,005	0,023	–	–
Anzahl Aufgaben/Zielsetzungen: Leistungsniveau über Niveaukomponente	–0,082	0,023	–0,100	0,028	–0,146	–0,053	<0,001	–	–
Anzahl Aufgaben/Zielsetzungen: Leistungsbreite über Niveaukomponente	0,018	0,009	0,014	0,007	0,005	0,028	0,044	–	–

Die Schätzung der Pfadanalyse erfolgte mit Hilfe des ML-Schätzers unter Anwendung von Bias-korrigierten Bootstrapping zur Schätzung der Konfidenzintervalle in Mplus (V. 8.3). Das Pfadmodell zeigt eine sehr gute Passung zu den Daten: χ^2–Test of Model Fit = 15,924; $df = 13$; $p = 0{,}253$; Comparative Fit Index $(CFI) = 0{,}996$; Tucker-Lewis Index $(TLI) = 0{,}994$, Root Mean Square Error of Approximation $(RMSEA) = 0{,}022$(90% *CI* [0,000; 0,054]); Standardized Root Mean Square Error of Residual index $(SRMR) = 0{,}031$. Die Tilde (~) beschreibt die Kovariate (bzw. Korrelation) zwischen Niveau- und Differenzierungskomponente (vgl. Abb. 6)

sich in ihrem Verhalten hingegen nicht von den nicht-gymnasialen Sekundarstufenlehrkräften.

Entgegen der Erwartung korrelierten Niveau- und Differenzierungskomponente hoch miteinander ($r = 0{,}498; 95\%\,CI\,[0{,}423; 0{,}561]; p < 0{,}001$). Leistungsstärkere Lerngruppen wurden auch als leistungshomogener eingeschätzt. Erwartungsgemäß zeigte sich ein hoher Zusammenhang zwischen den beiden Unterrichtsplanungsvariablen ($r = -0{,}683; 95\%\,CI\,[0{,}752; 0{,}609]; p < 0{,}001$). Ein höheres angelegtes Anforderungsniveau ging mit einer größeren Anzahl unterschiedlicher eingeplanter Vorgehensweisen einher.

Während 35,9 % der Varianz der Niveaukomponente durch Einflüsse des Leistungsniveaus ($b = 0{,}715; \beta = 0{,}596; z = 20{,}276; p < 0{,}001$) und in erheblich kleinerem Ausmaß der Leistungsbreite ($b = -0{,}161; \beta = -0{,}083; z = -2{,}537; p = 0{,}011$) aufgeklärt werden konnten, zeigte sich wider Erwarten ausschließlich das Leistungsniveau als signifikanter Prädiktor der Differenzierungskomponente ($b = 0{,}666; \beta = 0{,}482; z = 12{,}495; p < 0{,}001$), mit dem sich 23,3 % ihrer Varianz aufklären ließen.

Indirekte Effekte Einen Aufschluss über die Frage, welche Auswirkungen Leistungsniveau und -breite indirekt über Niveau- und Differenzierungskomponente als Mediatorvariablen auf die Unterrichtsplanungsvariablen haben, zeigte eine Analyse der indirekten Effekte (vgl. Tab. 9). Dabei erwiesen sich alle möglichen indirekten Effekte zwischen Leistungsniveau und -breite auf der einen und den Unterrichtsplanungsvariablen auf der anderen Seite als statistisch signifikant. Bemerkenswerterweise bestanden alle indirekten Effekte ausschließlich auf dem Pfad über die Niveaukomponente.

6 Diskussion

Die vorliegende Studie hatte das Ziel, die Güte der Urteilsgenauigkeit von Musiklehrkräften im für den Musikunterricht wichtigen Bereich der Musikpraxis am Beispiel des Klassengesangs zu untersuchen. Dabei interessierte auch deren potenzieller Einfluss auf Unterrichtsplanung. Es konnte gezeigt werden, dass Musiklehrkräfte zwar das Leistungsniveau von Lerngruppen in der Jahrgangsstufe 9 differenziert einschätzen können (Niveaukomponente), dass aber die Einschätzung der Leistungsbreite (Differenzierungskomponente) nicht hinreichend funktioniert. Eine multivariate Varianzanalyse erbrachte einen Beleg dafür, dass realistische Variationen der Leistungsbreite ohne Einfluss für die Einschätzung der Differenzierungskomponente blieben. Allerdings zeigte sich ein Interaktionseffekt zwischen Leistungsniveau und Leistungsbreite: Leistungsschwächere Lerngruppen werden grundsätzlich als heterogener wahrgenommen; eine Unterscheidung zwischen homogeneren und heterogeneren Lerngruppen scheint lediglich in leistungsstärkeren Gruppen zu gelingen (vgl. Abb. 4).

