Musikbezogenes Argumentieren

Testentwicklung und Kompetenzmodellierung

Inauguraldissertation
zur
Erlangung des Doktorgrades
der Humanwissenschaftlichen Fakultät
der Universität zu Köln
nach der Promotionsordnung vom 18.12.2018
vorgelegt von

Julia Ehninger
geboren in
Stuttgart

Köln, September 2021

Erstgutachter:	Prof. Dr. Christian Rolle
Zweitgutachter:	Prof. Dr. Jens Knigge

Diese Dissertation wurde von der Humanwissenschaftlichen Fakultät der Universität zu Köln im Januar 2022 angenommen.

Danksagung

In den letzten fünf Jahren haben mich zahlreiche Personen beim Zustandekommen dieser Arbeit unterstützt. Ohne den Austausch mit diesen großartigen Menschen hätte ich die Arbeit in dieser Form nicht schreiben können.

Mein herzlicher Dank gilt Christian Rolle und Jens Knigge, die beide durch ihre Unterstützung wesentlich zur Entstehung der Dissertation beigetragen haben. In den letzten fünf Jahren gab es zahlreiche konstruktive Diskussionen und Gespräche, auf die ich zurückblicken kann. Ich konnte mich dabei stets auf Eure Unterstützung verlassen!

Bei Michael Schurig möchte ich mich herzlich für den intensiven Austausch und die Hilfe bei der statistischen Datenanalyse bedanken. Die aufwändige Auswertung des Datenmaterials wäre nicht möglich gewesen ohne Alina Öttinger, die alle Testantworten der Hauptstudie mitkodiert hat und maßgeblich bei der Überarbeitung der Kodierregeln mitgewirkt hat. Ebenfalls möchte ich mich bei allen Schüler*innen, Lehrkräften und Eltern bedanken, die die Durchführung der Studie durch ihre Kooperation ermöglicht haben!

Ein besonderer Dank gebührt allen tollen Kolleg*innen, die über die Jahre immer wieder mit mir die Arbeit diskutiert und meine Texte kritisch gelesen haben. Ihr habt meine Promotionszeit unglaublich bereichert und inspiriert! Insbesondere möchte ich mich bedanken bei Linus Eusterbrock, Daniel Fiedler, Thomas Gottschalk, Annemarie Haberecht, Johannes Hasselhorn, Theresa Meyer, Anna Rizzi und Julia Weber.

Die Studienstiftung des deutschen Volkes hat die letzten drei Jahre meine Arbeit finanziell und ideell gefördert. Für das Vertrauen möchte ich mich herzlich bedanken. Ohne diese Unterstützung wäre es mir nicht möglich gewesen, mich so intensiv mit meiner Dissertation zu beschäftigen. Ebenfalls bedanke ich mich herzlich bei der Graduiertenschule der Humanwissenschaftlichen Fakultät sowie der Graduiertenschule für LehrerInnenbildung der Universität zu Köln für die Förderung mehrerer Tagungs- und Workshopteilnahmen.

Abschließend möchte ich mich bei Florian Herzog für sein offenes Ohr und die gemeinsame Zeit bedanken. Ebenfalls gebührt ein riesengroßer Dank meiner tollen Familie, auf die ich stets zählen kann.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Zusammenfassung

Musikbezogene Argumentationskompetenz bezeichnet die Kompetenz, Urteile über Musikstücke begründen zu können (Knörzer et al., 2016). Diese Kompetenz ist bedeutsam, wenn über Musik diskutiert wird oder wenn musikbezogene Urteile verhandelt werden. Während Argumentationskompetenz in anderen Fachdidaktiken in den letzten Jahren zunehmend empirisch untersucht wurde, gibt es innerhalb der Musikpädagogik bisher nur wenig empirische Forschung in diesem Bereich. Um diesem Desiderat zu begegnen, wurde in der Dissertation ein neues Messinstrument für musikbezogene Argumentationskompetenz entwickelt (*MARKO*, *M*usikbezogene *AR*gumentations*KO*mpetenz). Es war Ziel der Arbeit, die Kompetenz valide, reliabel und objektiv zu messen, um so die Kompetenzanforderungen beim musikbezogenen Argumentieren auf Basis empirischer Daten zu klären.

Ausgehend von einem theoretischen Kompetenzmodell (Rolle, 2013, 2017) wurden Testaufgaben entwickelt, die in zwei Pretests erprobt wurden ($N = 391$). Der Test, der in der Hauptstudie zum Einsatz kam, bestand aus 25 offenen Items. 440 Schüler*innen der neunten bis zwölften gymnasialen Jahrgangsstufe sowie Musikstudierende nahmen an den Datenerhebungen der Hauptstudie teil. Für jedes Item wurden Kodierregeln entwickelt, deren Interrater-Reliabilität sichergestellt wurde. Alle Testaufgaben Testaufgaben erfüllten wesentliche psychometrische Gütekriterien wie lokale stochastische Unabhängigkeit und Itemhomogenität. Auf Basis der Testaufgaben konnte ein statistisches Modell geschätzt werden (eindimensionales Partial Credit Model). So wurden vier Kompetenzniveaubeschreibungen aus den empirischen Daten abgeleitet. Während Personen auf dem niedrigsten Niveau ihre eigene Meinung über die Musik äußern, indem sie sich auf saliente, also besonders herausstechende, musikalische Merkmale bezogen, diskutieren Personen auf dem höchsten Niveau unterschiedliche Meinungen und berücksichtigen den sozialen und kulturellen Kontext der Musik. Die Dissertation zeigt, dass es möglich ist, musikbezogene Argumentationskompetenz mithilfe des entwickelten *MARKO*-Test zuverlässig zu messen. So leistet die vorliegende Arbeit einen wichtigen Beitrag bei der Modellierung musikbezogener Kompetenzen.

Die kumulative Dissertation besteht aus vier Publikationen (Ehninger, 2021; Ehninger et al., 2021a; Ehninger et al., 2021b; Ehninger & Rolle, 2020) sowie dem vorliegenden Manteltext. Bei Ehninger (2021) handelt es sich um eine Methodenreflexion über empirische Forschung zu musikbezogener Argumentationskompetenz. Die Entwicklung von Testaufgaben für den *MARKO*-Kompetenztest ist Gegenstand zwei weiterer Publikationen (Ehninger et al., 2021a; Ehninger & Rolle, 2020) und die zentralen Ergebnisse der Hauptstudie werden in Ehninger et al. (2021b) diskutiert.

Abstract

Music-related argumentative competence can be defined as the ability to justify judgments about music (Knörzer et al., 2016). This competence is relevant when music is discussed or when music-related judgments are negotiated. While argumentative competence has been increasingly empirically studied in other educational contexts in recent years, there has been little empirical research in music education. To address this lack of research, a new measurement instrument for music-related argumentation competence was designed in this dissertation. The aim of the thesis was to measure the competence in a valid, reliable, and objective way in order to specify the competence requirements based on empirical data.

Based on a theoretical competency model by Rolle (2013, 2017), test items were designed and tested in two pretests ($N = 391$). The final test consisted of 25 open-ended items. 440 ninth through twelfth grade high school students and university students participated in the main study. All test items fulfilled psychometric criteria such as inter-rater reliability, local stochastic independence, and item homogeneity. Based on the test items, a statistical model was estimated (one-dimensional partial credit model) and descriptions of four different proficiency levels were derived from the empirical data. While persons at the lowest proficiency level articulated their opinion about music referring to salient musical attributes, persons at the highest level considered different opinions of the music and discussed its social and cultural context. This dissertation makes an important contribution to empirical research on music-related competences. For the first time, it is possible to measure music-related argumentative competence with a competency test.

The cumulative dissertation consists of four publications (Ehninger, 2021; Ehninger et al., 2021a; Ehninger et al., 2021b; Ehninger & Rolle, 2020). Ehninger (2021) is a methodological reflection of empirical research on music-related argumentative competence. The design of test items for the *MARKO* competency test is the subject of two other publications (Ehninger et al., 2021a; Ehninger & Rolle, 2020) and the results of the main study are discussed in Ehninger et al. (2021b).

Publikationsübersicht

Publikation A

Ehninger, J. (2021). Wie lässt sich musikbezogene Argumentationskompetenz empirisch untersuchen? Über die empirische Erforschung einer facettenreichen Kompetenz. *Beiträge empirischer Musikpädagogik*, *12*. https://www.b-em.info/index.php/ojs/article/view/192

Publikation B

Ehninger, J. & Rolle, C. (2020). Musikbezogenes Argumentieren – Nur Geschmacksache? Über die Entwicklung eines Kompetenztests. In M. Schwarzbauer & K. Steinhauser (Hrsg.), *‚Nur' Geschmacksache? Der Umgang mit kreativen Leistungen im Musik- und Kunstunterricht* (S. 168–182). LIT.

Publikation C

Ehninger, J., Knigge, J. & Rolle, C. (2021a). Musikbezogene Argumentationskompetenz. Ein Werkstattbericht über die Entwicklung von Testaufgaben. In A. Budke & F. Schäbitz (Hrsg.), *Argumentieren und Vergleichen* (S. 93–112). LIT.

Publikation D

Ehninger, J., Knigge, J., Schurig, M. & Rolle, C. (2021b). A New Measurement Instrument for Music-Related Argumentative Competence: The MARKO Competency Test and Competency Model. *Frontiers in Education*, *6*(191). https://doi.org/10.3389/feduc.2021.668538

Datenverfügbarkeit

Die Item- und Skalendokumentation, alle verwendeten Datensätze sowie das *R*-Skript können hier abgerufen werden:

https://osf.io/zvp4b/?view_only=a8427eb5a0ce4b8384243a72e7e1aa1e

Einleitung

Wenn wir musikalische Erfahrungen oder Höreindrücke für andere nachvollziehbar machen wollen, passiert dies häufig über Sprache. Dabei kann es vorkommen, dass wir schildern, weshalb uns ein bestimmtes Musikstück besonders gefällt oder weshalb wir von einer Interpretation nicht überzeugt sind. Auch überall dort, wo Musik geprobt wird, ist Sprache ein wichtiges Medium der Verständigung. Dann wird zwischen den Musizierenden oftmals verbal ausgehandelt, wie Musik klingen soll, sei es, dass die Sänger*innen eines Chors die Anweisungen von Dirigent*innen umsetzen, eine Band einen Song entwickelt oder sich Mitglieder eines Streichquartetts auf ein Tempo einigen müssen. Auch im Musikunterricht spielt die verbale Beschäftigung mit Musik eine große Rolle. Hier wird Musik beschrieben, analysiert, interpretiert und diskutiert. Nicht selten stellen verbale Äußerungen im Unterricht oder in Klausuren sogar die (teilweise ausschließliche) Grundlage für Benotungen dar.

Die sprachliche Auseinandersetzung mit Musik ist ein zentrales Mittel für das Lehren und Lernen im Musikunterricht und ist deshalb auch fester Bestandteil von Schulcurricula. Die Fähigkeit, Musik kriteriengeleitet zu beurteilen ist bereits in vielen Lehrplänen für die Sekundarstufe I festgehalten.[1] In der Abiturprüfung im Fach Musik soll schließlich der Nachweis erbracht werden, „gestaltbildende Merkmale der Musik zu erkennen, zu beschreiben, zu analysieren, zu interpretieren, deren Wirkung und Bedeutung zu beschreiben und reflektierend zu beurteilen" (Kultusministerkonferenz, 2005, S. 7). Bei dieser diskursiven Auseinandersetzung mit dem Gegenstand Musik ist Argumentationskompetenz erforderlich. Diese kann in Bezug auf Musik definiert werden als Kompetenz, „verständlich, plausibel und differenziert ästhetische Werturteile über Musikstücke begründen zu können" (Knörzer et al., 2015, S. 148).

Die Fähigkeit, sich diskursiv mit Inhalten auseinanderzusetzen, ist nicht nur im Musikunterricht eine wichtige Voraussetzung für die Teilhabe am Unterrichtsgeschehen. Bildungssprachliche Handlungen wie das Argumentieren oder Erklären dienen dem Aufbau und der Aushandlung von Wissen (Morek & Heller, 2012; Morek et al., 2017). Es überrascht daher nicht, dass Argumentationskompetenz als Schlüsselkompetenz für den Lernerfolg von Schüler*innen gehandelt wird (Quasthoff et al., 2020b). Obwohl sich in den letzten Jahren zunehmend herauskristallisiert hat, dass Sprache für das Lernen in allen schulischen Fächern konstitutiv ist (u. a. Michalak et al., 2015; Quasthoff et al., 2020a; Schmölzer-Eibinger, 2013), gibt es bisher kaum theoretische und empirische Forschung über die Rolle von Sprache sowie sprachlichen Kompetenzen im Schulfach Musik (Bossen, 2017).

Als ein wesentliches Ziel von Musikunterricht legen Schulcurricula den Erwerb musikbezoge-

[1] u. a. Ministerium für Kultus, Jugend und Sport Baden-Württemberg, 2016; Ministerium für Schule und Berufsbildung des Landes Schleswig-Holstein, 2015; Ministerium für Schule und Bildung des Landes Nordrhein-Westfalen, 2019; Staatsinstitut für Schulqualität und Bildungsforschung München, 2021.

ner Kompetenzen fest. Bislang ist jedoch größtenteils ungeklärt, wie musikbezogene Kompetenzen gemessen werden können, wie diese strukturiert sind und über welche Kompetenzen Schüler*innen verfügen bzw. nicht verfügen (Hasselhorn & Knigge, 2018). Dies gilt ebenfalls für die Kompetenz, Urteile über Musik zu begründen. Während es seit den 2000er-Jahren eine Vielzahl an Studien gab, die fachliche Kompetenzen von Schüler*innen untersuchten (u. a. Beck & Klieme, 2007; Leutner et al., 2017; Stanat et al., 2017), gibt es bislang für das Unterrichtsfach Musik nur wenig empirische Forschung.

Diese Dissertation untersucht musikbezogene Argumentationskompetenz und begegnet dem Desiderat der musikpädagogischen Kompetenzforschung. Die Arbeit will Aufschluss darüber geben, wie die Kompetenz strukturiert ist. Bislang ist aus empirischer Sicht nur wenig darüber bekannt, auf welche Aspekte Personen Bezug nehmen, wenn sie musikbezogen argumentieren und welche Herausforderungen sich hierbei stellen. Die Dissertation verfolgte daher das Ziel, musikbezogene Argumentationskompetenz mithilfe eines eigens entwickelten Messinstruments zu modellieren, um so die Kompetenz auf Basis empirischer Daten beschreiben zu können. Zu diesem Zweck wurde im Rahmen der Dissertation ein Testinstrument für musikbezogenes Argumentieren entwickelt (*MARKO*; *M*usikbezogene *AR*gumentations*KO*mpetenz). Mit diesem Test wurden Daten von Schüler*innen der neunten bis zwölften Jahrgangsstufe sowie von Studierenden erhoben, um so die Struktur der Kompetenz in Form von Kompetenzniveaus zu beschreiben. Im Rahmen dieser kumulativen Dissertation sind vier Publikationen entstanden, die im folgenden Publikation A, B, C und D genannt werden (s. Übersicht auf S. 17). Die Publikationen sind in diesem Manteltext zusammenfassend dargestellt und werden thematisch wie methodisch in einen größeren Forschungszusammenhang eingeordnet. Zudem werden über die einzelnen Artikel hinaus reichende Ergebnisse dargestellt und diskutiert.

Der Manteltext gliedert sich in einen empirischen und theoretischen Teil. In Kapitel 1 gehe ich zunächst auf den Kompetenzbegriff und die Forschung zu musikbezogenen Kompetenzen ein. Kapitel 2 widmet sich dem zentralen Gegenstand, der in der vorliegenden Arbeit untersucht wird: Musikbezogene Argumentationskompetenz. Was genau unter Argumentieren verstanden wird, unterscheidet sich z. T. erheblich zwischen verschiedenen Disziplinen. In diesem Kapitel werden die Besonderheiten des musikbezogenen Argumentierens als einer Form ästhetischen Argumentierens thematisiert. In diesem Zusammenhang reflektiere ich auch die verschiedenen Möglichkeiten, musikbezogene Argumentationskompetenz empirisch zu untersuchen. Diese Methodenreflexion war Bestandteil von Publikation A (Ehninger, 2021). In Kapitel 3 werden schließlich die übergeordneten Ziele der Dissertation dargestellt, bevor der empirische Teil der Arbeit mit der Darstellung der verwendeten Methoden beginnt (Kapitel 4). Die Entwicklung von Testaufgaben für den Kompetenztest, das Testdesign sowie die Pilotstudie waren Gegenstand der Publikationen B und C und werden in Kapitel 5 zusammengefasst (Ehninger & Rolle, 2020; Ehninger et al., 2021a). In Kapitel 6 wird es schließlich um die Hauptstudie gehen, deren Ergebnisse in Publikation D veröffentlicht wurden (Ehninger et al., 2021b). Im Rahmen der Hauptstudie wurden Daten von 440 Schüler*innen der neunten bis zwölften gymnasialen Jahrgangstufe sowie von Musikstudierenden erhoben. Diese Daten wurden ausgewertet und so konnten Kompetenzniveaus für musikbezogenes

Argumentieren beschrieben und Kompetenzausprägungen gemessen werden. In diesem Zusammenhang werden ebenfalls Einflussfaktoren für die Ausprägung der Kompetenz diskutiert. Diese Analysen wurden zuvor noch nicht veröffentlicht. Die Arbeit schließt mit einer Diskussion der Ergebnisse und des Erkenntnisgewinns (Kapitel 7).

I.

Theoretischer Teil

1. Kompetenzorientierung in der Musikpädagogik

Seit dem „PISA-Schock“ Anfang der 2000er-Jahre hat sich das deutsche Bildungswesen umfassend verändert. Im Zuge der Bildungsreform in den Jahren 2003 und 2004 wurden nationale Bildungsstandards eingeführt, um die Qualität der schulischen Leistungen von Schüler*innen zu sichern. Zuvor hatte sich das deutsche Schulsystem hauptsächlich am *Input* orientiert. In Curricula war festgehalten, was Schüler*innen lernen sollten, also welchen Input sie erhalten sollten. Mit der Bildungsreform sollte sich Bildungspolitik nicht mehr am Input, sondern am *Output*, d. h. an den tatsächlichen Lernergebnissen von Schüler*innen orientieren (Klieme et al., 2003, S.11-12). Etwas vereinfacht ausgedrückt sollte in den Schulcurricula nun nicht mehr festgehalten werden, was Schüler*innen *lernen*, sondern was sie *können* sollten. Letzteres wurde nun mithilfe von Kompetenzen beschrieben. Mit der Kompetenzorientierung sollten Wissen und Können so vermittelt werden, dass keine „‚träge‘ und isolierten Kenntnisse und Fähigkeiten entstehen, sondern anwendungsfähiges Wissen und ganzheitliches Können“ (Klieme & Hartig, 2007, S. 13).

Der Paradigmenwechsel zur Kompetenzorientierung wurde in der Musikpädagogik kritisch diskutiert. Knapp zwanzig Jahre nach der Bildungsreform ist die Kompetenzorientierung jedoch zweifelsohne auch fester Bestandteil der deutschen Musikpädagogik. Dennoch gibt es bisher nur wenig empirische Forschung zu musikbezogenen Kompetenzen. Es ist bislang nur in Ansätzen geklärt, wie musikbezogene Kompetenzen gemessen werden können, wie sie strukturiert sind und über welche Kompetenzen Schüler*innen verfügen bzw. nicht verfügen (Knigge, 2014). Kapitel 1.2 gibt einen Überblick über die bisherige empirische Forschung zu musikbezogenen Kompetenzen. Zunächst werden jedoch die grundlegenden Begriffe *Kompetenz* und *Bildungsstandards* eingeführt (Kapitel 1.1).

1.1. Kompetenzen und Bildungsstandards

Wie bereits erwähnt, sollte sich das deutsche Schulsystem mit der Bildungsreform nicht mehr am Input, sondern am Output, also an den Leistungen und Lernergebnissen von Schüler*innen ausrichten (Klieme et al., 2003, S. 11-12). Was je nach Schulfach als Output zu verstehen ist, ist in Bildungsstandards festgehalten. Bildungsstandards orientieren sich an allgemeinen Bildungszielen und „legen fest, welche Kompetenzen die Kinder oder Jugendlichen bis zu einer bestimmten Jahrgangsstufe mindestens erworben haben sollen“ (Klieme et al., 2003, S. 9). Von der Kultusministerkonferenz wurden erstmals 2003 bzw. 2004 nationale Bildungsstandards für die Kernfächer

eingeführt, die mithilfe von Kompetenzanforderungen spezifiziert sind.[2] Mit Bildungsstandards sollte die Möglichkeit geschaffen werden, das Erreichen von Bildungszielen besser überprüfen zu können. Die Qualität von Bildung sollte so messbarer werden, als das zuvor der Fall war. Seit 2009 wird im Ländervergleich in den Jahrgangsstufen vier und neun geprüft, ob die erforderlichen Bildungsstandards erreicht wurden (Kultusministerkonferenz, 2021b).

Die Einführung von Bildungsstandards wurde – insbesondere im erziehungswissenschaftlichen und fachdidaktischen Diskurs – teilweise kritisch diskutiert, da der Kompetenzbegriff von verschiedenen Disziplinen mit unterschiedlichen theoretischen und normativen Bedeutungen verwendet wird (Klieme & Hartig, 2007, S. 13-14). Zur Identifikation eines für pädagogische Kontexte tragfähigen Kompetenzkonzepts sind für den nationalen wie auch internationalen Diskurs die Arbeiten Franz E. Weinerts maßgeblich. Nach seiner vielzitierten Definition bedeutet Kompetenz

> die bei Individuen verfügbaren oder durch sie erlernbaren kognitiven Fähigkeiten und Fertigkeiten, um bestimmte Probleme zu lösen, sowie die damit verbundenen motivationalen, volitionalen und sozialen Bereitschaften und Fähigkeiten, um die Problemlösungen in variablen Situationen erfolgreich und verantwortungsvoll nutzen zu können. (Weinert, 2001b, S. 27-28)

Im DFG-Schwerpunktprogramm „Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen“ wurde – wie von Weinert (2001a) für empirische Untersuchungen empfohlen – eine Trennung kognitiver und motivationaler Aspekte vorgenommen, um anschließend deren Wechselwirkung untersuchen zu können (Klieme & Leutner, 2006, S. 880). Daher definieren Klieme und Leutner (2006, S. 879) Kompetenzen als „kontextspezifische kognitive Leistungsdispositionen“. Nach Klieme et al. (2003, S. 21) sollen Kompetenzanforderungen systematisch in Kompetenzmodellen geordnet werden. Diese sollen dabei zwei Funktionen erfüllen: Sie beschreiben erstens „das Gefüge der Anforderungen, deren Bewältigung von Schülerinnen und Schülern erwartet wird“ und liefern zweitens „wissenschaftlich begründete Vorstellungen darüber, welche Abstufungen eine Kompetenz annehmen kann bzw. welche Grade oder Niveaustufen sich bei den einzelnen Schülerinnen und Schülern feststellen lassen“ können (Klieme et al., 2003, S. 74). Kompetenzmodelle werden häufig theoretisch hergeleitet (z. B. anhand fachdidaktischer Annahmen) und können mithilfe statistischer Verfahren empirisch überprüft werden (Klieme & Leutner, 2006).

Die empirische Bildungsforschung hat seit den 2000er-Jahren stark an Einfluss gewonnen. Nicht zuletzt deshalb, weil alle ein bis drei Jahre das Erreichen der nationalen Bildungsstandards im Rahmen internationaler oder nationaler Schulleistungsstudien überprüft wird (Überblick s. Stanat et al., 2019, S. 16). Das „Vermessen“ von Schüler*innen-Leistungen in Schulleistungsstudien wurde jedoch wiederholt kritisiert. Dabei wurde mitunter immer wieder der Vorwurf diskutiert,

[2] In der Primarstufe sind Bildungsstandards für die Fächer Deutsch und Mathematik festgelegt. Für den mittleren Schulabschluss bzw. die allgemeine Hochschulreife gibt es zudem Bildungsstandards für die erste Fremdsprache, Biologie und Physik (Kultusministerkonferenz, 2021a).

dass sich Schulleistungsstudien auf „das Messbare“ reduzieren (Tillmann, 2017, S. 11). Kompetenzen, die sich international nicht vergleichen lassen, wie z. B. literarästhetische Kompetenz im Fach Deutsch, spielen in internationalen Schulleistungsstudien keine Rolle.

Auch viele Musikpädagog*innen diskutierten die Kompetenzorientierung kritisch (z. B. Brenk, 2014; Rolle, 2014b; Vogt, 2008). Hasselhorn und Knigge stellen fest, dass in vielen Musik-Curricula das begriffliche Verständnis von Kompetenz nicht ausreichend expliziert wird (Hasselhorn & Knigge, 2018, S. 201; s. a. Knigge, 2014). Dennoch ist der Kompetenzbegriff mittlerweile aus der Musikpädagogik nicht mehr wegzudenken. Obwohl es bislang keine nationalen Bildungsstandards für das Fach Musik gibt, ist auch für Musik-Curricula eine eindeutige Orientierung am Konzept von Bildungsstandards bzw. an Kompetenzen festzustellen (Hasselhorn, 2015, Kap 3.2; Knigge & Lehmann-Wermser, 2008). In vielen Curricula lassen sich für das Schulfach Musik drei Kompetenzbereiche festmachen: (1) der Bereich *Rezeption*, der das Verstehen und Analysieren von Musik beinhaltet, (2) der Bereich *Produktion*, der das aktive Musizieren und Gestalten von Musik umfasst, und (3) *Reflexion*, das Beurteilen von Musik und das Herstellen von größeren Zusammenhängen (z. B. Hessisches Kultusministerium, 2016; Ministerium für Kultus, Jugend und Sport Baden-Württemberg, 2016; Ministerium für Schule und Bildung des Landes Nordrhein-Westfalen, 2019).[3] Folglich gibt es kompetenzorientierte musikdidaktische Konzepte und Konzeptionen wie etwa den „Aufbauenden Musikunterricht“ (Gies & Jank, 2015) sowie entsprechende Schulbücher (z. B. Detterbeck & Schmidt-Oberländer, 2015).

Bisher ist allerdings ein Desiderat an empirischer Forschung zu musikbezogenen Kompetenzen zu verzeichnen. Dies betrifft sowohl ihren Erwerb als auch ihre Struktur. Die meisten Kompetenzmodelle, die die Struktur musikbezogener Kompetenzen beschreiben – wie etwa die Kompetenzmodelle in den oben erwähnten Lehrplänen – haben theoretischen Status. Empirische Forschung zu Kompetenzmodellen im Fach Musik liegt bislang nur ausschnitthaft vor und wird im folgenden Kapitel thematisiert.

1.2. Kompetenzforschung in der Musikpädagogik

In den letzten Jahren gab es große Fortschritte in der Kompetenzmodellierung vieler schulischer Unterrichtsfächer (u. a. Leutner et al., 2017; Stanat et al., 2019)[4] und in der Modellierung professioneller Kompetenzen von Lehrkräften (u. a. Blömeke et al., 2010; Klemenz et al., 2019; König,

[3] Diese Darstellung beansprucht keine Vollständigkeit. Eine aktuelle umfassende Sichtung der Musiklehrpläne der einzelnen Bundesländer steht noch aus. Einige Bundesländer, wie z. B. Schleswig-Holstein oder Bayern beschreiben vier Kompetenzbereiche. In Bayern sind beispielsweise „Wahrnehmen und Erleben“ sowie „Analysieren und Einordnen“ zwei getrennte Kompetenzbereiche (Staatsinstitut für Schulqualität und Bildungsforschung München, 2021), während diese beiden Bereiche in Baden-Württemberg und Nordrhein-Westfahlen in einem Kompetenzbereich zusammengefasst sind (Ministerium für Kultus, Jugend und Sport Baden-Württemberg, 2016; Ministerium für Schule und Bildung des Landes Nordrhein-Westfalen, 2019). Niessen et al. (2008, S. 12-14) beschrieben bei einer umfassenden Analyse deutscher Schulcurricula ebenfalls drei Bereiche.

[4] Siehe Sälzer (2016, S. 39-40) und Stanat et al. (2019, S. 16) für einen Überblick über die in Deutschland regelmäßig stattfindenden Schulleistungsstudien.

2020; König et al., 2020). Empirische Forschung zu musikbezogenen Kompetenzen und Kompetenzerwerb ist allerdings nach wie vor rar. Es gibt zwar viele Studien, die in der sog. Expertiseforschung angesiedelt sind, jedoch liegt hier der Fokus meist auf überdurchschnittlichen Leistungen, die eher im außerschulischen Bereich erbracht werden (für einen Überblick s. Hasselhorn & Knigge, 2018). Für das Unterrichtsfach Musik gibt es national wie international bisher lediglich zwei Kompetenzmodelle, die empirisch erstellt und untersucht wurden. Die beiden Kompetenzmodelle wurden im Rahmen der durch die Deutsche Forschungsgemeinschaft (DFG) geförderten Projekte *KoMus* und *KOPRA-M* empirisch validiert.

Abbildung 1.
KoMus-Kompetenzmodell „Musik wahrnehmen und kontextualisieren“

Wahrnehmung und musikalisches Gedächtnis	**Anwendung musikalischen Wissens auf Basis der Wahrnehmung**		
Sound, Gestalt, Rhythmus, Melodie, Harmonie, Wirkung, Form, Tonhöhe, Lautstärke etc.	Kritische Bewertung von Musik und ihrer Ausführung		
	Verbalisierung/ Terminologie	Notation	historisch-kultureller Kontext
Dimension 1	D2	D3	D4

Anmerkung. Vereinfachte Darstellung aus Hasselhorn und Knigge (2018, S. 201); s. a. Jordan et al. (2012, S. 514) für eine ausführliche Darstellung.

Im Projekt *KoMus* (Jordan et al., 2012; Knigge, 2011) wurden ein Modell und ein entsprechender Test für den rezeptiven Kompetenzbereich „Musik wahrnehmen und kontextualisieren“ entwickelt. Der Kompetenztest richtete sich an Schüler*innen der sechsten Jahrgangsstufe und es konnte ein vierdimensionales Modell mit den Dimensionen „Wahrnehmung/musikalisches Gedächtnis“, „Verbalisierung/Fachterminologie“, „Notation“ und „historischer und kultureller Kontext“ empirisch bestätigt werden (Abbildung 1). Die einzelnen Dimensionen wurden in zwei bzw. drei Kompetenzniveaus unterteilt (Jordan et al., 2012, S. 514). Das Projekt *KOPRA-M* beschäftigte sich wiederum mit der Entwicklung und Validierung eines Modells zur Messung von musikpraktischen Kompetenzen (Hasselhorn, 2015). Hier konnten die Kompetenzdimensionen „Gesang“, „Instrumentales Musizieren“, und „Rhythmusproduktion“ bestätigt werden (u. a. Hasselhorn, 2015, S. 139).

Die Kompetenztests von *KoMus* und *KOPRA-M* wurden computerbasiert im Klassenverband durchgeführt und nutzten Kenntnisse und Methoden aus der Testtheorie bei der Konstruktion und Auswertung des Tests bzw. zur Beschreibung der Kompetenzmodelle. Ein Kompetenztest muss bestimmte statistische Anforderungen erfüllen, damit man mithilfe des Testergebnisses einer Person auf die theoretisch angenommene Kompetenz schließen kann. Nur wenn diese Anforderungen erfüllt sind, ist es zulässig, ein Kompetenzmodell aus empirischen Daten abzuleiten. Statistische Anforderungen dieser Art werden in der vorliegenden Arbeit in Kapitel 4 näher beschrieben. In besagtem Kapitel wird die sog. Item-Response-Theorie näher erläutert, die auch im Rahmen der beiden Musik-Forschungsprojekte *KoMus* und *KOPRA-M*, sowie in Schulleistungsstudien wie *PISA* oder *TIMMS* Anwendung findet. Die Methoden der Item-Response-Theorie sind zentral, um

Testergebnisse angemessen interpretieren zu können. Die beiden Projekte *KoMus* und *KOPRA-M* übertrugen diese statistischen Methoden erstmals auf Kompetenzmodellierung im Fach Musik. Für den Hochschulbereich legte Wolf (2016) ein empirisch validiertes Kompetenzmodell für analytisches Hören (Gehörbildung) vor.

In der Schulleistungsforschung wird häufig nicht nur die Leistung von Schüler*innen untersucht, sondern auch die Einflussfaktoren, die für das Zustandekommen der Leistung entscheidend sind, wie etwa Motivation oder soziale Herkunft (z. B. Stanat et al., 2019, Kap. 7-10). Für das Fach Musik gibt es bisher nur wenig Forschung darüber, welche externen Faktoren musikbezogene Kompetenzen beeinflussen. Erste Studien, die den Einfluss von außerschulischem Musikunterricht, Motivation, Geschlecht oder musikalischem Selbstkonzept untersuchen, liegen u. a. von Fiedler und Hasselhorn (2020), Harnischmacher und Knigge (2017), Hasselhorn und McElvany (2016) sowie Lill et al. (2019) vor. Heß (2018) sowie Fiedler und Hasselhorn (2020) zeigten, dass Mädchen häufig ein höheres musikalisches Selbstkonzept haben als Jungen. In den Studien, die musikbezogene Kompetenzen untersuchten, zeigte sich ein Zusammenhang zwischen besseren Ergebnissen im Kompetenztest und musikalischer Vorerfahrung, die beispielsweise durch Instrumentalunterricht bestand (Hasselhorn, 2015; Hasselhorn & McElvany, 2016; Jordan, 2014, S. 141-143).

Zusammenfassend kann festgehalten werden, dass in den letzten Jahren der Kompetenzbegriff auch in der Musikpädagogik auf curricularer und didaktischer Ebene zentral geworden ist. Empirische Forschung zu musikbezogenen Kompetenzen und Kompetenzmodellen ist jedoch vergleichsweise wenig vorhanden.

2. Musikbezogene Argumentationskompetenz

In diesem Kapitel wird der Gegenstand der vorliegenden Arbeit erörtert: musikbezogene Argumentationskompetenz. Im Rahmen eines kurzen Ausflugs in die Argumentationstheorie gehe ich zunächst darauf ein, was unter Argumentieren verstanden werden kann (Kapitel 2.1). Das Argumentieren in ästhetischen Fächern unterscheidet sich in vielerlei Hinsicht vom Argumentieren in anderen Disziplinen. Auf die wesentlichen Merkmale von ästhetischer Argumentation und ästhetischer Wahrnehmung gehe ich in Kapitel 2.2 ein. Kapitel 2.3 widmet sich Argumentationskompetenz und in Kapitel 2.4 stelle ich das theoretische Modell für musikbezogene Argumentationskompetenz vor, das der vorliegenden empirischen Arbeit zugrunde liegt. Abschließend widme ich mich in Kapitel 2.5 der Frage, wie man musikbezogene Argumentationskompetenz empirisch erforschen kann.

2.1. Argumentationstheorie

Die Argumentationstheorie kann auf eine lange Geschichte zurückblicken, die bis in die Antike zurückreicht und verschiedene Disziplinen berührt, darunter formale Logik, informelle Logik, Rhetorik und Dialektik.[5] Was dabei jeweils unter Argumentation verstanden wird, unterscheidet sich in den einzelnen Disziplinen teilweise erheblich.

In der formalen Logik geht es in erster Linie um die objektive Gültigkeit eines Arguments, also darum, ob ein Argument logisch korrekt schlussfolgert. Hierfür wird die Beziehung zwischen der Schlussfolgerung eines Arguments, die auch Konklusion genannt wird, und den Gründen, die die Schlussfolgerung stützen, untersucht. Die Gründe, die die Konklusion bzw. Schlussfolgerung stützen, werden Prämissen genannt. Ob ein Argument aus logischer Perspektive korrekt ist, hängt ausschließlich von der Beziehung zwischen den Prämissen und der Konklusion ab (Salmon, 1973, S. 12-13).