Die Varianz der Unterrichtsplanungsvariablen in dieser Studie war ausschließlich systematisch beeinflusst durch die Niveaukomponente, nicht hingegen durch die Differenzierungskomponente. Allerdings war auch die Varianz der Differen-

zierungskomponente lediglich durch das Leistungsniveau, nicht jedoch durch die Leistungsbreite der Lerngruppen beeinflusst. Diese Befunde sind für das Schulfach Musik von besonderer Relevanz, da adaptiver Unterricht sowohl auf der Makro- als auch auf der Mikroebene nur dann funktionieren kann, wenn die lernrelevanten Unterschiede zwischen den Schüler:innen ausreichend bekannt sind. Dabei gilt die aktuelle Leistung als besonders wichtiger lernrelevanter Unterschied (vgl. Helmke 2017, S. 252). Die Leistungsbreite der Lerngruppe eben auch im Bereich der Musikpraxis zu kennen, im Idealfall sogar konkrete Kenntnisse über die individuellen Leistungsstände zu haben, ist für einen adaptiven Musikunterricht daher unerlässlich. Die Ergebnisse dieser Studie weisen aber darauf hin, dass Musiklehrkräfte anhand von Gruppenleistungen kaum in der Lage sind, die Leistungsbreite zielgenau einzuschätzen. Viel mehr wird eine vermeintliche Einschätzung der Leistungsbreite anhand des mittleren Leistungsniveaus getroffen, dann aber für unterrichtsrelevante Planungsentscheidungen nicht einbezogen. Eine Anpassung des Unterrichtsgeschehens und der -materialien an die individuellen Leistungsunterschiede scheint daher nahezu unmöglich, bzw. kann sich nur auf einem sehr groben Niveau bewegen. Dabei sollte darauf hingewiesen werden, dass die Passung zwischen pädagogischer Diagnostik und den Unterrichtsangeboten auf der Grundlage von normativen Überlegungen zu Unterrichtsstrategien basiert. Ob das Einplanen von mehr verschiedenen Aufgaben bei leistungsheterogeneren Gruppen tatsächlich die beste Lösung ist, kann durchaus kritisch hinterfragt werden. Für diesen Beitrag ist aber vor allem entscheidend, dass im Sinne der Grunddefinition von adaptivem Unterricht als Anpassung des Unterrichts an die lernrelevanten Unterschiede zwischen den Schüler:innen überhaupt ein Einfluss der Differenzierungskomponente auf Unterrichtsplanungsvariablen festzustellen sein sollte.

Der Befund, dass Musiklehrkräfte ihre Informationen für beide untersuchten Komponenten der Urteilsgenauigkeit aus derselben Grundeigenschaft ziehen, nämlich dem Leistungsniveau der Lerngruppe, sollte als Hinweis verstanden werden, dass hier in der Lehrkräftebildung noch Optimierungspotenzial vorhanden ist. Diese offensichtliche Fehlvorstellung sollte gerade für derartige Unterrichtsbereiche, in denen die individuelle Leistung nicht nur von der eigenen Kompetenzausprägung, sondern insbesondere auch vom Kompetenzgefüge der Lerngruppe abhängig ist, thematisiert werden, um hier Beurteilungsfehler zu reduzieren. Solche Unterrichtsbereiche beschränken sich dabei nicht nur auf die Musikpraxis im Musikunterricht. Sie betreffen möglicherweise auch Teile des Sportunterrichts, wenn Teamsportarten Unterrichtsgegenstand sind, sie betreffen aber auch die Qualität von Gruppenarbeiten oder Diskussionen im Unterricht. Gerade wenn hier keine soziale, sondern eine sachliche Bezugsnorm angelegt werden soll, erhöht sich die Wahrscheinlichkeit von Fehlurteilen. Um auch für solche Unterrichtssituationen die pädagogische Diagnostik besser zu schulen, könnten in der Zukunft digitale Unterstützungssysteme in Form von virtuellen Klassenräumen eingesetzt werden (vgl. Kaiser et al. 2012), die vermutlich fachspezifisch sehr unterschiedlich ausgestaltet werden müssen.

Musikpraxis wird im schulischen Musikunterricht in der Regel im Klassenverband oder in Kleingruppen durchgeführt, nur in den seltensten Fällen wird individuell musiziert. Im Studium werden angehende Musiklehrkräfte sowohl künstlerisch als auch pädagogisch ausgebildet. Dabei kann bezogen auf die in dieser Studie untersuchte

Situation durchaus ein Konflikt zwischen diesen beiden Schwerpunkten entstehen. Der künstlerische Anteil des Studiums zielt auf eine ästhetische Praxis ab, wie wir sie auch im außerschulischen Musikleben anfinden. Musizierende Gruppen wie Chöre oder Orchester werden als einheitliche Gruppen, als Klangkörper interpretiert, es zählt vor allem das musikalische Gesamtergebnis. Natürlich wird darauf geachtet, musikalische Arbeit so zu gestalten, dass einzelne Störfaktoren wie individuelle falsche Töne etc. identifiziert und behoben werden können. Wenn aber eine schwächere Leistung das musikalische Gesamtergebnis nicht negativ beeinflusst, wird an dieser Stelle oftmals nicht mehr korrigiert. Diese Sichtweise auf Lerngruppen als Klangkörper findet sich auch in zahlreichen curricularen Beschreibungen wieder (vgl. Hasselhorn 2017), geht im Zweifelsfall aber an dem Grundsatz der individuellen Kompetenz- oder Leistungsentwicklung vorbei. Diesen möglichen Konflikt gilt es, bereits in der ersten Phase der Lehrkräftebildung zu thematisieren, um ein optimales Lernen aller Schüler:innen ermöglichen zu können. Künstlerischer Anspruch sollte dabei die pädagogischen Ziele des Musikunterrichts nicht unterlaufen.

Unter der Voraussetzung, dass im Musikunterricht die individuelle Kompetenzentwicklung auch für die Musikpraxis über der ästhetischen Qualität eines musikalischen Gesamtergebnisses steht, deuten die Ergebnisse dieser Studie darauf hin, dass besonders im Bereich der pädagogischen Diagnostik Entwicklungsbedarf besteht. Dass im vorliegenden Datensatz keinerlei Alterseffekt gefunden werden konnte, kann dahingehend interpretiert werden, dass diagnostisches Verhalten von Musiklehrkräften in Studium und Vorbereitungsdienst erlernt und anschließend nur noch angewendet, aber nicht mehr systematisch weiterentwickelt wird. Es scheint hier keinen auf Berufserfahrung basierenden Effekt zunehmender Expertise zu geben. Grund hierfür könnten mangelnde Supervisions- und Feedbackmöglichkeiten für Lehrkräfte sein. Es gibt kaum eine Möglichkeit, die Güte der eigenen pädagogischen Diagnostik zu überprüfen.