Dieses Verständnis von Argumentation wurde spätestens seit Mitte des zwanzigsten Jahrhunderts mit der pragmatischen Wende in der Argumentationstheorie erweitert. In der formalen Logik spielt der Kontext, in dem das Argument steht, keine Rolle. Ebenso wenig wird die Überzeugungskraft von Argumenten beachtet. In alltäglichen Situationen argumentiert man häufig genau dann, wenn die Konklusion fraglich erscheint (Tetens, 2004, S. 24). Aus logischer Perspektive bleiben verbalsprachliche, situative und pragmatische Faktoren unberücksichtigt, die in Kommunikations- und Interaktionsprozessen aber durchaus eine Rolle spielen (Eemeren et al., 2014, S. 149). Mit der

[5] Bei diesem Kapitel handelt es sich um eine ausschnitthafte Darstellung. Für einen Überblick über die verschiedenen Disziplinen innerhalb der Argumentationstheorie s. Eemeren et al. (2014).

pragmatischen Wende in der Argumentationstheorie gewannen andere theoretische Ansätze an Bedeutung, wie z. B. die informelle Logik (Johnson & Blair, 1994, 2000) oder die Pragma-Dialektik (Eemeren & Grootendorst, 2003). Diesen Ansätzen ist gemein, dass sie ein erweitertes Verständnis von Argumentation an den Tag legen. So gehen Eemeren et al. (2014, S. 3) in ihrer Definition vom alltagssprachlichen Gebrauch des Wortes „Argumentation“ aus und unterstreichen dabei das kommunikative und dialogische Moment von Argumentation. Laut den Autor*innen beschreibt das Wort „Argumentation“ in vielen europäischen Sprachen sowohl einen Prozess als auch ein Produkt.[6] Außerdem wird unter „Argumentation“ eine „konstruktive Bemühung“ verstanden, eine Meinungsverschiedenheit zu überwinden (Eemeren et al., 2014, S. 3).[7] Die Autor*innen definieren „Argumentation“ schließlich wie folgt:

> Argumentation is a communicative and interactional act complex aimed at resolving a difference of opinion with the addressee by putting forward a constellation of propositions the arguer can be held accountable for to make the standpoint at issue acceptable to a rational judge who judges reasonably.[8] (Eemeren et al., 2014, S. 7)

Innerhalb der Argumentationstheorie wird auch diskutiert, inwiefern es Maßstäbe für das Argumentieren gibt, die universell für verschiedene Fachbereiche wie Recht, Medizin oder Kunst gelten. Stephen E. Toulmin (2003) geht davon aus, dass die Gültigkeit eines Arguments von *bereichsspezifischen* und *bereichsunabhängigen* Aspekten abhängt. Während beispielsweise die Form eines Arguments bereichsunabhängig ist, können sich Gründe, die zur Stützung eines Arguments angeführt werden, je nach Fachbereich unterscheiden (Toulmin, 2003, S. 103). Beispielsweise wird sich eine Anwältin bei der Verteidigung einer Klientin auf Gesetzestexte berufen. Eine Restaurant- oder Konzertkritikerin hingegen wird sich kaum auf Gesetzestexte in ihrer Argumentation stützen können. Hier gelten andere Maßstäbe des Argumentierens. Kriterien, die zur Beurteilung der Gültigkeit einer Argumentation herangezogen werden, sind nach Toulmins Ansicht bereichsspezifisch (Toulmin, 2003, S. 104-105). Aber worauf bezieht sich eine Konzertkritikerin, wenn sie einen Artikel verfasst? Mit dieser und anderen Fragen beschäftigen sich ästhetische Argumentationstheorien.

[6] Zu diesen europäischen Sprachen zählen u. a. Deutsch, Französisch, Niederländisch, Italienisch und Portugiesisch (Eemeren et al., 2014, S. 3). Im Englischen meint das Wort „argumentation“ laut den Autor*innen in erster Linie einen Prozess (Eemeren et al., 2014, S. 4).

[7] Die Autor*innen unterscheiden hier zwischen dem Gebrauch des Wortes im Englischen und in anderen europäischen Sprachen. Im Englischen wird mit dem Wort „argument“ zudem Streit assoziiert (Eemeren et al., 2014, S. 4). Die Aussage „they had an argument last night“ bezeichnet einen Streit und nicht eine sachliche Auseinandersetzung.

[8] Je nach Theorietradition wird die Bedeutung des in der Definition erwähnten „rational judge who judges reasonably“ unterschiedlich ausgelegt. Mit „rational judge“ können eine oder mehrere reale Personen gemeint sein, die adressiert werden, oder auch eine abstrakte Instanz, die jeweils analytisch zu definieren ist (Eemeren et al., 2014, S. 40).

2.2. Ästhetische Argumentation und ästhetische Wahrnehmung

Auf Wikipedia gibt es einen Eintrag mit dem Titel „List of Music Considered the Worst“ (2021). Mehrere Personen haben Songs zusammengetragen, die in verschiedenen Umfragen und Kritiken zur „worst music ever made“ gekürt wurden. Die Einschätzungen verschiedenster Personen wurden hier also zusammengestellt, um zu argumentieren, dass beispielsweise „Don’t Worry, Be Happy“ von Bobby McFerrin zu den schlechtesten Songs aller Zeit gehört. Obwohl sich der Eintrag neben objektiven Einschätzungen auch auf den subjektiven Eindruck mehrerer Personen beruft, beansprucht er allgemeine Gültigkeit, nicht zuletzt, weil er auf der Online-Enzyklopädie Wikipedia erschienen ist. Dass ein Song wie „Don’t Worry Be Happy“ auf dieser Liste steht, erscheint zunächst verwunderlich, schließlich handelt es sich um einen sehr bekannten Song. Kann man von einem Song dieser Popularität wirklich behaupten, dass er zu den schlechtesten Songs aller Zeiten gehört? Vermutlich würden viele Menschen das Gegenteil behaupten. An diesem Beispiel lässt sich eine Besonderheit ästhetischer Argumentation zeigen: Ein ästhetisches Urteil erhebt Anspruch auf Allgemeingültigkeit, obwohl es auf der subjektiven Einschätzung einer Person basieren kann.

Ob und inwiefern ein ästhetisches Urteil allgemeine Geltung beanspruchen kann, wird in der Philosophie mindestens seit dem Ende des 17. Jahrhunderts diskutiert (Kurbacher-Schönborn, 2007).[9] Wenn Personen derart unterschiedlich über einen Song urteilen, liegt die Schlussfolgerung nahe, dass ein ästhetischer Gegenstand nicht über eine bestimmte ästhetische Eigenschaft verfügt, sondern dass diese Eigenschaft dem Gegenstand bloß zugeschrieben wird. Laut Immanuel Kant sind Geschmacksurteile zwar subjektiv, doch wer ein ästhetisches Urteil fällt, spricht so, als sei eine ästhetische Eigenschaft „eine Beschaffenheit des Gegenstandes und das Urteil logisch“ (Kant, 1790, § 1, 6). Übertragen auf das Eingangsbeispiel bedeutet dies, dass die Autor*innen des Wikipedia-Artikels so schreiben, als sei „schlechter Song“ per se eine Eigenschaft des Liedes „Don’t Worry Be Happy“, obwohl sie diese Eigenschaft dem Lied lediglich zugeschrieben haben. Wer ästhetisch urteilt, erhebt trotz der Subjektivität des Urteils „Anspruch auf Gültigkeit für jedermann“ (Kant, 1790, § 6), was Kant auch als „subjektive Allgemeingültigkeit“ bezeichnet (Kant, 1790, § 8). Allerdings kann man die Zustimmung zum Urteil nur „ansinne[n]“ (Kant, 1790, § 8). Wenn man ein Lied wie „Don’t Worry Be Happy“ mag, wird man sich kaum durch Überredung vom Gegenteil überzeugen lassen.

Wir können nur ein Urteil über einen ästhetischen Gegenstand fällen, wenn wir diesen zuvor wahrgenommen haben (z. B. Sibley, 1965, S. 137). Die Bewertung eines ästhetischen Objekts, wie z. B. eines Songs, passiert demnach in der Wahrnehmung des Objekts und nicht durch eine logische Operation (Kleimann, 2005, S. 112). Christian Rolle bezieht sich auf Kant und plädiert dafür,

[9] Bei diesem Kapitel handelt es sich um ausschnitthafte Darstellungen ästhetischer Theorien. Für einen ausführlichen Überblick s. Reicher (2005).

ästhetische Urteile als Empfehlungen zu verstehen (Rolle, 1998, S. 31). Ästhetische Urteile sind demnach eine Empfehlung für eine andere Person, den Gegenstand so wahrzunehmen, wie man selbst es tut. Wenn Werturteile geäußert werden – etwa, dass „Don't Worry, Be Happy" einer der schlechtesten Songs aller Zeiten ist – versuchen wir unsere*n Gesprächspartner*in dazu anzuregen, den Song so wahrzunehmen, wie wir es tun. Eine Empfehlung für genau diese Wahrnehmung des Songs wird im eingangs erwähnen Wikipedia-Artikel ausgesprochen. Die Autor*innen empfehlen im Artikel, eine ganze Liste an Songs so wahrzunehmen, wie sie es tun, nämlich als die schlechtesten Songs aller Zeiten. Der Geltungsanspruch des ästhetischen Urteils ist laut Rolle erst dann eingelöst, wenn sich die ästhetische Wahrnehmung „als lohnenswert oder als nicht lohnenswert für andere herausgestellt hat" (Rolle, 1998, S. 32). Obwohl ästhetische Werturteile mit einem Anspruch auf Allgemeingültigkeit formuliert werden, kann man sich nicht auf verbindliche Kriterien berufen, diesen Anspruch zu rechtfertigen (Kleimann, 2005, S. 107).

Wie bisher dargestellt, spielt ästhetische Wahrnehmung eine zentrale Rolle im ästhetischen Urteil. Unsere Wahrnehmung kann sich dabei laut Ursula Brandstätter sowohl auf die äußere Wirklichkeit als auch auf unsere inneren Befindlichkeiten beziehen (Brandstätter, 2008, S. 99). In Bezug auf ästhetische Wahrnehmung unterscheidet Martin Seel (1991) drei verschiedene Formen ästhetischer Wahrnehmung: korresponsive, kontemplative und imaginative Wahrnehmung. *Korresponsive* Wahrnehmung ist mit dem menschlichen Bedürfnis verbunden, die Lebensumgebung sinnhaft zu gestalten (Seel, 1993, S. 34): Man hört Musik, die zur aktuellen Lebenssituation, zu den aktuellen Bedürfnissen passt. Mit *kontemplativer* Wahrnehmung bezeichnet Seel einen ästhetischen Wahrnehmungsmodus, bei dem die wahrnehmende Person *nicht* daran interessiert ist, dem ästhetischen Objekt Sinn zuzuschreiben, und die Musik im Moment erlebt. Es geht also um eine „sinn*abstinente* Aufmerksamkeit" (Seel, 1993, S. 35). Zudem identifiziert Seel einen Modus der *imaginativen* Wahrnehmung, der neue Sichtweisen auf die Wirklichkeit ermöglicht (Seel, 1993, S. 37). Seel betont, dass diese Wahrnehmungsformen seltenen isoliert vorkommen.

Es wird auch diskutiert, ob ästhetische Wahrnehmung ins Verbalsprachliche übersetzt werden kann (Brandstätter, 2008, S. 37). Christopher Wallbaum etwa wirft die Frage auf, ob man über kontemplative Erfahrungen überhaupt sprechen könne. Wer kontemplativ wahrnehme, betrachte einen Gegenstand im Moment und schreibe ihm keinen Sinn zu. Durch die kommunikative Bezugnahme, die in der Beschreibung kontemplativer Wahrnehmung geschieht, werde jedoch Sinn gestiftet (Wallbaum, 2009, S. 215). Die Reichhaltigkeit ästhetischer Wahrnehmung scheint sich nicht in sprachliche Begriffe übersetzen zu lassen. Brandstätter (2008, S. 107) spricht in diesem Zusammenhang auch von einer „Diskrepanz zwischen sinnlicher Wahrnehmung und begrifflicher Verarbeitung".

Trotz dieser Diskrepanz kommunizieren wir häufig über ästhetische Objekte. Martin Seel und Bernd Kleimann beschreiben verschiedene Formen ästhetischer Argumentation bzw. Kommunikation. Seel skizziert dabei drei Formen ästhetischer Argumentation und berücksichtigt dabei die Beziehung der urteilenden Person zum ästhetischen Gegenstand (Seel, 1985, S. 237-244). Im *Kommentar* wird der Gegenstand beschrieben bzw. werden ihm ästhetische Eigenschaften zugeschrieben. Eine ästhetische Bewertung des Gegenstandes ist dort nicht vorgesehen (Seel, 1985, S. 243-

244). Bei der *Konfrontation* werden Angaben über die subjektive Wirkung eines ästhetischen Objekts gemacht (Seel, 1985, S. 244). Im Gegensatz zum Kommentar sind konfrontative Einschätzungen „primär emotional und als solche gar nicht zu begründen“ (Seel, 1985, S. 245). Auch in der Konfrontation steht die ästhetische Qualität eines Objekts nicht im Mittelpunkt; es wird nur artikuliert, „*daß* es sich um ein [...] Objekt ästhetischen Interesses handelt“ (Seel, 1985, S. 245). In der *Kritik* werden schließlich Konfrontation und Kommentar miteinander verbunden (Seel, 1985, S. 237, 256).[10] Kleimann (2005) wiederum erweitert Seels Unterscheidung und schlägt eine kleinteiligere Systematik von sechs ästhetischen Kommunikationsformen vor: *spontane Wertung*, *Beschreibung*, *Kommentar*, *Charakterisierung*, *Interpretation* sowie *reflektierte Wertung*.

Wenn man über ästhetische Objekte kommuniziert, versucht man also, die eigene ästhetische Erfahrung verständlich zu machen und das Gegenüber davon zu überzeugen, den ästhetischen Gegenstand so wahrzunehmen, wie man selbst es tut. Die Bedeutung dieser kommunikativen Auseinandersetzung ist laut Rolle und Wallbaum für ästhetische Bildungsprozesse bedeutsam (u. a. Rolle, 1999, 2014a; Rolle & Wallbaum, 2011; Wallbaum, 2009). In argumentativen Auseinandersetzungen über Musik, die die Autoren auch „ästhetischen Streit“ nennen, geht es darum, ästhetische Wahrnehmung zu kommunizieren (Rolle & Wallbaum, 2011, S. 509). Im ästhetischen Streit werden verschiedene Hör- oder Spielweisen von Musik empfohlen, die schließlich intersubjektiv verhandelt werden. In diesem Geschehen werden musikalische Bildungsprozesse möglich (Rolle, 2014a, S. 1). Laut Kleimann (1998, S. 71) ist ein argumentativer Austausch über Musik allerdings nur möglich, wenn die Argumentierenden ähnliche Hörgewohnheiten haben und daher gewisse musikalische Grunderfahrungen teilen. Musikbezogene Urteile können nur dann gerechtfertigt werden, wenn die Gründe, die beim Argumentieren herangezogen werden, für alle Seiten nachvollziehbar sind (Kleimann, 1998, S. 74). Daher wird entsprechend betont, dass Musikstücke stets im Kontext einer bestimmten Musikkultur oder Musikpraxis stehen. Rolle und Wallbaum verweisen in diesem Zusammenhang auf Kleimann: Wenn eine Musikkultur bzw. -praxis nicht geteilt wird, kann eine argumentierende Person versuchen, andere Personen dazu zu bewegen, „Wahrnehmungsschemata so umzubilden, daß eine Verständigung und argumentative Auseinandersetzung über Musik allererst möglich wird“ (Kleimann, 1998, S. 74).

Zusammenfassend kann Folgendes festgehalten werden: Wer ästhetisch argumentiert, hat zuvor einen ästhetischen Gegenstand wahrgenommen. Das ästhetische Urteil ist daher ein wesentliches Resultat der ästhetischen Wahrnehmung. Das ästhetische Urteil bezieht sich also immer sowohl auf die subjektiven Sinneseindrücke als auch auf das ästhetische Objekt. Auch wenn es keine verbindlichen Kriterien für die Gültigkeit eines ästhetischen Arguments gibt, beansprucht es dennoch allgemeine Gültigkeit. Doch auf welche Kriterien berufen sich Personen, wenn sie ästhetisch urteilen und welche Kompetenzen werden beim ästhetischen Argumentieren benötigt? Diese Fragen versucht Rolles (2013) theoretisches Modell in Bezug auf das Fach Musik zu klären. Bevor ich ausführlicher auf dieses musikspezifische Kompetenzmodell in Kapitel 2.4 eingehe, erläutere ich im folgenden Kapitel 2.3 zunächst die Bedeutung von Argumentationskompetenz im schulischen

[10] Wallbaum (2009, S. 218-227) überträgt Seels Unterscheidungen auf den musikdidaktischen Kontext.

Kontext.

2.3. Argumentationskompetenz

Argumentieren spielt als sprachliche Schlüsselkompetenz eine wichtige Rolle für das schulische Lernen und ist fächerübergreifend in Schulcurricula und Bildungsstandards relevant (Budke & Meyer, 2015; Morek et al., 2017; Vollmer & Thürmann, 2010). Im Fachunterricht werden Lerninhalte durch Sprache vermittelt und Sprache dient als Kommunikationsmittel in Unterrichtssituationen (Michalak et al., 2015, S. 5). Sprachliche Fähigkeiten sind somit fächerübergreifend notwendig für die Teilnahme und Teilhabe an Unterrichtsgesprächen und entscheidend für den Bildungserfolg von Schüler*innen (Morek & Heller, 2012; Quasthoff, 2009; Quasthoff et al., 2020b; Schmölzer-Eibinger, 2013). In bildungssprachlichen Handlungen wie dem Argumentieren wird Wissen konstruiert und diskursiv verhandelt (Morek, 2016, S. 125). Insbesondere diskursive sprachliche Handlungen wie Erklären oder Argumentieren sind in schulischen Lernkontexten für die Konstruktion, Vermittlung und Überprüfung von Wissen bedeutsam (Quasthoff et al., 2020a, S. 20). Fachliches und sprachliches Lernen bedingen sich demnach gegenseitig. Wenn eine Person komplexe inhaltliche Zusammenhänge sprachlich darstellen kann, dann zeigt sie, dass sie sich die dafür notwendigen Inhalte kognitiv und kommunikativ angeeignet hat (Morek & Heller, 2012, S. 75).

Sprachliche Fähigkeiten hängen stark mit der sozialen Herkunft von Schüler*innen zusammen (Stanat et al., 2017). Bildungsbenachteiligung entsteht hier durch eine fehlende Passung zwischen schulischen Anforderungen und herkunftsbedingten sprachlich-diskursiven Praktiken (Quasthoff et al., 2020a, S. 14). Die Fähigkeit, etwas zu erklären oder zu begründen, wird in den einzelnen Schulfächern häufig einfach vorausgesetzt und im Unterricht selbst nicht geschult (Schmölzer-Eibinger, 2013, S. 27). Vor diesem Hintergrund gab es in den letzten Jahren eine Vielzahl an empirischen Studien, die Argumentationskompetenz oder verwandte Kompetenzen in verschiedenen Schulfächern untersucht haben (u. a. Chng et al., 2014; Domenech et al., 2017; Erath et al., 2018; Frederking et al., 2012; Gronostay, 2019; Krelle, 2014; Krelle & Willenberg, 2008; Morek, 2016; Neumann & Lehmann, 2008; Prediger & Hein, 2017; Quasthoff & Domenech, 2016; Quasthoff et al., 2017; Quasthoff et al., 2020b).