Berücksichtigt werden sollte, dass in der vorliegenden Studie die Niveau- und die Differenzierungskomponente anders als üblich erhoben wurde. In der Regel werden Lehrkräfte aufgefordert, die individuellen Leistungen ihrer Schüler:innen zu konkreten Testaufgaben einzuschätzen. Aus diesen Einschätzungen werden dann die Komponenten der Urteilsgenauigkeit berechnet. Hier dagegen wurden die Lehrkräfte aufgefordert, die beiden erhobenen Komponenten direkt mit einem Gesamturteil einzuschätzen. Wir halten dieses Vorgehen für den untersuchten Unterrichtsbereich für sinnvoll, da überprüfbare Einzelleistungen hier an Validität einbüßen (s. oben). Außerdem ist es auch aus Gründen der ökologischen Validität sinnvoll. Lehrkräfte in Mathematik und Deutsch können ihre Urteile aus einer Vielzahl individueller Leistungsproben (mündlich wie schriftlich) erstellen. Diese Fächer finden in der Regel 4- bis 5-stündig statt. Das Fach Musik wird z. B. in der Jahrgangsstufe 9 in den meisten Bundesländern einstündig unterrichtet. Bei 40 Schulwochen ergibt das eine Gesamtunterrichtszeit von 30 Zeitstunden. Legt man als obere Abschätzung an, dass ein Drittel der Unterrichtszeit für Musikpraxis verwendet wird, werden dafür 10 Zeitstunden im ganzen Schuljahr aufgewendet. In dieser Zeit findet aber auch die Erarbeitung statt, nicht nur die musikpraktische Präsentation. Musiklehrkräfte haben daher schon aus Zeitgründen kaum eine Möglichkeit, individuelle Leistungsproben systematisch zu sammeln, um daraus eine Bewertung abzuleiten. Sie müssen darauf

vertrauen können, dass sie gut genug ausgebildet wurden, um aus den Gruppenleistungen die richtigen Schlüsse für ihre pädagogische Diagnostik ziehen zu können. Wir verstehen daher den vorliegenden Beitrag als einen ersten Vorschlag, die bislang gängige methodische Praxis zur Erfassung von Urteilsgenauigkeit dahingehend zu erweitern, dass auch Unterrichtsfächer bzw. Bereiche dieser Fächer, die sich dem üblichen methodischen Zugang entziehen, erfassbar zu machen. Die vorliegenden Ergebnisse müssen dafür selbstverständlich repliziert und möglicherweise erweitert werden.

Die hier vorgestellten Ergebnisse sind möglicherweise auch von übergeordnetem theoretischem Interesse. Leistungsniveau und Leistungsbreite sind zwei unterschiedliche, nicht zwingend voneinander abhängige Kategorien. Diese Unabhängigkeit der Kategorien wird durch die klassische Erhebungsform möglicherweise nicht vollständig abgebildet. Wir sehen in den hier präsentierten Daten, dass diese Kategorien auch in den Vorstellungen der Lehrkräfte offenbar nicht unabhängig voneinander angewendet werden können. So zeigen die Analysen, dass leistungsstärkere Lerngruppen als homogener und leistungsschwächere Lerngruppen als heterogener eingeschätzt werden. Diese Fehlvorstellung sollte im Sinne individueller Förderung unbedingt korrigiert werden, sonst erhöht sich gerade für Fächer wie Musik die Gefahr, dass der Unterricht vorwiegend an den Leistungen der Leistungsspitze ausgerichtet wird.

In diesem Zusammenhang überraschend sind auch die Einschätzungen der Lehrkräfte, wie viele Schüler:innen der Jahrgangsstufe 9 ihrer Ansicht nach in der Lage sind, die in der Studie vorgeschlagenen Unterrichtsaufgaben zu bewältigen. Die Lehrkräfte schätzen im Mittel, dass nur knapp die Hälfte der Jugendlichen in diesem Alter in der Lage ist, das Begleit-Ostinato, das lediglich aus zwei Tönen im Quartabstand besteht und keinerlei ernste rhythmische Herausforderung darstellt, intonationssauber zu singen. Diese Einschätzung erzwingt die Frage, wie wichtig das Singen für den Musikunterricht eigentlich ist. Falls Kompetenzerwerb im Singen Ziel von schulischem Musikunterricht ist, erscheint diese Einschätzung wie ein Eingeständnis, dass die anvisierten Ziele in der Mehrheit der Fälle nicht erreicht werden. Falls Singen lediglich im Unterricht eingesetzt wird, um andere Ziele zu erreichen, darf man durchaus die Frage stellen, ob dies möglich ist, wenn sich die Leistungen im Singen auf einem derart niedrigen Niveau bewegen.