Je nach Schulfach unterscheiden sich die fachspezifischen Besonderheiten beim Argumentieren. Während sich Schüler*innen im Mathematikunterricht Argumentation beispielsweise in Form von Berechnungen oder mathematischen Beweisen aneignen, werden im Deutschunterricht Zeitungsartikel analysiert und im Politikunterricht ethische Perspektiven diskutiert (s. a. Budke & Meyer, 2015, S. 14; Jahnke & Ufer, 2015; Manzel & Weißeno, 2017). Alexandra Budke definiert Argumentationskompetenz in Anlehnung an Weinerts Kompetenzdefinition (s. a. Kapitel 1.1) dahingehend, dass Schüler*innen

> über Fähigkeiten und Fertigkeiten verfügen, mündliche und schriftliche Argumentationen in verschiedenen fachlichen Kontexten zu verstehen, eigene Argumentationen

> zu produzieren und in der Interaktion mit anderen auf Argumentationen angemessen zu reagieren, sowie auch, dass sie die damit verbundenen motivationalen, volitionalen und sozialen Bereitschaften aufweisen, diese Argumentationsfähigkeiten in variablen Situationen erfolgreich und verantwortungsvoll zu nutzen. (Budke, 2013, S. 360)[11]

In der Sprachwissenschaft wird betont, dass ein wesentlicher Anteil an Argumentationskompetenz darin besteht, „übersatzmäßige – globale – Strukturen kohärent" aufzubauen (Quasthoff & Domenech, 2016, S. 22). Quasthoff et al. definieren in diesem sprachwissenschaftlichen Zusammenhang Argumentationskompetenz als „gattungsspezifische Fähigkeit, in Begründungsdiskursen äußerungsübergreifende Einheiten im Mündlichen und Schriftlichen kontextuell angemessen verstehen, platzieren, sequenziell aufbauen und sprachlich markieren zu können" (Quasthoff et al., 2020b, S. 86).[12]

Wegen der Fülle an verschiedenen Forschungstraditionen im Zusammenhang mit Argumentation bezeichnen Chrysi Rapanta et al. Argumentationskompetenz als „ill-defined concept" (Rapanta et al., 2013, S. 493). Die Autor*innen sichteten in ihrem Review-Artikel 97 empirische Studien aus dem erziehungswissenschaftlichen Kontext, um das Begriffsverständnis von Argumentationskompetenz zu analysieren. Zusammenfassend schlagen die Autor*innen eine dreiteilige Konzeptualisierung argumentativer Kompetenz vor. Danach schließt Argumentationskompetenz (1) Wissensbereiche ein, die dem deklarativen Wissen zuzuordnen sind („know what skills"), (2) Wissensbereiche, die dem prozeduralen Wissen angehören („know-how skills") und (3) Wissen darüber, wie Wissen zustande kommt („know-be skills") (Rapanta et al., 2013, S. 491-492).[13]

Aus dieser knappen Darstellung wird deutlich, dass es sich bei Argumentationskompetenz um eine facettenreiche Kompetenz handelt. Je nach Fach unterscheiden sich die Anforderungen beim Argumentieren. Für das musikbezogene Argumentieren hat Rolle ein theoretisches Kompetenzmodell entwickelt, das die Anforderungen beim musikbezogenen Argumentieren beschreibt.

2.4. Theoretisches Modell für musikbezogene Argumentationskompetenz

Wie in den vorangegangenen Kapiteln dargestellt, unterscheiden sich die Anforderungen beim Argumentieren je nach Fach. Auch Toulmin (2003) spricht von *bereichsspezifischen* und *bereichsunabhängigen* Aspekten, die beim Argumentieren eine Rolle spielen (s. a. Kapitel 2.1). In Kapitel 2.2

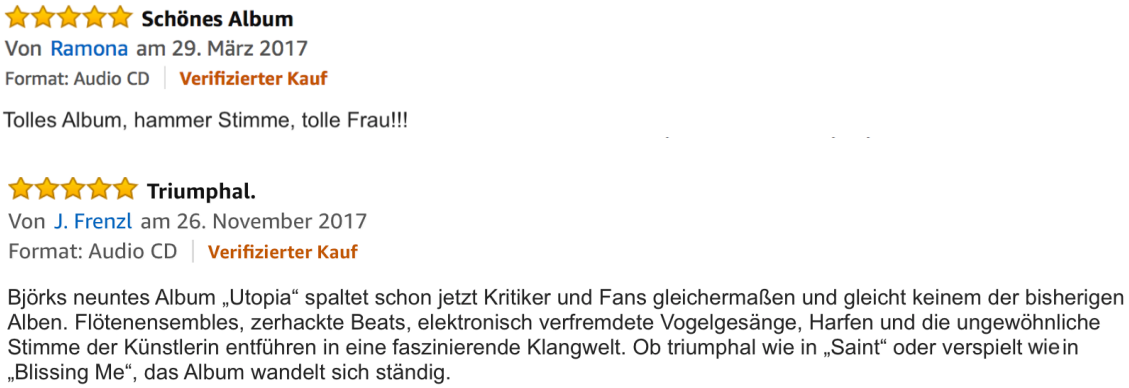
[11] Budke (2013) schlägt diese Definition in Hinblick auf den Geografieunterricht vor. In Budke und Meyer (2015, S. 14) wird dieselbe Definition für den allgemeinen Schulkontext vorgeschlagen.

[12] Quasthoff sieht Argumentationskompetenz als eine spezifische Ausprägung von Diskurskompetenz und schlägt ein Modell vor, das Diskurskompetenz gattungs- und medialitätsübergreifend beschreibt (Quasthoff, 2009; Quasthoff & Domenech, 2016).

[13] Rapanta et al. (2013, S. 511) nennen diese drei Teilbereiche auch metakognitive, metastrategische und epistemologische Kompetenzen. Diese Begrifflichkeiten gehen auf Deanna Kuhn (1999) zurück, die in Bezug auf „Critical Thinking" zwischen metakognitivem, metastrategischem und epistemologischem Wissen unterscheidet.

wurde erläutert, dass sich ästhetisches Argumentieren vom logischen Argumentieren unterscheidet. Während es beim logischen Argumentieren um logische Beziehungen zwischen Prämissen und Konklusion geht, sind ästhetische Argumente häufig auch subjektiv, obwohl sie in vielen Fällen allgemeine Gültigkeit beanspruchen. Dieser Grad an Subjektivität ist darauf zurückzuführen, dass ästhetische Urteile immer auch das Resultat ästhetischer Wahrnehmung sind. Demzufolge gibt es keine verbindlichen Kriterien für die Gültigkeit ästhetischer Argumente. Aber wodurch zeichnet sich die Überzeugungskraft von ästhetischen Argumenten aus, wenn sie in einem logischen Sinne nicht objektiv richtig oder falsch sein können? Obwohl es keine verbindlichen Kriterien für die Gültigkeit von ästhetischen Argumenten gibt, können sie unterschiedlich überzeugend sein. Dies sollen die folgenden beiden Kaufempfehlungen für eine CD bei einem Online-Händler verdeutlichen (Abbildung 2):

Abbildung 2.
Kaufempfehlungen für eine CD



Während die erste Kaufempfehlung ein Geschmacksurteil ist („Tolles Album“), bezieht sich die zweite Kaufempfehlung auf Merkmale der Musik (z. B. „zerhackte Beats, elektronisch verfremdete Vogelgesänge“) und auf den Ausdruck der Musik (z. B. „verspielt“). Außerdem wird auf die Meinung von „Kritikern und Fans“ verwiesen. Beide Kaufempfehlungen werben dafür, die CD so wahrzunehmen, wie die Autor*innen. Allerdings versucht die zweite Kaufempfehlung, die Hörweise des*der Autor*in nachvollziehbar zu machen. Dies geschieht durch die differenzierte Beschreibung des Albums.

Rolle stellt fest, dass es beim ästhetischen musikbezogenen Argumentieren darum geht, andere Personen für neue Hörweisen zu gewinnen, indem man dafür wirbt, die Musik, die zur Debatte steht, mit anderen Ohren zu hören (Rolle, 2017, S. 131). Dabei fallen laut Rolle objektive Gründe nicht ins Gewicht:

> Eine Musik ist nicht gut, weil sie diese oder jene Eigenschaften hat; sie ist nicht gelungen, weil sie so oder so gemacht ist, sondern weil sie mich in dieser oder jener Hinsicht anspricht. Wie sie mich anspricht, gilt es nachvollziehbar zu machen, wenn

> ich andere von meinem Urteil überzeugen möchte. Dazu muss ich auf Eigenschaften der Musik Bezug nehmen und sie in der Art und Weise, in der ich sie wahrnehme, beschreiben. (Rolle, 2017, S. 131)

Rolles Auffassung nach zeigt sich musikalisch-ästhetische Kompetenz mitunter darin, wie verständlich eine Person über Musik sprechen kann und ob ihre Gründe nachvollziehbar sind, wenn ästhetische Urteile begründet werden (Rolle, 2008, S. 79). Er definiert musikbezogene Argumentationskompetenz als die Kompetenz, „verständlich, plausibel und differenziert ästhetische Werturteile über Musikstücke begründen zu können“ (Knörzer et al., 2015, S. 148).

Rolle schlägt ein Kompetenzmodell vor, das die Struktur musikbezogener Argumentationskompetenz in sieben Niveaustufen beschreibt (Rolle, 2013, 2017) (Abbildung 3).[14] Er nimmt einerseits an, dass sich musikbezogene Urteile auf das Objekt beziehen, also z. B. auf die Qualität einer Komposition oder auf die Angemessenheit einer musikalischen Interpretation (Rolle, 2017, S. 132). Ein musikbezogenes Urteil bezieht sich aber auch auf die subjektive Empfindung, die durch die Musik hervorgerufen wird. Nicht zuletzt versteht Rolle Argumentieren als ein dialogisches Geschehen und berücksichtigt in seinem Modell, inwiefern sich Personen beim Argumentieren mit Meinungen anderer Personen bzw. anderen Hörweisen auseinandersetzen. Auf dem niedrigsten Niveau 1 beziehen sich Personen ausschließlich auf das persönliche Gefallen bzw. Missfallen. Ein Beispiel hierfür ist die erste Kaufempfehlung aus Abbildung 2: „Tolles Album, hammer Stimme, tolle Frau“. Personen auf Niveau 3 können sich auf „objektive Eigenschaften der Musik“ beziehen und auf Niveau 4 Bezug „auf den eigenen Eindruck des Ausdrucks der Musik“ nehmen (Rolle, 2017, S. 141). Die zweite Kaufempfehlung aus Abbildung 2 kann als ein Beispiel für eine Argumentation auf einem höheren Niveau gelten. Auf den Niveaus 5 bis 7 können Personen schrittweise auch andere Hörweisen in ihre Argumentation einbeziehen und musikkulturelle Konventionen bzw. Besonderheiten berücksichtigen. Das Modell wird im Zusammenhang mit dem Testdesign des *MARKO*-Tests in Kapitel 5 noch ausführlicher diskutiert.

Es gibt erste Arbeiten, die Rolles Modell empirisch operationalisierten (Knörzer et al., 2015; Knörzer et al., 2016). Knörzer et al. (2016) baten in einer qualitativen Studie Versuchspersonen darum, zwei Versionen zweier Musikstücke aus dem Pop- und Klassik-Bereich miteinander zu vergleichen. Die Versuchsteilnehmenden hatten 60 Minuten Zeit, in einer offenen Stellungnahme zu begründen, welche der beiden Versionen ihnen besser gefiel. Die Stichprobe wurde nach Expertise der Versuchsteilnehmenden in drei Gruppen (Noviz*innen, Semi-Expert*innen und Expert*innen) eingeteilt und es konnten deutliche Unterschiede zwischen den drei Gruppen in Bezug auf ihre Argumentationsweisen festgestellt werden (Knörzer et al., 2016, S. 13). Personen mit höherer Expertise bezogen sich in ihrer Argumentation häufiger auf kontextspezifisches Hintergrundwissen und Merkmale des Stückes, während subjektive Aspekte im Gegensatz zu den Noviz*innen eine geringere Rolle spielten. Außer den Arbeiten von Knörzer et al. gibt es nur wenig empirische Forschungsarbeiten zum musikbezogenen Argumentieren. Gottschalk und Lehmann-Wermser (2013)

[14] Rolle bezieht sich in seinem Kompetenzmodell u. a. auf Arbeiten von Michael J. Parsons (1987) aus dem Bereich der Bildenden Kunst und auf das „Reflective Judgment Model“ von Patricia M. King und Karen S. Kitchener (1994, 2004).

untersuchten in einer Design-Based Research-Studie die musikalisch-ästhetische Diskursfähigkeit von Schüler*innen der neunten Klasse. Außerdem gibt es einige laufende Dissertationsvorhaben, die sich Argumentations- oder Diskursfähigkeiten von Schüler*innen widmen (Gottschalk, i. Vorb.; Haberecht, i. Vorb.; Meyer, i. Vorb.). In Bezug auf musikbezogene Argumentationskompetenz besteht allerdings derzeit ein eindeutiges Desiderat an empirischen Forschungsarbeiten. Das folgende Kapitel widmet sich deshalb der Frage, wie man musikbezogene Argumentationskompetenz empirisch untersuchen kann. Dabei werden methodische Herangehensweisen empirischer Forschungsarbeiten aus unterschiedlichen Fachbereichen reflektiert.

Abbildung 3.
Theoretisches Modell musikbezogener Argumentationskompetenz (Rolle, 2017, S. 141) (englische Version s. Rolle, 2013, S. 146)

Ebene	Die Urteilenden können...
7	*Ebene des ästhetischen Diskurses* ... musikbezogene Urteile begründen in Reflexion unterschiedlicher ästhetischer Konventionen, Hörweisen und musikkultureller Praxen; dabei unterschiedliche Perspektiven einnehmen und Kritik anderer in die eigene Perspektive einbeziehen.
6	*Ebene ästhetischer Urteile* ... musikbezogene Urteile begründen unter Verweis auf formale und expressive Eigenschaften der Musik sowie stilistische und musikkulturelle Besonderheiten, um die eigene Sicht- bzw. Hörweise nachvollziehbar zu machen und zu anderen Perspektiven ins Verhältnis zu setzen.
5	*Ebene konventioneller Urteile* ... musikbezogene Urteile begründen unter Bezugnahme sowohl auf musikalische Parameter wie subjektive Eindrücke und unter Berufung auf kulturspezifische, häufig technisch-handwerkliche Kriterien. [... to raise counterarguments against judgments of the opponent (without criticising the presented arguments and justifications) (Rolle, 2013, S.146)]
4	*Subjektivistische Ebene* ... musikbezogene Urteile begründen unter der Bezugnahme auf den eigenen Eindruck des Ausdrucks der Musik; Begründungen, die andere anführen, sind deren Begründungen, die die eigene Interpretation und Einschätzung nicht in Frage stellen können.
3	*Objektivistisch-geschmacksrelativistische Ebene* ... musikbezogene Urteile begründen unter Bezugnahme auf objektive Eigenschaften der Musik, ohne dass daran Zweifel aufkommen könnte; Differenzen sind ein Zeichen unterschiedlichen Geschmacks, über den man nicht streiten kann.
2	*Autoritätsbezogene Ebene* ... musikbezogene Urteile äußern und auf Nachfrage unter Verweis auf Autoritäten bzw. Kenntnisse aus zweiter Hand begründen; unterschiedliche Meinungen bedeuten keinen Dissens; Gründe, die andere anführen, werden nicht als Begründungen wahrgenommen.
1	*Ebene unmittelbarer Präferenzen* ... Musik wahrnehmen und Gefallen bzw. Missfallen bekunden; das Urteil ist Teil der musikalischen Praxis; andere Einschätzungen werden kaum wahrgenommen; eine Begründung ist nicht nötig.

Zunehmend differenzierte Fähigkeit, Musik wahrzunehmen und zu beschreiben.
Zunehmende Kenntnis von Musik unterschiedlicher Stilistiken und Kulturen.
Zunehmende Fähigkeit, sich in komplexen Argumentationszusammenhängen zu bewegen.

2.5. Wie lässt sich musikbezogene Argumentationskompetenz empirisch untersuchen? Publikation A[15]

Dieser Abschnitt fasst Publikation A zusammen (Ehninger, 2021). In der Publikation wurden verschiedene methodische Herangehensweisen zur empirischen Erforschung von musikbezogener Argumentationskompetenz reflektiert.[16] In der Musikpädagogik gibt es bisher kaum empirische Arbeiten zu musikbezogener Argumentationskompetenz, weshalb Forschungsarbeiten aus verschiedenen Fächern analysiert wurden, die Argumentationskompetenz oder eine verwandte Kompetenz empirisch untersuchen (Erath et al., 2018; Frederking et al., 2011a; Frederking et al., 2011b; Frederking et al., 2012; Gottschalk & Lehmann-Wermser, 2013; Krelle & Willenberg, 2008; Morek, 2016; Neumann, 2007; Neumann & Lehmann, 2008; Prediger et al., 2016; Willenberg et al., 2007). Anschließend an eine kritische Reflexion der jeweiligen Arbeiten wurden in der Publikation Perspektiven für die musikpädagogische empirische Forschung aufgezeigt, um zu verdeutlichen, wie innerhalb der Musikpädagogik fachspezifische, aber auch fächerübergreifende Aspekte von Argumentationskompetenz untersucht werden können. Dabei wurde deutlich, dass Argumentationskompetenz eine facettenreiche Kompetenz ist und mit einem Forschungsdesign jeweils nur Teilaspekte der Kompetenz empirisch untersucht werden können.

Die Studien, die im Rahmen der Publikation analysiert wurden, untersuchen das Verstehen von Argumenten (*Rezeption*), das schriftliche Hervorbringen von Argumenten (*Produktion*) und/oder das interaktive Geschehen, das beim Argumentieren stattfindet (*Interaktion*). Argumentation ist immer auch abhängig von der Medialität (s. a. Ehninger, 2021, S. 19): In einem formalen Beschwerdebrief wird beispielsweise anders argumentiert als in einer Gruppendiskussion im Klassenzimmer. Eine Übersicht über die untersuchten Forschungsarbeiten findet sich in Tabelle 1. Das übergeordnete Erkenntnisinteresse der untersuchten Studien lag auf unterschiedlichen Schwerpunkten. Dazu zählten neben den bereits erwähnten Fokussen auf Produktion, Rezeption und Interaktion der Erwerb der Fähigkeit zu argumentieren.