Aufgrund des gewählten Experimentaldesigns haben wir darauf verzichtet, Musiklehrkräfte solche Schüler:innen einschätzen zu lassen, die sie bereits gut kennen. Da sowohl zum Lesen (z. B. Oerke et al. 2016) als auch zu mathematischen Leistungen (z. B. Stang und Urhahne 2016) Ergebnisse vorliegen, die nahelegen, dass sich diagnostische Urteile im Verlauf eines Schuljahres bezogen auf eine konkrete Lerngruppe zumindest in Teilen leicht zu verbessern scheinen, wäre es zukünftig sinnvoll, auch diesen Faktor bei der Einschätzung von Leistungen im Fach Musik zu berücksichtigen. Hierfür müssten insbesondere für solche Unterrichtsbereiche wie die Musikpraxis, die kaum individuell ausführbar sind, möglicherweise weitergehende methodische Überlegungen angestellt werden. Dabei ist ebenfalls zu diskutieren, wie eine möglichst optimale Passung zwischen gruppenorientierten Unterrichtsinhalten wie dem Gruppenmusizieren und differenzierenden Unterrichtsangeboten aussehen kann. Diese Besonderheit betrifft sicher nicht ausschließlich das Schulfach Musik, tritt dort aber besonders häufig auf.

Zusatzmaterial online Datensätze und Auswertungssytaxen der im Ergebnisteil beschriebenen Auswertungen sind als zusätzliche Informationen in der Online-Version dieses Artikels (https://doi.org/10.1007/s11618-022-01105-z) enthalten.

Funding Open Access funding enabled and organized by Projekt DEAL.

Interessenkonflikt J. Hasselhorn, F. Platz und C. Harnischmacher geben an, dass kein Interessenkonflikt besteht.

Literatur

Arnold, K.-H. (2008). Förderung schulfachlicher Fähigkeiten. Vorbemerkung. In K.-H. Arnold, O. Graumann & A. Rakhkochkine (Hrsg.), *Handbuch Förderung* (S. 256–257). Weinheim: Beltz.

Brunner, P. (2020). Peter und der Wolf. Ein Musikstück mit allen Sinnen kennenlernen. *Musik in der Grundschule*, *2*, 40–45.

Busch, B. (2010). Einfach musizieren!? Instrumentaler Gruppenunterricht in der Grundschule. In T. Greuel, U. Kranefeld & E. Szczepaniak (Hrsg.), *Jedem Kind (s) ein Instrument–Die Musikschule in der Grundschule* (S. 71–84). Aachen: Shaker.

Dartsch, M., & Heß, C. (2018). Instrumentalspiel als didaktisches Handlungsfeld. In M. Dartsch, J. Knigge, A. Niessen, F. Platz & C. Stöger (Hrsg.), *Handbuch Musikpädagogik* (S. 302–310). Münster: Waxmann.

Fischer, C., Rott, D., Veber, M., Fischer-Ontrup, C., & Gralla, A. (2014). *Individuelle Förderung als schulische Herausforderung*. Berlin: Friedrich-Ebert-Stiftung.

Gibbons, P. (2015). *Scaffolding language. Scaffolding learning* (2. Aufl.). Portsmouth: Heinemann.

Göllner, M., & Niessen, A. (2016). Planungsanpassung als adaptive Maßnahme in musikpädagogischen Lernsituationen im Spiegel qualitativer Interviews. In J. Knigge & A. Niessen (Hrsg.), *Musikpädagogik und Erziehungswissenschaft* (S. 121–135). Münster: Waxmann.

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). Aachen: 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). Aachen: Shaker.

Häcker, T. (2017). Individualisierter Unterricht. In T. Bohl, J. Budde & M. Rieger-Ladich (Hrsg.), *Umgang mit Heterogenität in Schule und Unterricht* (S. 275–290). Bad Heilbrunn: Klinkhardt.

Hardy, I., Decristan, J., & Klieme, E. (2019). Adaptive teaching in research on learning and instruction. *Journal for Educational Research Online*, *11*(2), 169–191.

Harnischmacher, C. (2012). *Subjektorientierte Musikerziehung. Eine Theorie des Lernens und Lehrens von Musik*. Augsburg: Wißner.