In den untersuchten Studien wurde die argumentative Leistung einer oder mehrerer Personen beobachtet. Morek et al. (2017, S. 13) sprechen in diesem Zusammenhang auch von *sprecherzentrierten* und *interaktionsorientierten* Perspektiven. Bei sprecherzentrierten Perspektiven liegt der Fokus auf einer Person. Hier wird entweder das Produkt, das eine Person hervorbringt, untersucht, oder die Art und Weise, wie die Person Argumente rezipiert. Für die Leistungserfassung einer Person wurden in den untersuchten Studien in erster Linie schriftliche Erhebungsverfahren verwendet. Dazu zählen die Kompetenztests zu literarästhetischer Urteils- bzw. Verstehenskompetenz sowie die Kompetenztests zu Argumentation und Schreiben im Fach Deutsch (Frederking et al., 2012; Frederking et al., 2011a; Frederking et al., 2011b; Krelle & Willenberg, 2008; Neumann,

[15] Publikation A: Ehninger, J. (2021). Wie lässt sich musikbezogene Argumentationskompetenz empirisch untersuchen? Über die empirische Erforschung einer facettenreichen Kompetenz. *Beiträge empirischer Musikpädagogik*, *12*. https://www.b-em.info/index.php/ojs/article/view/192.

[16] Zur besseren Lesbarkeit verzichte ich weitestgehend auf die Kennzeichnung von indirekten Zitaten.

Tabelle 1.
Ausgewählte empirische Studien über Argumentationskompetenz bzw. verwandte Konstrukte

Kategorie	Studien	Untersuchtes Konstrukt	Auswertungsmethode
Produktion	Knörzer et al. (2015, 2016)	Musikbezogenes Argumentieren	Inhaltsanalyse
	Neumann & Lehmann (2008); Neumann (2007)	Schreibkompetenz im Fach Deutsch	Kompetenzmodellierung, Probabilistische Testtheorie
Rezeption	Frederking et al. (2011a, 2011b, 2012)	Literarästhetische Urteils- bzw. Verstehenskompetenz	Kompetenzmodellierung, Probabilistische Testtheorie
	Krelle & Willenberg (2008); Willenberg, Gailberger & Krelle (2007)	Argumentation im Fach Deutsch	Kompetenzmodellierung, Probabilistische Testtheorie
Interaktion	Morek (2016)	Komplexität von Diskursanforderungen in Unterrichtsgesprächen	Gesprächsanalyse
	Erath et al. (2018); Prediger et al. (2016)	Erklären in Unterrichtsgesprächen	Gesprächsanalyse, epistemische Matrix
	Gottschalk & Lehmann-Wermser (2013)	Musikbezogene Diskursfähigkeit	Analyseschema nach Grundler (2011)

Anmerkung. Die Tabelle ist auch in Ehninger (2021, S. 13) abgedruckt.

2007; Neumann & Lehmann, 2008; Willenberg et al., 2007). Geschlossene Aufgabenformate mit Multiple- oder Forced-Choice-Items wurden v. a. eingesetzt, um die Rezeption von Argumenten zu untersuchen. Offene Aufgabenformate, wie das Schreiben eines Beschwerdebriefs, eignen sich besonders, um die Produktion von Argumenten zu erfassen. Pohl (2014, S. 288) spricht in diesem Zusammenhang von der Gefahr, einem „monologischen Reduktionismus“ zu verfallen, da das dialogische Moment von Argumentation in diesen Erhebungsettings bestenfalls angedeutet werden kann. In mündlichen Erhebungssettings kann man nur begrenzt auf die argumentative Leistung einer Person schließen, da hier mehrere Personen miteinander agieren. Jedoch können in mündlichen Settings kollektive Argumentationsprozesse sowie Interaktion untersucht werden (Erath et al., 2018; Gottschalk & Lehmann-Wermser, 2013; Morek, 2016; Prediger et al., 2016).

Die Vor- und Nachteile der jeweiligen methodischen Zugriffe werden in Ehninger (2021, S. 20-24) ausführlich diskutiert und es werden Perspektiven aufgezeigt, die sich daraus für die empirische Erforschung musikbezogener Argumentationskompetenz ergeben. Hierbei war die Unterscheidung von sowohl *fachspezifischen* als auch von *fächerübergreifenden* Aspekten zentral. Beim Argumentieren ist beispielsweise fächerübergreifend die Fähigkeit erforderlich, die Struktur eines Arguments zu verdeutlichen und das Argument adäquat im Gesprächs- oder Textverlauf zu kontextualisieren. Fachspezifische und fächerübergreifende Aspekte können jedoch nicht immer eindeutig auseinandergehalten werden. Beim Argumentieren spielen ebenfalls *fachsprachli-*

che Fähigkeiten eine Rolle. Fachsprache besteht dabei nicht nur aus „fachspezifischen Begriffen", sondern bedient sich an „sprachliche[n] Handlungsmuster[n], Wortgruppen, Formulierungen und Abkürzungen, die für ein bestimmtes Fach charakteristisch sind" (Bossen, 2019, S. 65).[17]

Die folgende Liste gibt eine Übersicht über die *fachspezifischen* und *fächerübergreifenden* Aspekte beim Argumentieren, die in den diskutierten Studien untersucht wurden und wagt eine Übertragung der fachspezifischen Aspekte auf den musikpädagogischen Kontext.[18] Die Darstellung erhebt keinen Anspruch auf Vollständigkeit.

Fachspezifische (= fachlich-inhaltliche) Aspekte beim musikbezogenen Argumentieren

- Merkmale des Musikstückes,
- Subjektive Aspekte (die den eigenen Eindruck der Musik betreffen),
- Kontextspezifisches Hintergrundwissen,
- Medienbezogene Aspekte,
- Reflexion musikbezogener Wahrnehmung,
- Verwendung von Fachterminologie,
- ...

Fächerübergreifende Aspekte (nicht nur, aber auch) beim musikbezogenen Argumentieren

- Kommunikative und situative Anforderungen,
- Sprachliche Realisierung (mündlich/schriftlich),
- Struktur eines Arguments,
- Perspektivwechsel,
- Sprachliche Anforderungen auf Wort-, Satz- und Textebene (z. B. pragmatisch, semantisch, syntaktisch, lexikalisch, sprachsystematisch),
- ...

In der Publikation wurde schließlich diskutiert, dass sich je nach Erkenntnisinteresse unterschiedliche Forschungsmethoden eignen, um die verschiedenen fachspezifischen bzw. fächerübergreifenden Aspekte zu untersuchen, die beim musikbezogenen Argumentieren eine Rolle spielen. Dabei wurden verschiedene Perspektiven für die musikpädagogische Forschung aufgezeigt (Ehninger, 2021, S. 20-24).

[17] Bossen stellt in Bezug auf Fachsprache im Fach Musik ein Theoriedefizit fest und trägt wesentliche Merkmale von Fachsprache im Fach Musik zusammen (Bossen, 2019, S. 64-99).

[18] Die nachfolgende Aufzählung ist ebenfalls in Ehninger (2021, S. 19) abgedruckt.

3. Forschungsziele

Beim Argumentieren handelt es sich um eine kognitiv wie sprachlich anspruchsvolle Handlung. Dies trifft auch auf das Argumentieren im Fach Musik zu. In Kapitel 2.5 habe ich dargestellt, dass es sich bei musikbezogener Argumentationskompetenz um eine facettenreiche Kompetenz handelt, bei der sowohl fachspezifische als auch fächerübergreifende Aspekte eine Rolle spielen. Es wird unmöglich sein, alle Aspekte, die beim musikbezogenen Argumentieren von Bedeutung sind, mit einer Forschungsmethode in den Blick zu bekommen. Neben dem Desiderat an empirischer Forschung zum musikbezogenen Argumentieren gibt es nach wie vor nur wenig Arbeiten, die musikbezogene Kompetenzen modellieren (s. a. Kapitel 1.2). Diese Dissertation will einen Beitrag im Bereich der Modellierung musikbezogener Kompetenzen leisten und widmet sich außerdem der empirischen Forschung musikbezogener Argumentationskompetenz.

Die Studie stützt sich auf theoretische Arbeiten von Rolle (2013, 2017), der ein theoretisches Kompetenzmodell für musikbezogenes Argumentieren entwickelte. Es wird darin theoretisch angenommen, dass sich Personen in ihrem Vermögen musikbezogen zu argumentieren unterscheiden. Es fehlt jedoch bisher an empirischen Arbeiten, die Aufschluss darüber geben, wie diese Kompetenz strukturiert ist. Auf welche Aspekte nehmen Personen Bezug, wenn sie musikbezogenen Argumentieren? Wo liegen die Herausforderungen beim musikbezogenen Argumentieren? Um diesem Desiderat zu begegnen, wurde in der vorliegenden Arbeit ein Kompetenztest für musikbezogenes Argumentieren entwickelt. Mit diesem Kompetenztest wurden Daten erhoben, die dazu beitragen sollen, dass argumentative Kompetenz genauer beschrieben werden kann.

Die zentralen Ziele der Dissertation lassen sich wie folgt zusammenfassen:

(a) Theoriegeleitete Entwicklung eines Kompetenztests für musikbezogenes Argumentieren,

(b) Kompetenzmessung und Beschreibung von Kompetenzniveaus für musikbezogenes Argumentieren auf Basis der empirischen Daten, die im Kompetenztest erhoben wurden.

Der Testentwicklung (Ziel (a)) lagen theoretische Annahmen aus Rolles Kompetenzmodell zugrunde (Rolle, 2013, 2017; s. a. Kapitel 2.4). Der entwickelte Kompetenztest sollte wesentliche Gütekriterien der quantitativen Forschung erfüllen. Musikbezogene Argumentationskompetenz sollte demnach objektiv, zuverlässig und valide gemessen werden (zu Haupt- und Nebengütekriterien in der Testkonstruktion s. a. Bühner, 2021, Kap. 8). Die Messung von musikbezogener Argumentationskompetenz sowie die Beschreibung der Kompetenz in Form von Kompetenzniveaus ist Gegenstand von Ziel (b).

Quantitative Methoden, die auch im Rahmen zahlreicher Schulleistungsstudien wie u. a. *PISA, DESI, VERA* zum Einsatz kamen, eignen sich besonders, um den genannten Zielen nachzugehen. Die Methoden, die mitunter in solchen Schulleistungsstudien Anwendung finden, sind u. a. der *Item-Response-Theorie* zuzuordnen. Mit diesem methodischen Forschungsansatz können Daten aus Kompetenztests ausgewertet werden und so kann empirisch dargelegt werden, wie eine Kom-

petenz strukturiert ist.

In den folgenden Kapiteln widme ich mich zunächst den Methoden, die in der vorliegenden Arbeit verwendet wurden (Kapitel 4) und gehe auf wesentliche Annahmen der Item-Response-Theorie sowie auf die oben erwähnten Gütekriterien ein. Anschließend stelle ich in Kapitel 5 das Testdesign und die Pilotstudie vor. In Kapitel 6 wird es um die Hauptstudie gehen. Hier wird die Kompetenz modelliert und es werden Kompetenzniveaubeschreibungen aus den empirischen Daten abgeleitet. Außerdem werden verschiedene Einflussfaktoren auf die Ausprägung von musikbezogener Argumentationskompetenz dargestellt. Schließlich werden die Ergebnisse der Dissertation in Kapitel 7 diskutiert.

II.

Empirischer Teil

4. Methoden

Das methodische Vorgehen der vorliegenden Arbeit orientierte sich an der *Item-Response-Theorie*. Die Item-Response-Theorie (IRT), auch Probabilistische Testtheorie genannt, kam in zahlreichen Schulleistungsstudien zum Einsatz (z. B. *DESI*, *PISA*, *TIMMS*). Mithilfe von IRT können Fähigkeitsausprägungen von Personen quantitativ dargestellt werden. Beispielsweise kann man so mit Daten eines Kompetenztests die zugrundeliegende Kompetenz quantitativ erfassen. Um musikbezogene Argumentationskompetenz mit IRT in den Blick zu bekommen, wurde in der vorliegenden Arbeit ein Test für musikbezogenes Argumentieren entwickelt. Bei der Entwicklung eines Kompetenztests ist es zunächst wichtig, dass der Messgegenstand (hier: musikbezogene Argumentationskompetenz) auf Basis einer Theorie genau definiert wird (Bühner, 2021, S.14). In Kapitel 4.1 gehe ich zunächst auf Grundannahmen der Item-Response-Theorie ein, um dann in Kapitel 4.2 verschiedene Analyseverfahren näher zu beschreiben, die in der vorgelegten Studie zum Einsatz kamen.

4.1. Item-Response-Theorie (IRT)

In der Item-Response-Theorie (IRT) geht man davon aus, dass man die Eigenschaft bzw. Fähigkeit einer Person, die für das Ergebnis eines Tests verantwortlich ist, nicht direkt beobachten kann. Genau diese Fähigkeit bzw. Eigenschaft ist jedoch für das Testverhalten der Person verantwortlich. Wenn ein*e Schüler*in in einer Musik-Klassenarbeit 20 von 25 möglichen Punkten erzielt, sind diese 20 Punkte ein Indikator für die zugrundeliegende Fähigkeit der Schülerin bzw. des Schülers im Schulfach Musik. Das Ergebnis der Klassenarbeit (20 Punkte) kann man also direkt beobachten. Eine solche direkt beobachtbare Variable bezeichnet man auch *manifeste Variable*. Die eigentliche Fähigkeit einer Schülerin bzw. eines Schülers im Schulfach Musik kann man jedoch nicht direkt, sondern nur indirekt über das Ergebnis der Klassenarbeit beobachten. Eine Variable wie diese nennt man *latente Variable*. Die latente Variable (Fähigkeit im Schulfach Musik) wird daher über eine manifeste Variable (Ergebnis der Klassenarbeit) beschrieben. Das beobachtete Testverhalten einer Person ist also lediglich ein Indikator für die zugrundeliegende Fähigkeit (Moosbrugger et al., 2020, S. 252). Dieser Zusammenhang zwischen den beobachteten Testwerten einer Person und der zugrundeliegenden Fähigkeit (des latenten Merkmals) wird in der IRT durch Wahrscheinlichkeitsfunktionen beschrieben.

In der IRT wird angenommen, dass die Wahrscheinlichkeit, eine Aufgabe in einem Test richtig zu lösen einerseits von der Schwierigkeit der Aufgabe abhängt und andererseits von der Fähigkeit der Person, die die Aufgabe löst. Die Schwierigkeit der Aufgabe wird durch den Itemschwierigkeitsparameter σ beschrieben und die Fähigkeit der Person durch den Personenfähigkeitsparameter θ. So kann für jede Person und jede Aufgabe in einem Test die Wahrscheinlichkeit bestimmt werden, mit der diese Person diese Aufgabe löst (Moosbrugger et al., 2020, S. 262). Dieser Zusammenhang

wird mithilfe einer logistischen Wahrscheinlichkeitsfunktion dargestellt. Dabei spielt die Differenz der Personenfähigkeit θ und der Itemschwierigkeit σ eine besondere Rolle ($\theta - \sigma$). Je ‚besser' eine Person im Gegensatz zum Item ist, desto höher ist die Wahrscheinlichkeit, das Item zu lösen. Dieser Zusammenhang ist im dichotomen Rasch-Modell (1PL) von zentraler Bedeutung. Das dichotome Rasch-Modell wurde für Items entwickelt, die zwei Antwortmöglichkeiten haben, z. B. ‚richtig' oder ‚falsch'. Die Antwort auf das entsprechende Item kann dann beispielsweise mit 1 für ‚richtig' und 0 für ‚falsch' kodiert werden. Das statistische Modell wird in der Literatur häufig durch die untenstehende Modellgleichung beschrieben (Trendtel et al., 2016a, S. 190; s. a. Bühner, 2021, S. 254–257 für die Herleitung der Gleichung des Rasch-Modells). $P(X_i = 1)$ beschreibt die Wahrscheinlichkeit einer Person v, ein Item i zu lösen:

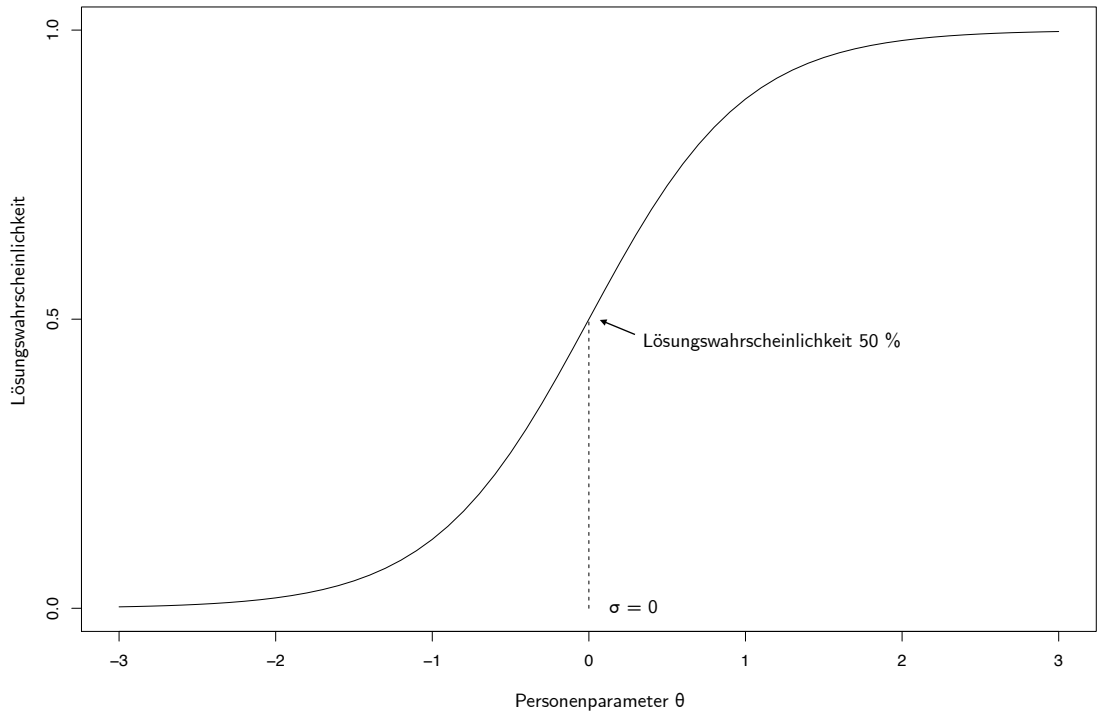
$$P(X_i = 1|\theta_v) = \frac{exp(\theta_v - \sigma_i)}{1 + exp(\theta_v - \sigma_i)}$$

In dieser Gleichung kann man erkennen, dass die Lösungswahrscheinlichkeit ausschließlich von der Fähigkeit der Person (Personenfähigkeitsparameter θ) und der Schwierigkeit des Items abhängt (Itemschwierigkeitsparameter σ). Die Differenz dieser beiden Parameter ist, wie bereits erwähnt, die entscheidende Größe für die Lösungswahrscheinlichkeit. Somit sind beide Parameter differenzskaliert und lassen sich auf einer gemeinsamen Skala abbilden (Bühner, 2021, S. 262). Diese Skala wird auch *Logit-Skala* genannt, wobei deren Wertebereich theoretisch von $-\infty$ bis $+\infty$ gehen kann, de facto jedoch meist zwischen -3 und $+3$ liegt. Negative Werte stehen für Personen mit geringen Fähigkeiten bzw. leichte Items. Positive Werte beschreiben Personen mit hohen Fähigkeiten bzw. schwere Items (Bühner, 2021, S. 252). Die Lösungswahrscheinlichkeit für ein Item in Abhängigkeit von der Personenfähigkeit und der Itemschwierigkeit kann auch in der sog. *Itemcharakteristischen Funktion* (IC-Funktion) dargestellt werden (Abbildung 4). Auf der y-Achse ist die Lösungswahrscheinlichkeit dargestellt (der Wertebereich geht von 0 bis 1), auf der x-Achse ist der Personenfähigkeitsparameter abgetragen. Abbildung 4 zeigt eine IC-Funktion für ein Item mit der Schwierigkeit $\sigma = 0$. Für eine Person mit der Fähigkeit $\theta = 0$ besteht demnach eine 50 %ige Wahrscheinlichkeit, das Item richtig zu lösen.