Harnischmacher, C., Hasselhorn, J., Schulz-Heidorf, K., & Temming, L. (2021). Individuelle Förderung, Autonomieförderung, Motivation und Musikinteresse in der Familie als Prädiktoren des Kompetenzerlebens und der Zensurenrelevanz im Musikunterricht. *Beiträge empirischer Musikpädagogik*, *12*, 1–19.

Hartinger, A. (2005). Verschiedene Formen der Öffnung von Unterricht und ihre Auswirkungen auf das Selbstbestimmungsempfinden von Grundschulkindern. *Zeitschrift für Pädagogik*, *51*, 397–414.
Hasselhorn, J. (2015). *Messbarkeit musikpraktischer Kompetenzen bei Schülerinnen und Schülern*. Münster: Waxmann.
Hasselhorn, J. (2017). Musikpraktische Kompetenzen – Theoretische Grundlagen und Ableitungen für die Unterrichtspraxis. In B. Hofmann (Hrsg.), *Plan\mäl\ßig. Schulmusik unter den Vorzeichen von Bildungsstandards und Kompetenzorientierung* (S. 27–44). Innsbruck: Helbling.
Hasselhorn, J., & Lehmann, A. C. (2015). Leistungsheterogenität im Musikunterricht. Eine empirische Untersuchung zu Leistungsunterschieden im Bereich der Musikpraxis in Jahrgangsstufe 9. In J. Knigge & A. Niessen (Hrsg.), *Theoretische Rahmung und Theoriebildung in der musikpädagogischen Forschung* (S. 163–176). Münster: Waxmann.
Hasselhorn, J., & McElvany, N. (2016). Die Bedeutung außerschulischer Prädiktoren für schulrelevante musikpraktische Kompetenzen. In N. McElvany, R. Strietholt, H. G. Holtappels & W. Bos (Hrsg.), *Jahrbuch der Schulentwicklung* (Bd. 19, S. 168–205). Weinheim: Juventa.
Hasselhorn, J., & Wolf, A. (2018). Assessment, Bewertung und Musikkritik. In A. C. Lehmann & R. Kopiez (Hrsg.), *Handbuch Musikpsychologie* (S. 389–410). Göttingen: Hogrefe.
Hasselhorn, M., & Gold, A. (2017). *Pädagogische Psychologie* (4. Aufl.). Stuttgart: Kohlhammer.
Hasselhorn, J., Hasselhorn, S., Altenmüller, E., & Hasselhorn, M. (2012). Aufführungsangst bei Studierenden in den Fächern Gesang und Klavier. Verändert sie sich im Laufe der Ausbildung? *Beiträge empirischer Musikpädagogik*, *3*(2), 1–15.
Helmke, A. (2017). *Unterrichtsqualität und Lehrerprofessionalität* (7. Aufl.). Seelze: Klett.
Hesse, I., & Latzko, B. (2017). *Diagnostik für Lehrkräfte* (3. Aufl.). Opladen: Budrich.
Jank, W. (2018). Unterrichtsgestaltung: Schulischer Musikunterricht. In M. Dartsch, J. Knigge, A. Niessen, F. Platz & C. Stöger (Hrsg.), *Handbuch Musikpädagogik* (S. 355–360). Münster: Waxmann.
Jank, W. (2021). *Musikdidaktik. Praxishandbuch* (Bd. 9). Berlin: Cornelsen.
Jank, W., & Meyer, H. (2017). Zur Unterrichtsplanung. In W. Jank (Hrsg.), *Musikdidaktik* (6. Aufl., S. 132–141). Berlin: Cornelsen.
Jordan, A.-K. (2014). *Empirische Validierung eines Kompetenzmodells für das Fach Musik. Teilkompetenz Musik wahrnehmen und Kontextualisieren*. Münster: Waxmann.
Kaiser, J., Helm, F., Retelsdorf, J., Südkamp, A., & Möller, J. (2012). Zum Zusammenhang von Intelligenz und Urteilsgenauigkeit bei der Beurteilung von Schülerleistungen im Simulierten Klassenraum. *Zeitschrift für Pädagogische Psychologie*, *26*(4), 251–261.
Karst, K., & Förster, N. (2017). Ansätze zur Modellierung diagnostischer Kompetenz. In A. Südkamp & A.-K. Praetorius (Hrsg.), *Diagnostische Kompetenz von Lehrkräften* (S. 19–66). Münster: Waxmann.
Kim, H.-Y. (2013). Statistical notes for clinical researchers: assessing normal distribution using skewness and kurtosis. *Restorative Dentistry and Endontics*, *38*, 52–54. https://doi.org/10.5395/rde.2013.38.1.52.
Klieme, E., & Warwas, J. (2011). Konzepte der individuellen Förderung. *Zeitschrift für Pädagogik*, *57*(6), 805–818.
Kline, R. B. (2015). *Principles and practice of structural equation modeling* (4. Aufl.). New York: Guilford.
König, J., Buchholz, C., & Dohmen, D. (2015). Analyse von schriftlichen Unterrichtsplanungen: Empirische Befunde zur didaktischen Adaptivität als Aspekt der Planungskompetenz angehender Lehrkräfte. *Zeitschrift für Erziehungswissenschaft*, *18*(2), 275–404.
Kranefeld, U., & Heberle, K. (2016). „Dankeschön! Was war das Problem?" Zur Rekonstruktion mikroadaptiver Handlungsroutinen im Musikunterricht. In J. Knigge & A. Niessen (Hrsg.), *Musikpädagogik und Erziehungswissenschaft* (S. 137–153). Münster: Waxmann.
Kranefeld, U., Heberle, K., & Pankoke, C. (2015). Zur videographischen Erfassung von Passungsprozessen im Musik-unterricht – Methodologische Überlegungen und fallanalytische Perspektiven. *Beiträge empirischer Musikpädagogik*, *6*(2), 2–19.
Kunter, M., & Trautwein, U. (2013). *Psychologie des Unterrichts*. Paderborn: Schöningh.
Legrand, R. (2012). *Der Pädago-Gig. Musikvermittlung als Projekt*. Hannover: ifmpf.
Leuders, T., Dörfler, T., Leuders, J., & Philipp, K. (2018). Diagnostic competence of mathematics teachers: unpacking a complex construct. In T. Leuders, K. Philipp & J. Leuders (Hrsg.), *Diagnostic competence of mathematics teachers* (S. 3–31). Heidelberg: Springer. https://doi.org/10.1007/978-3-319-66327-2_1.
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 (Hrsg.), *Advancing music education through assessment: honoring culture, diversity, and practice* (S. 137–154). Chicago: GIA.