Im Rasch-Modell wird konventionell die Itemschwierigkeit für ein bestimmtes Item am Wendepunkt der IC-Funktion bestimmt. Dort beträgt die Lösungswahrscheinlichkeit 50 % (Abbildung 4). Auf der x-Achse ist also nicht nur die Personenfähigkeit, sondern auch die Itemschwierigkeit abgebildet (Koller et al., 2012, S. 13). Das Antwortverhalten einer Person darf nur von dieser einen Fähigkeit abhängen. Diesen Sachverhalt bezeichnet man auch als *Itemhomogenität*. Außerdem darf das Lösen bzw. Nichtlösen einer Aufgabe nicht von einer vorangegangenen Aufgabe beeinflusst werden (*lokale stochastische Unabhängigkeit*). Aufgrund der Annahme dieser lokal stochastischen Unabhängigkeit können bei der Berechnung der gemeinsamen Lösungswahrscheinlichkeit die Einzelwahrscheinlichkeiten miteinander multipliziert werden (Moosbrugger et al., 2020, S. 261).

Das dichotome Raschmodell stellt das einfachste IRT-Modell für nur zwei Antwortkategorien (z. B. ‚richtig' und ‚falsch') dar. Es gibt jedoch Items, bei denen auch Punkte für eine teilweise richtige Lösung vergeben werden. So könnte eine Aufgabe ‚falsch', ‚teilweise' oder ‚richtig'

Abbildung 4.
Beispiel einer IC-Funktion eines dichotomen Items mit dem Schwierigkeitsparameter $\sigma = 0$



Anmerkung. Die gestrichelte Linie zeigt, dass bei einer Personenfähigkeit von $\theta = 0$ die Lösungswahrscheinlichkeit 0.5 (bzw. 50 %) beträgt.

gelöst werden. Hier liegt ein *polytomes* Item vor, das z. B. mit 0 für ‚falsch', 1 für ‚teilweise' und 2 für ‚richtig' kodiert wird. Für diesen Fall wird das *Partial-Credit-Modell* (PCM) bzw. ordinale Rasch-Modell verwendet, das eine Erweiterung des dichotomen Rasch-Modells ist. Im PCM wird angenommen, dass eine höhere Kategorie (z. B. 2 für ‚richtig') schwieriger zu erreichen ist als eine niedrigere (z. B. 1 für ‚teilweise richtig'). Für das PCM gilt die folgende Modellgleichung (Trendtel et al., 2016a, S.194, Formel modifiziert):

$$P(X_i = x_i|\theta_v) = \frac{exp(\sum_{h=0}^{x_i}(\theta_v - \sigma_{ih}))}{\sum_{k=0}^{H_i} exp(\sum_{h=0}^{k}(\theta_v - \sigma_{ih}))}$$

Auch aus dieser Modellgleichung geht die zentrale Bedeutung der Differenz der Personenfähigkeit θ und der Itemschwierigkeit σ hervor. Das Item i hat die Kategorien 0 bis H_i. σ_{ih} ist der Itemschwierigkeitsparameter, der bei Item i für die Kategorie h beobachtet wird. θ ist wieder der Personenfähigkeitsparameter einer Person v. In der vorliegenden Arbeit kamen vorwiegend Items mit mehr als zwei Antwortkategorien zum Einsatz, weshalb das PCM angewandt wurde.

In einigen Publikationen wird die IRT der *Klassischen Testtheorie* (KTT) gegenübergestellt. Bei der KTT handelt es sich um eine Messfehlertheorie. Dieser liegt die Annahme zugrunde, dass

Messungen fehlerbehaftet sind. Ein wesentliches Ziel der KTT ist daher, den wahren Wert einer Messung vom Messfehler zu trennen (Moosbrugger et al., 2020, S. 254). Während die KTT vorwiegend für Items mit kontinuierlichem Antwortformat entwickelt wurde und das Verhältnis zwischen Itemvariable und latentem Konstrukt als lineare Beziehung darstellt, widmet sich die IRT der Modellierung dichotomer (oder polytomer) Antwortkategorien und der Schätzung latenter Personen- und Itemparameter durch meist logistische Zusammenhänge (Moosbrugger et al., 2020, S. 252-253). Entgegen früheren Darstellungen sollten KTT und IRT „nicht als konkurrierende Ansätze verstanden werden, sondern als zwei Methoden, die sich gegenseitig vorteilhaft ergänzen“ (Moosbrugger et al., 2020, S. 271). Beide Theorien weisen enge Beziehungen auf (für einen Überblick s. Moosbrugger et al., 2020, S. 268–271).

4.2. Analyseverfahren zur Modellprüfung

Aus der oben vorgestellten Modellgleichung können bestimmte Erwartungen an die Testitems in einem Datensatz abgeleitet werden (Koller et al., 2012, S. 61). Werden diese Erwartungen nicht erfüllt, dann kann ein IRT-Modell nicht für einen bestimmten Datensatz gelten. Es gibt verschiedene Kriterien, anhand derer man beurteilen kann, ob die erhobenen Daten zum IRT-Modell passen. Man kann hierbei die Passung einzelner Items mithilfe verschiedener statistischer Analysen untersuchen oder auch den globalen Modellfit überprüfen (Trendtel et al., 2016a, S. 208). In diesem Unterkapitel werden verschiedene Verfahren zur Modellprüfung beschrieben, die in der vorliegenden Arbeit angewandt wurden. Im Wesentlichen geht es bei der Modellprüfung darum herauszufinden, ob ein bestimmtes Modell die Daten gut beschreiben kann. Die Analyseverfahren prüfen also, inwiefern die erhobenen Daten vom IRT-Modell abweichen. Dabei gibt es nicht die „eine“ Analyse, die stimmen muss, damit man von der Passung eines IRT-Modells ausgehen kann. Vielmehr werden verschiedene Analyseverfahren abgewogen.

Im vorangegangenen Unterkapitel wurde bereits die *lokale stochastische Unabhängigkeit* erwähnt, die in IRT-Modellen gelten muss. Die Wahrscheinlichkeit ein Item zu lösen, darf nur von den beiden erwähnten Parametern (Personenfähigkeit und Itemschwierigkeit) abhängen. Außerdem müssen die Items in einem Test auf *eine* dahinterliegende latente Variable schließen lassen. Es wäre beispielsweise denkbar, dass ein Testitem versehentlich nicht musikbezogene Argumentationskompetenz, sondern Lesekompetenz misst. In diesem Fall wäre die *Itemhomogenität* nicht gegeben. Itemhomogenität bedeutet, dass alle Items des Tests, bzw. einer Testhälfte, dieselbe Fähigkeit messen wie der gesamte Test. Für jede beliebige Teilstichprobe müssen also die Itemschwierigkeitsparameter gleich ausfallen. Erst wenn das gewährleistet ist, wird für alle Personen in der jeweiligen Stichprobe dieselbe Fähigkeit oder Eigenschaft gemessen (Bühner, 2021, S. 272). Das bezeichnet man auch als *Personenhomogenität*. Es könnte sein, dass Personen in der Stichprobe verschiedene Strategien benutzen, um die Testitems zu lösen (Bühner, 2021, S. 257). In diesem Fall wäre die Personenhomogenität verletzt. Es sollten also nur die Items für die Berechnung des IRT-Modells ausgewählt werden, die zur Messung eines Konstrukts psychometrisch am besten geeignet sind (Kelava & Moosbrugger, 2020a, S. 155). Die damit verbundenen Analysen und das

„Aussortieren“ nicht geeigneter Items bezeichnet man *Itemselektion*.

In den nächsten Abschnitten werden verschiedene Analyseverfahren vorgestellt, die zur Überprüfung der Modellpassung in der vorliegenden Arbeit zum Einsatz kamen. In diesen Analyseverfahren spielen Personen- und Itemhomogenität eine Rolle (s. Bühner, 2021, S. 258 für eine Übersicht).

Andersen-Test, grafische Modellkontrolle und Wald-Test

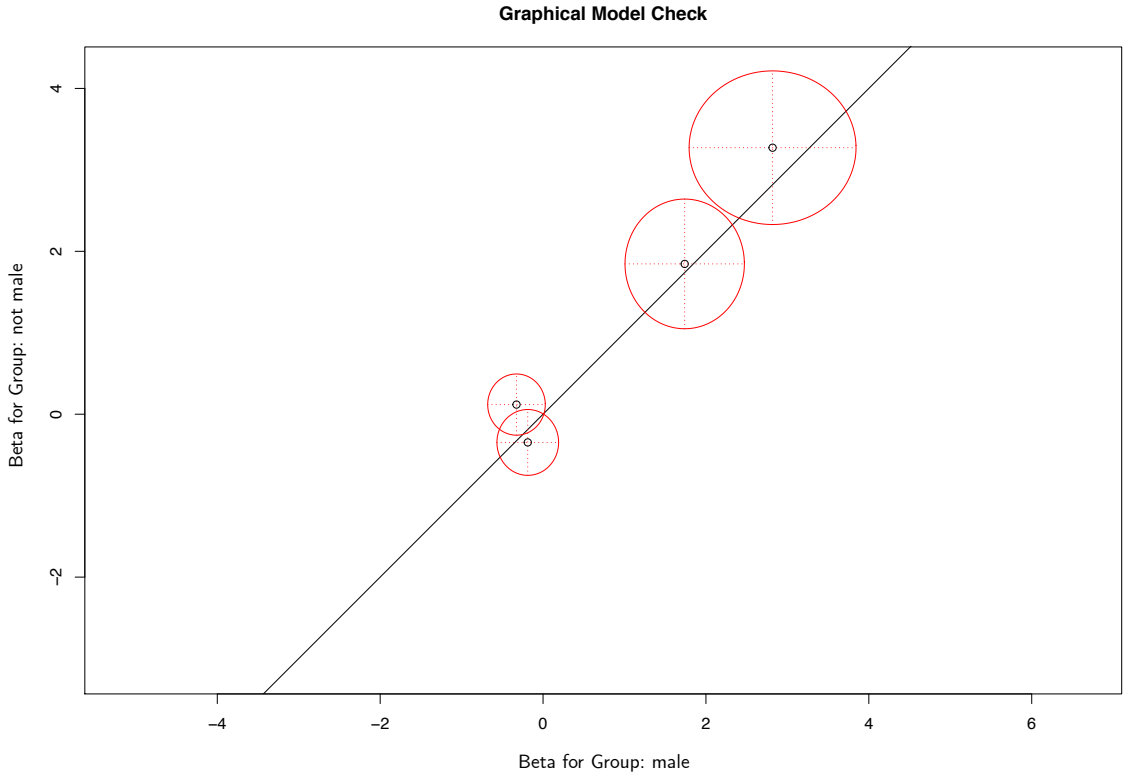
Wie bereits erläutert gibt es verschiedene Testverfahren, die prüfen, inwieweit die erhobenen Daten vom IRT-Modell abweichen. Für viele dieser Testverfahren sind allerdings die Voraussetzungen seitens der Stichprobe häufig nicht gegeben (Bühner, 2021, S. 277). Um zu überprüfen, ob die Daten der vorliegenden Studie vom IRT-Modell abweichen, wurde u. a. der *Andersen-Test* (bedingter Likelihood-Quotienten-Test) durchgeführt. Der Andersen-Test überprüft die Personenhomogenität der Stichprobe. Hierfür wird die Stichprobe in mindestens zwei Teilstichproben eingeteilt. Die Teilung erfolgt anhand eines Teilungskriteriums (z. B. dem Geschlecht, dem Median der latenten Variable oder anhand einer Zufallsvariable). Im Andersen-Test wird nun überprüft, ob die Itemschwierigkeitsparameter in den Teilstichproben gleich sind (Bühner, 2021, S. 265). Wenn man die Stichprobe in zwei Gruppen teilt, halbiert sich allerdings entsprechend die Anzahl der Personen und so kann es passieren, dass bestimmte Antwortkategorien in den Teilstichproben nicht besetzt sind, was zu Problemen bei der Schätzung führen kann (Bühner, 2021, S. 267). Wenn der Andersen-Test signifikant wird und so die Personenhomogenität innerhalb der Stichprobe nicht gegeben ist, kann mithilfe der *grafischen Modellkontrolle* überprüft werden, welche Items dazu führen, dass der Andersen-Test signifikant wird (Bühner, 2021, S. 277). Bei der grafischen Modellkontrolle wird ebenfalls analysiert, ob ein Test, dessen Stichprobe in zwei Teilstichproben geteilt wird, dieselbe Fähigkeit innerhalb der Teilstichproben misst. Hierbei kommt allerdings, im Gegensatz zum Andersen-Test, kein inferenzstatistischer Hypothesentest zum Einsatz. Die grafische Modellkontrolle ist somit zwar anschaulich, liegt jedoch im subjektiven Ermessen.

In Abbildung 5 ist eine solche grafische Modellkontrolle dargestellt. Die gesamte Stichprobe wurde in zwei Teilstichproben geteilt. In der einen Gruppe sind nur die männlichen, in der anderen die nicht-männlichen Versuchspersonen. An der x-Achse sind die Itemschwierigkeitsparameter für die männlichen Versuchspersonen aufgetragen („Beta for Group: male“); an der y-Achse die der nicht-männlichen Versuchspersonen. Jeder schwarze Punkt in Abbildung 5 entspricht einem Item, dessen Koordinaten die geschätzten Schwierigkeitsparameter des Items in den beiden Gruppen sind. Ein Item mit identischen Itemschwierigkeitsparametern in den beiden Subgruppen würde auf der Geraden liegen. Die Ellipsen zeigen die Konfidenzintervalle der geschätzten Parameter an. Wenn die Ellipse die Gerade schneidet, kann angenommen werden, dass sich die Itemschwierigkeiten in den beiden Gruppen nicht unterscheiden (Koller et al., 2012, S. 81). Zu große Konfidenzintervalle können auch dafür verantwortlich sein, dass die Itemschwierigkeitsparameter zwischen den Teilstichproben unterschiedlich sind (Bühner, 2021, S. 272).

Zusätzlich zur grafischen Modellkontrolle und zum Andersen-Test kam in der vorgelegten Studie auch der *Wald-Test* zum Einsatz. Mit diesem Test kann man auf Itemebene die Gleichheit der

Itemschwierigkeitsparameter von zwei Teilstichproben messen (Kelava & Moosbrugger, 2020b, S. 397). Hier werden also die Unterschiede zwischen den Gruppen für einzelne Items überprüft (Koller et al., 2012, S. 62). Beim Andersen-Test hingegen handelt es sich um einen globalen Test, bei dem alle Items simultan geprüft werden (Koller et al., 2012, S. 67).

Abbildung 5.
Beispiel einer grafischen Modellkontrolle



Anmerkung. An der x-Achse sind die Itemschwierigkeitsparameter für die männlichen Versuchspersonen aufgetragen („Beta for Group: male"); an der y-Achse die der nicht-männlichen Versuchspersonen („Beta for Group: not male"). Jeder schwarze Punkt steht für ein Item. Wenn ein Punkt auf der Geraden liegen würde, wären die Itemschwierigkeitsparameter in beiden Gruppen gleich. Die roten Ellipsen um die Items stehen für die Konfidenzintervalle.

Zur Überprüfung der Modellgeltung kann auch ein parametrisches *Bootstrap-Verfahren* angewandt werden. Dieses Verfahren hat sich jedoch als wenig teststark erwiesen, weshalb Modellverletzungen häufig nicht erkannt werden (Bühner, 2021, S. 263). Wenn mehrere Modelle zur Auswahl stehen, um die empirischen Daten zu beschreiben, kann man anhand *informationstheoretischer Maße* überprüfen, welches Modell am besten zu den Daten passt. Informationstheoretische Maße (z. B. AIC und BIC) geben an, welches Modell hinsichtlich Passung und Sparsamkeit zu bevorzugen ist (Bühner, 2021, S. 278). Die Wahl fällt dann auf das Modell mit den niedrigsten informationstheoretischen Maßen. Allerdings überprüfen informationstheoretische Maße die Modellpassung selbst nicht.

Itemfit

Itemfitwerte drücken aus, ob ein bestimmtes Item zu den Annahmen des IRT-Modells passt. Hierfür wird bestimmt, inwieweit die beobachteten empirischen Werte von den theoretisch erwarteten Werten abweichen. Die Abweichung zwischen beobachteten und erwarteten Werten bezeichnet man *Residuum*. Die in dieser Studie verwendeten Itemfit-Maße *Infit* (weighted mean square statistic) und *Outfit* (unweighted mean square statistic) basieren auf der Berechnung von Residuen und beschreiben somit, inwieweit die empirischen Daten die theoretischen Annahmen des IRT-Modells erfüllen. Infit- und Outfit-Werte gewichten die Residuen unterschiedlich. Der Outfit ist sensibler gegenüber unerwarteten Beobachtungen bei sehr leichten oder sehr schweren Items. Die Infitwerte geben hingegen an, ob Personen auf Items ungewöhnlich antworten, die ihrer Fähigkeit entsprechen (Bühner, 2021, S. 274). Niedrige Infit- und Outfit-Werte weisen darauf hin, dass es weniger Variabilität gibt, als vom Modell vorhergesagt wird, hohe Werte zeigen das Gegenteil (Freunberger et al., 2016, S. 238).

Sowohl Infit-, als auch Outfit-Werte haben einen Erwartungswert von 1. Ein Wert von 1 bedeutet, dass der theoretisch erwartete Wert dem empirisch beobachteten Wert entspricht. Beispielsweise bedeutet somit ein Wert von 1.30, dass die beobachteten Daten 30 % mehr variieren, als im theoretischen Modell angenommen (Bond & Fox, 2015, S. 269). Es gibt verschiedene Empfehlungen, wie weit ein Itemfit-Maß von 1 abweichen darf, um den Anforderungen des IRT-Modells zu genügen. Bond und Fox (2015, S. 270) empfehlen Werte zwischen 0.75 und 1.3, Ames und Penfield (2015, S. 45) Werte zwischen 0.5 und 1.5. Der Outfit ist dabei sensibler gegenüber unerwarteten Beobachtungen bei sehr leichten oder schweren Items (Bond & Fox, 2015, S. 270). In Large-Scale Assessment Studien wird Outfit-Werten daher kaum Bedeutung beigemessen und es werden meist nur die Infit-Werte berichtet (Trendtel et al., 2016a, S. 209).

Differential Item Functioning (DIF)

DIF-Analysen untersuchen die Testfairness eines Messinstruments. Um das Prinzip zu verdeutlichen, nehmen wir an, dass Mädchen und Jungen mit denselben Fähigkeiten einen Musik-Test bearbeiten. In einem Item geht es um Videospiel-Musik und es fällt auf, dass Jungen bessere Ergebnisse bei der Aufgabe erzielen als Mädchen. Da bereits bekannt ist, dass Mädchen und Jungen dieselben Fähigkeiten haben, liegt eine mögliche Erklärung für das unterschiedliche Abschneiden darin, dass Jungen besser mit dem Kontext von Videospiel-Musik vertraut sind. Das Testitem benachteiligt also systematisch Mädchen. Man sagt deshalb auch, dass dieses Item ‚DIF hat'. Ein Test sollte jedoch gleich fair für alle Testpersonen ein.

Im IRT-Modell haben Personen mit derselben Personenfähigkeit θ dieselbe Wahrscheinlichkeit, ein bestimmtes Item zu lösen. DIF tritt dann auf, wenn zwei Gruppen mit derselben Personenfähigkeit θ verschiedene Wahrscheinlichkeiten haben, ein Item zu lösen. Es gibt unterschiedliche Möglichkeiten, DIF zu bestimmen (s. ausführlich bei Wu et al., 2016, Kap. 11). Eine Möglichkeit bei dichotomen Items ist, die Itemschwierigkeiten getrennt voneinander für die verschiedenen Gruppen zu bestimmen und diese anschließend zu vergleichen. In der vorliegenden Studie wurde

ein anderes Vorgehen gewählt, da die meisten Items mehrkategoriale Antwortformate enthielten. Es wurde deshalb ein sog. *Facetten-Modell* berechnet. Hier wird zusätzlich zum Itemschwierigkeitsparameter σ und zum Personenparameter θ ein ‚item-by-group' Interaktionsterm in die Gleichung aufgenommen, die in Kapitel 4.1 vorgestellt wurde. Die Wahrscheinlichkeit, ein Item zu lösen, hängt also nicht nur vom Personenfähigkeitsparameter θ und dem Itemschwierigkeitsparameter σ ab, sondern auch vom Interaktionsterm. Der Interaktionsterm passt die Itemschwierigkeit aufgrund der Zugehörigkeit einer Person zu einer bestimmten Gruppe an (Wu et al., 2016, S. 216). Bei den DIF-Analysen richtete sich die vorliegende Studie nach den Richtlinien des Educational Testing Service (ETS). Ein vernachlässigbarer DIF liegt bei einer Effektstärke von ≤ 0.426 vor und ein starker DIF bei einer Effektstärke von ≥ 0.638 (Signifikanzniveau .05) (Trendtel et al., 2016b, S. 127-131).

Itemschwierigkeit

In einem Kompetenztest soll möglichst das gesamte Fähigkeitsspektrum von Versuchsteilnehmenden erfasst werden. Dafür braucht man Items, die unterschiedlich schwierig sind. Neben „mittelschweren" Items benötigt man leichte Items, die von vielen Personen gelöst werden, und schwere Items, die nur von wenigen sehr fähigen Personen beantwortet werden. Es gibt verschiedene Methoden, die Itemschwierigkeit in IRT-Modellen zu berechnen (s. ausführlich in Wu et al., 2016, Kap. 9). Im *Partial-Credit-Modell* (PCM) sind vor allem die sog. *Thurstonian Thresholds* (τ) entscheidend. Das PCM findet bei mehrkategorialen bzw. polytomen Items Anwendung und wurde in Abschnitt 4.1 vorgestellt. Ein Testitem könnte z. B. nicht nur ‚falsch' oder ‚richtig', sondern ‚falsch', ‚teilweise' oder ‚richtig' gelöst werden. In diesem Fall liegt ein mehrkategoriales bzw. polytomes Antwortformat vor. Man könnte beispielsweise ‚falsch' mit 0, ‚teilweise' mit 1 und ‚richtig' mit 2 kodieren. Es wird angenommen, dass eine höhere Kategorie (z. B. 2 für ‚richtig') schwieriger zu erreichen ist als eine niedrigere Kategorie (z. B. 1 für ‚teilweise richtig'). Thurstonian Thresholds (τ) geben nicht die Schwierigkeit für ein komplettes Testitem an, sondern die Schwierigkeit einer Item*kategorie* (also z. B. wie wahrscheinlich es ist, ein Testitem ‚teilweise' zu lösen). Die Thurstonian Threshold τ_k einer Kategorie k ist also definiert als die 50 %-ige Wahrscheinlichkeit, Kategorie k oder höher zu erreichen (Wu et al., 2016, S. 170). Dabei ist es wichtig, dass die Kategorien k richtig geordnet sind, denn im PCM wird angenommen, dass eine Person eine höhere Fähigkeit zeigen muss, um eine höhere Itemkategorie zu erreichen. Die aufeinanderfolgenden Itemkategorien (z. B. ‚falsch', ‚teilweise' und ‚richtig gelöst') sollen aufeinanderfolgende Abschnitte einer Fähigkeit abbilden (Bühner, 2021, S. 326). Thurstonian Thresholds liegen in der Regel in einem Wertebereich von -3 bis $+3$.

In der vorliegenden Arbeit ist außerdem die *klassische Itemschwierigkeit* relevant, die in der Klassischen Testtheorie Anwendung findet. Die klassische Itemschwierigkeit ist die relative Häufigkeit einer Itemkategorie. Nehmen wir an, von 100 Personen haben 30 Personen ein Item ‚falsch', 60 ‚teilweise' und 10 ‚richtig' gelöst. In diesem Fall läge die relative Häufigkeit der einzelnen Itemkategorien bei 30 %, 60 % und 10 % (bzw. 0.30, 0.60 und 0.10). Ein Testitem, das von fast allen

Versuchspersonen korrekt oder gar nicht gelöst wird, liefert nur wenige Informationen zur Erklärung der zugrundeliegenden Kompetenz. Die klassische Itemschwierigkeit sollte deshalb in einem Bereich zwischen 0.05 und 0.95 liegen (Kelava & Moosbrugger, 2020a, S. 155). Wenn die klassische Itemschwierigkeit weniger als 0.05 beträgt, besteht die Möglichkeit, diese Itemkategorie mit der darunterliegenden Kategorie zusammenzufassen.

Interrater-Reliabilität

Man unterscheidet zwischen Items mit *freiem* und *gebundenem* Antwortformat (Bühner, 2021, S. 45). Bei Items mit gebundenem Antwortformat (z. B. Multiple-Choice Items) sind Antwortoptionen bereits vorgegeben. Freie Aufgabenformate sind dadurch gekennzeichnet, dass Antwortmöglichkeiten nicht vorgegeben sind und von den Versuchspersonen selbst erzeugt werden müssen (beispielsweise in Form einer schriftlichen Antwort). In der vorliegenden Arbeit kamen hauptsächlich Items mit freiem Antwortformat zum Einsatz. Die teilnehmenden Personen produzierten also freie Antworten bei der Beantwortung der Items. Diese Texte der Personen wurden von Rater*innen mithilfe von Kodierregeln bewertet. Diese Kodierregeln legten fest, welche Kriterien in einer Antwort erfüllt sein sollten, um eine bestimmte Anzahl von Punkten in einer Aufgabe zu erhalten. Nun musste gewährleistet werden, dass die Rater*innen die Aufgaben zuverlässig bewerten. Die Rater*innen sollten also zu einem gewissen Grad in ihren Bewertungen übereinstimmten, um die Zuverlässigkeit bzw. die Reliabilität der Messung zu gewährleisten. Es gibt verschiedene Möglichkeiten, die Interrater-Reliabilität zu bestimmen (s. ausführlich bei Wirtz & Caspar, 2002). Das gewählte Analyseverfahren hängt dabei u. a. von der Anzahl der Rater*innen ab. In der vorliegenden Arbeit haben zwei Rater*innen die Testitems bewertet. Zur Berechnung der Interrater-Reliabilität wurde Cohens κ verwendet.[19] Werte $\geq .75$ gelten als sehr gute Übereinstimmung, Werte $\geq .60$ als gute Übereinstimmung (Wirtz & Caspar, 2002, S. 59).

4.3. Bestimmung von Kompetenzniveaus

Wie in Kapitel 4.1 beschrieben, hängt in IRT-Modellen die Lösungswahrscheinlichkeit eines Items im Wesentlichen von zwei Parametern ab: der Itemschwierigkeit σ und der Personenfähigkeit θ. Diese beiden Parameter werden auf einer gemeinsamen Skala abgebildet, die auch *Logit-Skala* genannt wird (s. a. Kapitel 4.1). Mithilfe der Logit-Skala können die Kompetenzen beschrieben werden, die bei der Bearbeitung von Testitems benötigt werden. Die Werte auf der Skala werden zu Kompetenzen in Beziehung gesetzt, die im Test benötigt werden. Hierfür wird die Skala zunächst in Abschnitte unterteilt, die die Kompetenzniveaus darstellen. Die Festlegung von Schwellen zwischen Kompetenzniveaus wird auch *Standard-Setting* genannt. Es gibt verschiedene Methoden,

[19] Da die Daten in der vorliegenden Studie ordinales Messniveau aufweisen, wurde der gewichtete Kappa-Koeffizient verwendet (linear weights). Wenn die Rater*innen bei der Bewertung um nur eine Kategorie auseinanderliegen, bedeutet dies eine bessere Übereinstimmung, als wenn sie bspw. um zwei Kategorien auseinanderliegen. Dies wird im gewichteten Kappa-Koeffizient berücksichtigt (s. ausführlich Warrens, 2013).

Schwellen zwischen Kompetenzniveaus festzusetzen (Bond & Fox, 2015, Kap. 9; Luger-Bazinger et al., 2016; Rauch & Hartig, 2020). Die Wahl einer geeigneten Methode hängt meist von der durchgeführten Studie ab und existierende Methoden können je nach Bedarf abgeändert werden (Luger-Bazinger et al., 2016, S. 84). Bei der Bestimmung der Kompetenzniveaus diskutiert ein Panel aus Expert*innen anhand von inhaltlichen Kriterien, an welcher Stelle die Schwellen zwischen den Kompetenzniveaus festgelegt werden sollten.

In der vorliegenden Studie wurden die Kompetenzniveaus in Anlehnung an die *Bookmark-Methode* bestimmt (Lewis et al., 2012; Pitoniak & Cizek, 2016).[20] Die Bookmark-Methode ist eine der populärsten Standard-Setting Methoden (Luger-Bazinger et al., 2016, S. 109). Das methodische Vorgehen wurde für die vorliegende Studie etwas modifiziert. Grund hierfür waren die begrenzten personellen Ressourcen. Zunächst wurden die Items bzw. die Itemkategorien nach ihrer aufsteigenden Schwierigkeit geordnet. Nun wurden zunächst die Kompetenzen diskutiert, die Personen benötigen, um ein bestimmtes Item zu lösen (Karantonis & Sireci, 2006, S. 4-5). Hierbei waren mitunter die Kodierregeln der Items sowie Annahmen aus dem theoretischen Kompetenzmodell (Rolle, 2013) zentral. Die Kodierregeln enthielten bereits viele Informationen darüber, welche Kompetenzen benötigt wurden, um eine bestimmte Itemkategorie zu lösen. Ausgehend von diesen a priori festgelegten Annahmen diskutierten Christian Rolle, Jens Knigge und ich, welche kognitiven Prozesse beim Lösen eines Items eine entscheidende Rolle spielten. Dabei wurde jede einzelne Itemkategorie wiederholt besprochen. Auf diese Weise setzten wir in einem sich wiederholenden Prozess Schwellen für die Kompetenzniveaus. An einer Schwelle zwischen zwei Niveaus kann eine Person mit den entsprechenden Fähigkeiten die Items unterhalb der Schwelle lösen, aber noch nicht die Items oberhalb der Schwelle. Die Fähigkeit, ein Item zu lösen, bedeutet in Schulleistungsstudien i. d. R. eine 65 %ige Lösungswahrscheinlichkeit.[21]

[20] Dieses Vorgehen wird auch beschrieben in Ehninger et al. (2021b).

[21] In vielen Schulleistungsstudien (u. a. *TIMSS*, *DESI*) wird die 65 %ige Lösungswahrscheinlichkeit eines Items bzw. einer Itemkategorie bei der Festlegung von Schwellen zwischen Kompetenzniveaus vorausgesetzt (Rauch & Hartig, 2020, S. 416). Für alle Personen, die über dieser 65 %-Schwelle liegen, kann man demnach davon ausgehen, dass sie diese Aufgabe i. d. R. korrekt lösen.

5. Testdesign und Pilotstudie (Publikation B und C)[22]

Die Aufgaben für den *MARKO*-Test wurden in zwei Pilotstudien entwickelt. Bei der Entwicklung des Tests waren theoretische Annahmen zum musikbezogenen Argumentieren zentral (Kapitel 2.4). Der Test sollte sich an Schüler*innen der neunten bis zwölften gymnasialen Jahrgangsstufe sowie an Musikstudierende richten, um ein möglichst breites Leistungsspektrum zu erfassen. In Kapitel 5.1 wird der Prozess der Aufgabenentwicklung vorgestellt und zwei Beispielaufgaben aus dem Test werden in Kapitel 5.2 erläutert.

5.1. Entwicklung der Testaufgaben[23]

Die Entwicklung der Items für den Kompetenztest erfolgte ausgehend von Rolles theoretischem Kompetenzmodell (Rolle, 2013, 2017; s. a. Abbildung 2.4 auf S. 37). Es lagen bereits erste Testaufgaben vor, die in empirischen Studien zu Rolles Modell erprobt worden waren (Knörzer, 2012; Knörzer et al., 2015; Knörzer et al., 2016). Außerdem dienten Testaufgaben aus dem Projekt *KoMus* als Orientierung (Jordan et al., 2012; Knigge, 2011) sowie Inhalte aus Schulcurricula und Schulbüchern. Die Entwicklung von Testaufgaben fand im Rahmen von zwei Pilotstudien statt. Dieser Aufgabenentwicklungsprozess ist in Abbildung 6 schematisch dargestellt.

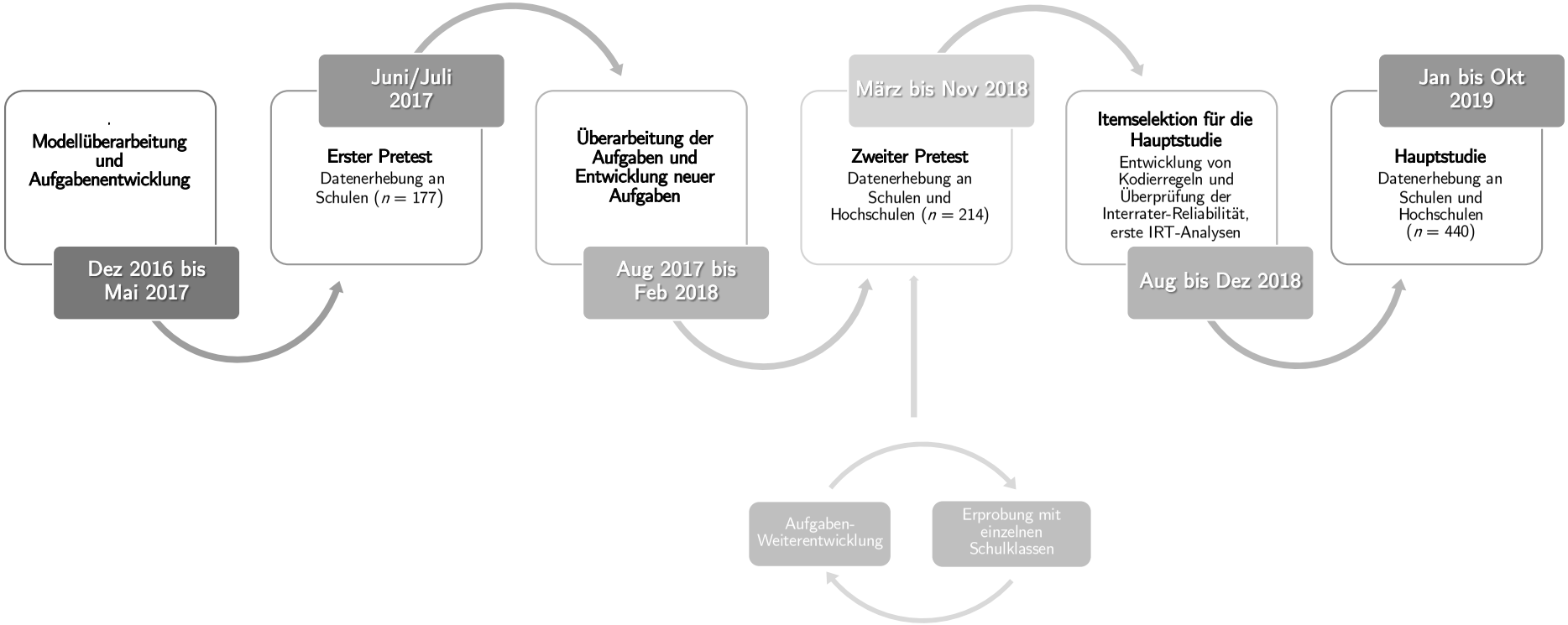
Die Testaufgaben sollten verschiedene Aspekte musikbezogener Argumentationskompetenz erfassen, die in Rolles theoretischem Modell bedeutsam sind. Rolle ging davon aus, dass Personen beim musikbezogenen Argumentieren auf verschiedene Aspekte Bezug nehmen und dass diese Bezugnahmen unterschiedlich anspruchsvoll sind. Während Urteilende auf Niveau 1 lediglich „Gefallen bzw. Missfallen bekunden“, nehmen Personen auf Niveau 2 Bezug auf Autoritäten (Rolle, 2017, S. 141). Personen auf Niveau 3 beziehen sich auf „objektive Eigenschaften der Musik“ (Rolle, 2017, S. 141) und auf Niveau 4 ist der „eigene Eindruck des Ausdrucks der Musik“ relevant.[24] Auf Niveau 5 ‚verschmelzen‘ gewissermaßen Niveau 3 und 4. Hier nehmen Urteilende sowohl Bezug auf „musikalische Parameter“ sowie auf „subjektive Eindrücke“ (Rolle, 2017, S. 141). Ab Niveau 6 können Personen auch „musikkulturelle Besonderheiten“ berücksichtigen und den eige-

[22] Publikation B: Ehninger, J. & Rolle, C. (2020). Musikbezogenes Argumentieren – Nur Geschmacksache? Über die Entwicklung eines Kompetenztests. In M. Schwarzbauer & K. Steinhauser (Hrsg.), *‚Nur‘ Geschmacksache? Der Umgang mit kreativen Leistungen im Musik- und Kunstunterricht* (S. 168–182). LIT.
Publikation C: Ehninger, J., Knigge, J. & Rolle, C. (2021a). Musikbezogene Argumentationskompetenz. Ein Werkstattbericht über die Entwicklung von Testaufgaben. In A. Budke & F. Schäbitz (Hrsg.), *Argumentieren und Vergleichen* (S. 93–112). LIT.