Messner, R., & Blum, W. (2019). Der Mythos des offenen Unterrichts – unter Einbeziehung von Befunden aus dem DISUM-Projekt. In U. Steffens & R. Messner (Hrsg.), *Unterrichtsqualität. Konzepte und Bilanzen gelingenden Lehrens und Lernens* (S. 57–90). Münster: Waxmann.

Moeller, H., & Castringius, S. (2005). Aufführungsangst als gesundheitliches Risiko bei Musikern – Ursachen, Therapie und Prävention. In R. Oerter & T. H. Stoffer (Hrsg.), *Spezielle Musikpsychologie* (S. 525–554). Göttingen: Hogrefe.

Müller, K., Gartmeier, M., & Prenzel, M. (2013). Kompetenzorientierter Unterricht im Kontext nationaler Bildungsstandards. *Bildung und Erziehung*, *66*, 127–144.

Oerke, B., McElvany, N., Ohle, A., Ullrich, M., & Horz, H. (2016). Verbessert sich die diagnostische Urteilsgenauigkeit von Lehrkräften bei längerem Kontakt mit der Klasse? *Psychologie in Erziehung und Unterricht*, *63*, 34–47.

Ott, T. (2018). Konzeptionen und zentrale Orientierungen für schulischen Musikunterricht. In M. Dartsch, J. Knigge, A. Niessen, F. Platz & C. Stöger (Hrsg.), *Handbuch Musikpädagogik* (S. 284–288). Münster: Waxmann.

Platz, F. (2018). Formen der Leistungserfassung und -rückmeldung. In M. Dartsch, J. Knigge, A. Niessen, F. Platz & C. Stöger (Hrsg.), *Handbuch Musikpädagogik* (S. 377–384). Münster: Waxmann.

Platz, F., & Kopiez, R. (2012). When the eye listens: a meta-analysis of how audio-visual presentation enhances the appreciation of music performance. *Music Perception*, *30*(1), 71–83.

Platz, F., & Kopiez, R. (2022). Stage behavior, impression management, and charisma. In G. McPherson (Hrsg.), *Oxford handbook of music performance* (S. 84–102). Oxford: Oxford University Press.

Platz, F., Wolf, A., & Hasselhorn, J. (2021). Lässt sich die Lernwirksamkeit von Musikunterricht durch den Einsatz neuer (digitaler) Medien steigern? In K. Martin & S. Stick (Hrsg.), *Musikpädagogik in Zeiten von Globalisierung und Digitalisierung* (S. 82–102). Weimar: HfM Franz Liszt.

Praetorius, A.-K., & Südkamp, A. (2017). Eine Einführung in das Thema der diagnostischen Kompetenz von Lehrkräften. In A. Südkamp & A.-K. Praetorius (Hrsg.), *Diagnostische Kompetenz von Lehrkräften* (S. 13–18). Münster: Waxmann.

Praetorius, A.-K., Hetmanek, A., Herppich, S., & Ufer, S. (2017). Herausforderungen bei der empirischen Erforschung diagnostischer Kompetenz. In A. Südkamp & A.-K. Praetorius (Hrsg.), *Diagnostische Kompetenz von Lehrkräften* (S. 95–101). Münster: Waxmann.

Prediger, S., & Pöhler, B. (2015). The interplay of micro- and macro-scaffolding: an empirical reconstruction for the case of an intervention on percentages. *ZDM Mathematics Education*, *47*(7), 1179–1194.

Rey, T., Lohse-Bossenz, H., Wacker, A., & Heyl, V. (2018). Adaptive Planungskompetenz bei angehenden Lehrkräften in der zweiten Phase der Lehrerbildung. Befunde einer Pilotierungsstudie aus Baden-Württemberg. *heiEDUCATION journal*, *1*(2), 127–150.

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

Schrader, F.-W., & Helmke, A. (1987). Diagnostische Kompetenz von Lehrern: Komponenten und Wirkungen. *Empirische Pädagogik*, *1*, 27–52.

Shandro, S. (2015). A comparison of standardized curricula for independent voice instruction. *Journal of Singing*, *71*(4), 497–506.

Stadler-Elmer, S. (2005). Entwicklung des Singens. In R. Oerter & T. H. Stoffer (Hrsg.), *Spezielle Musikpsychologie* (S. 123–154). Göttingen: Hogrefe.

Stang, J., & Urhahne, D. (2016). Stabilität, Bezugsnormorientierung und Auswirkungen der Urteilsgenauigkeit. *Zeitschrift für Pädagogische Psychologie*, *30*(4), 251–262.

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

Urhane, D., & Wijnia, L. (2021). A review on the accuracy of teacher judgments. *Educaional Research Review*, *32*, 100374.

Venus, D. (1969). *Unterweisungen im Musikhören*. Wuppertal: Henn.

Watkins, J. G., & Farnum, S. E. (1954). *The Watkins-Farnum performance scale: form A*. Winona: Hal Leonard.

Weinert, F. E. (1997). Notwendige Methodenvielfalt: Unterschiedliche Lernfähigkeit der Schüler erfordern variable Unterrichtsmethoden des Lehrers. *Friedrich Jahresheft: Lernmethoden-Lehrmethoden-Wege zur Selbstständigkeit*, *15*, 50–52.

Weston, R., & Gore, P. A. (2006). A brief guide to structural equation modeling. *The Counseling Psychologist*, *34*(5), 719–751. https://doi.org/10.1177/0011000006286345.

Zill, E. (2016). „Wow, das klingt schon richtig gut...". Eine qualitative Studie zu musikalisch-ästhetischen Erfahrungen von Schülern in produktionsorientierten Projekten. In J. Knigge & A. Niessen (Hrsg.), *Musikpädagogik und Erziehungswissenschaft* (S. 231–247). Münster: Waxmann.