[23] Bei Kapitel 5.1 handelt es sich um eine Zusammenfassung von Ehninger et al. (2021a; Abschnitt 3). Zur besseren Lesbarkeit verzichte ich weitestgehend auf die Kennzeichnung von indirekten Zitaten.

[24] Eine Arbeitsgruppe bestehend aus Thomas Gottschalk, Jens Knigge, Christian Rolle und mir überarbeitete Rolles theoretisches Modell im Zuge der Aufgabenentwicklung. Dabei haben sich teilweise andere Begrifflichkeiten durchgesetzt. Rolle (2017) spricht im theoretischen Modell von „objektive[n] Eigenschaften der Musik“ sowie von „musikalischen Parameter[n]“. Während der Testentwicklung hat es sich durchgesetzt, stattdessen von „musikalischen Merkmalen“ zu sprechen. Anstatt der Begriffe „Eindruck des Ausdrucks der Musik“ bzw. „subjektive Eindrücke“ wurde „Wirkung von Musik“ benutzt.

Abbildung 6.
Schematische Darstellung des Testdesigns und der Pilotstudie



Anmerkung. Diese Grafik ist auch enthalten in Ehninger et al., 2021a, S. 100.

nen Standpunkt „zu anderen Perspektiven ins Verhältnis setzen“ (Rolle, 2017, S. 141).

Die entwickelten Aufgaben operationalisierten jeweils unterschiedliche theoretische Annahmen aus Rolles Modell. In einigen Aufgaben wurden die Versuchsteilnehmenden darum gebeten, auf objektive Eigenschaften bzw. musikalische Merkmale Bezug zu nehmen (Niveau 3 und 5 im theoretischen Modell). Der Verweis auf den Ausdruck bzw. die Wirkung von Musik war ebenfalls Gegenstand von Testaufgaben (Niveau 4 und 5 im theoretischen Modell). Außerdem konzentrierten sich einige Aufgaben auf das Verhältnis des eigenen Standpunktes zu anderen Sichtweisen (Niveau 6 und 7 im theoretischen Modell). Alle Aufgaben, die im Test um Einsatz kamen, wurden mithilfe von Kodierregeln bewertet. Diese Kodierregeln enthielten Informationen darüber, wie die Aufgaben bewertet werden sollten; d. h. welche Anzahl von Punkten für eine Antwort vergeben werden sollte. Die Versuchspersonen konnten je nach Aufgabe zwischen 0 und 3 Punkten erreichen. Wenn eine Person auf eine Aufgabe ausschließlich mit einem Gefallensurteil antwortete, bekam sie 0 Punkte. Dies entsprach Niveau 1 im theoretischen Modell.[25] Zwei Kodierregeln werden exemplarisch an zwei Beispielaufgaben in Kapitel 5.2 vorgestellt.

An den beiden Pilotstudien nahmen 391 Personen teil und es wurden insgesamt 60 Testitems erprobt. Die Datenerhebungen fanden zur Zeit des regulären Musikunterrichts statt und dauerten in der Regel 90 Minuten. Die Testaufgaben standen auf der Plattform *moodle* zur Verfügung. So konnten die teilnehmenden Schüler*innen und Studierenden einzeln an Computern bzw. Laptops arbeiten. Sie hörten Musik über Kopfhörer oder sahen sich Videos an. In den Testaufgaben wurden sie darum gebeten, ihr musikbezogenes Urteil in einer schriftlichen Antwort zu begründen. In den Testaufgaben sollten die Testpersonen beispielsweise ihre Einschätzung begründen, ob ein Musikstück eine bestimmte Atmosphäre erzeugte. Es gab außerdem Aufgaben, in denen eine Filmszene mit verschiedenen Musiken unterlegt war. Die teilnehmenden Personen sollten darlegen, welche Musik ihrer Meinung nach die Filmszene am geeignetsten unterlegte und dies begründen. In anderen Aufgaben wiederum sollten die Versuchsteilnehmenden einen Kommentar zu einer Diskussion unter einem YouTube-Video verfassen, einer Schüler*innen-Band Feedback geben oder auf eine Konzertkritik in einer Zeitung reagieren (s. a. Ehninger et al., 2021b, S. 4; Ehninger & Rolle, 2020, S. 175). Die meisten Aufgaben, die im Rahmen der ersten Pilotierung entwickelt wurden, mussten verworfen oder überarbeitet werden (s. ausführlicher in Ehninger et al., 2021a). Einer der Hauptgründe hierfür war, dass viele Aufgabenstellungen – in Anlehnung an die Aufgaben von Knörzer et al. (2015, 2016) – bewusst recht offen formuliert worden waren. Eine Frage wie beispielsweise *„Gefällt Dir das Musikstück? Begründe Deine Meinung“* war jedoch zu unklar formuliert. Die teilnehmenden Schüler*innen meldeten zurück, dass sie nicht genau wussten, was sie bei der Beantwortung der Aufgaben beachten sollten. Diese Aufgabenstellungen mussten daher präzisiert werden.

Auf Grundlage des Feedbacks von Lehrkräften und Versuchsteilnehmenden wurden ab der zwei-

[25] In der Testentwicklung wurden auch Aufgaben erprobt, die Niveau 2 des theoretischen Modells überprüfen sollten. Auf Niveau 2 beziehen sich Personen auf Autoritäten bei der Urteilsbegründung. Solche Begründungsmuster konnten in der vorliegenden Studie allerdings nicht gefunden werden (s. a. Diskussion in Kapitel 7.2).

ten Pilotierung die Testaufgaben ständig überarbeitet. Insgesamt wurden 60 Testitems erprobt, von denen 25 in der Hauptstudie zum Einsatz kamen. Bei der Pilotierung stellte sich heraus, dass offene Aufgabenformate besonders geeignet waren, um musikbezogene Argumentationskompetenz zu erfassen. In offenen Aufgaben können Argumente von den Versuchspersonen selbst produziert werden. Im Gegensatz dazu können Personen in geschlossenen Aufgabenformaten keine Argumente selbst hervorbringen. Durch die offenen Aufgabenformate variierte jedoch die Bearbeitungsdauer des Tests erheblich. Während einige Personen ausführliche Antworten schrieben, verfassten andere Versuchsteilnehmende nur wenige Sätze. Die Beispiele im nächsten Kapitel werden zeigen, dass Antworten, die mehr Punkte bekommen, häufig länger sind als Antworten, die weniger Punkte bekommen. Aus diesem Grund wurde den Versuchsteilnehmenden mitgeteilt, dass sie sich Zeit lassen sollten bei der Beantwortung der Testaufgaben und dass es nicht wichtig sei, alle Testaufgaben zu bearbeiten. Zwei solcher Testaufgaben sowie deren Kodierregeln werden im folgenden Kapitel vorgestellt.

5.2. Beispielaufgaben und Kodierregeln[26]

Beide Aufgaben, die in diesem Kapitel vorgestellt werden, kamen in dieser Form in der Hauptstudie zum Einsatz. Die Aufgabe „Star Wars“ sollte erfassen, wie sich Personen auf musikalische Merkmale und auf die Wirkung der Musik bezogen. Die Aufgabe „Eurovision Song Contest“ bezog sich wiederum auf das Verhältnis des eigenen Standpunktes zu anderen Meinungen. Hier sollten Personen Stellung zu einer Diskussion auf YouTube beziehen. Um die jeweilige Aufgabe zu beantworten, schrieben die Versuchsteilnehmenden Texte. Diese Texte mussten mithilfe von Kodierregeln bewertet werden, die für jede Aufgabe passgenau entwickelt wurden.[27] Im Folgenden werden die beiden Beispielaufgaben „Star Wars“ (Kapitel 5.2.1) und „Eurovision Song Contest“ (Kapitel 5.2.2) sowie deren Kodierregeln vorgestellt.[28]

5.2.1. Beispielaufgabe „Star Wars“

In der Aufgabe „Star Wars“ sollten sich die Versuchsteilnehmenden auf musikalische Merkmale sowie auf die Wirkung der Musik beziehen (Abbildung 7). Die Bezugnahme zu musikalischen Merkmalen und zur Wirkung bzw. zum Ausdruck der Musik ist in Rolles Modell auf den Niveaus 3 bis 5 relevant. Das beobachtete Antwortverhalten der Versuchsteilnehmenden unterschied sich jedoch von den theoretischen Annahmen. Im theoretischen Modell wird zwischen einer Bezugnahme auf „objektive Eigenschaften“ (Niveau 3) und dem „Eindruck des Ausdrucks der Musik“

[26] Die beiden Beispielaufgaben sowie die Kodierregeln wurden ebenfalls in Ehninger et al. (2021a) vorgestellt. Indirekte Zitate für die bessere Lesbarkeit des Kapitels nicht gekennzeichnet.

[27] Wegen des Copyrights von Musik- und Videodateien können die Aufgaben des *MARKO*-Tests leider nicht vollständig veröffentlicht werden. Eine entsprechend zensierte Version ist online verfügbar (Link auf S. 17).

[28] Eine weitere Beispielaufgabe findet sich in Ehninger und Rolle (2020, S. 174-176). In dieser Aufgabe sollten Testpersonen einer Schüler*innen-Band Feedback für ihre musikalische Performance geben.

Abbildung 7.
Beispielaufgabe „Star Wars“

Musik wird in Filmen häufig dazu verwendet, um eine bestimmte Stimmung zu vermitteln.

In den Science Fiction Filmen „Star Wars“ spielen viele Filmszenen in Raumschiffen, die sich im Weltall in einer entfernten Galaxie bewegen. Im folgenden Musikstück soll die Atmosphäre im Weltraum dargestellt werden.

Hier siehst du ein Bild aus der Filmszene: [Hier steht im Test ein Bild]

Findest du, dass es im Musikausschnitt gelungen ist, die „Weltraum-Atmosphäre“ darzustellen? Begründe deine Antwort und gehe dabei auf die musikalischen Mittel ein, die der Filmkomponist verwendet hat.

Anmerkung. Die Versuchsteilnehmenden hörten einen Ausschnitt aus der Filmmusik von Star Wars (Episode I, „Arrival at Naboo“, 1:14-1:48 aus der Aufnahme des London Symphony Orchestras). In der Aufgabe war außerdem ein Bild aus der Filmszene abgebildet, das hier aus urheberrechtlichen Gründen entfernt werden musste. Das Bild zeigte den Blick aus dem Cockpit eines Raumschiffes auf einen Planeten. Eine Darstellung des Items findet sich ebenfalls in Ehninger et al. (2021a, S. 103) und in Ehninger et al. (2021b, S. 4).

(Niveau 4) unterschieden (Rolle, 2017, S. 141). In den empirischen Daten konnten diese beiden Bezugnahmen jedoch nur selten isoliert voneinander beobachtet werden. Außerdem fiel über die gesamte Bandbreite des Tests auf, dass sich Personen häufig auf ‚oberflächliche‘ musikalische Merkmale wie z. B. „laut“, „leise“, „schnell“ oder „langsam“ bezogen. Diese ‚oberflächlichen‘ musikalischen Merkmale werden in der vorliegenden Arbeit *saliente* musikalische Merkmale genannt. Saliente musikalische Merkmale sind besonders auffallende musikalische Merkmale wie z. B. prägnante Lautstärkeunterschiede.[29] Im theoretischen Modell ist jedoch nur allgemein von „objektiven Eigenschaften“ der Musik bzw. „musikalischen Parametern“ die Rede (Rolle, 2017, S. 141). Es war also sinnvoll, die Bezugnahme zu musikalischen Merkmalen in den Kodierregeln auszudifferenzieren. Die Kodierregeln für diese Aufgabe wurde in mehreren Interpretationswerkstätten diskutiert und verfeinert. In den finalen Kodierregeln der „Star Wars“-Aufgabe bekamen Versuchspersonen einen Punkt, wenn sie sich entweder nur auf die Wirkung der Musik oder auf saliente Merkmale bezogen (in der Kodierregel als „einfache und oberflächliche Merkmale“ bezeichnet). Mehr Punkte wurden vergeben, wenn Personen auf die Wirkung der Musik sowie auf musikalische Merkmale Bezug nahmen, die über saliente Merkmale hinausgingen. Musikalische Merkmale, die keine salienten Merkmale sind, werden in der vorliegenden Arbeit als *spezifische* musikalische Merkmale bezeichnet. Tabelle 2 zeigt die Kodierregeln für das Item „Star Wars“ sowie Beispielantworten aus den Testdaten.

Zwei Rater*innen kodierten ca. 15 % der erhobenen Daten dieser Aufgabe. Wenn die Übereinstimmung in einem guten Bereich lag, kodierte aus ökonomischen Gründen nur eine Person die

[29] In der Wahrnehmungspsychologie bezeichnet *Salienz* die „Deutlichkeit“ von Reizen (Ansorge, 2021). Ein salientes Merkmal ist demnach ein Merkmal, das besonders auffällig ist und direkt ‚ins Auge springt‘.

Tabelle 2.
Kodierregeln der Aufgabe „Star Wars“

Pkt.	Kodierregel	Beispiele
0	Tautologische oder keine Begründung	„Ja, auf Grund der Atmosphäre die im Weltraum herrscht. Dies hat der Komponist sehr gut daargestellt [sic]" (VP_661)
1	Personen beziehen sich nur auf die Wirkung. Falls musikalische Merkmale aufgeführt werden (und gar ein kausaler Bezug zwischen erzeugter Wirkung und musikalischen Merkmalen hergestellt wird) erfolgt dies unter Bezugnahme auf „einfache“ und oberflächliche Merkmale („helle Töne“, „lange Töne“, „laut“, „leise“, „Instrumente, die Spannung erzeugen“, ...) [...].	„Ich finde schon, da es sich spannend und unbekannt anhört, was meiner Meinung nach gut zur Weltraum-Atmospähre [sic] passt“ (VP_714).
2	Personen stellen einen Bezug zwischen der erzeugten Wirkung und musikalischen Merkmalen her. Sobald Instrumente (z.B. „ruhige Streicher“) genannt werden, sind zwei Punkte zu vergeben.	„Ja, ich finde es sehr gelungen. die flächenartigen Klänge beschreiben die unendliche Weite des Universums[.] die Syntheshizer [sic] geben dem Stück das futuristische des Weltalls [.] einzelne hohe Töne zur verdeutlichung [sic] der Sterne“ (VP_589).
3	Personen stellen einen Bezug zwischen der erzeugten Wirkung und musikalischen Merkmalen her. Es erfolgt eine detailliertere Beschreibung (z. B. der Spielweise der Instrumente, des Formverlaufs, etc.)	„Ich finde die Umsetzung gelungen, weil die langen Töne (der Violine) ein Gefühl von Weite vermitteln und trotzdem (wegen der Höhe) vor allem am Anfang ziemlich aufgeregt und dramatisch klingen. Die schnellen (Xylophon?)-Töne, die die Tonleiter herauf und herunter gehen, haben einen leuchtenden Klang und erinnern an Sterne. Der Tusch am Anfang könnte andeuten, dass sich dem Zuschauer da gerade die ganze Sicht auf die spektakuläre Umgebung öffnet“ (VP_610).

Anmerkung. Die Rechtschreibung und Zeichensetzung der Schüler*innen-Beispiele wurde nicht korrigiert. Die Kodierregeln für diese Aufgabe sind ebenfalls in Ehninger et al. (2021a, S. 104) abgedruckt.

restlichen Daten. Die Interrater-Reliabilität für das „Star Wars“-Item lag mit $\kappa = .78$ in einem sehr guten Bereich.[30]

5.2.2. Beispielaufgabe „Eurovision Song Contest“

Das Item „Eurovision Song Contest“ widmete sich dem Verhältnis des eigenen Standpunktes zu anderen Sichtweisen. Bei diesem Item spielte der soziale Kontext von Musik eine Rolle. Die Versuchsteilnehmenden sahen einen Ausschnitt aus der Gewinnerperformance des Eurovision Song Contests von 2018. In dieser Aufgabe sollten sie Stellung zu einer Diskussion beziehen, die unterhalb des Videos auf YouTube zu sehen war (Abbildung 8).

[30] Für die Berechnung der Interrater-Reliabilität wurde Cohens gewichteter Kappa-Koeffizient mit linear weights verwendet (s. a. Kapitel 4.2).

Abbildung 8.
Beispielaufgabe „Eurovision Song Contest“

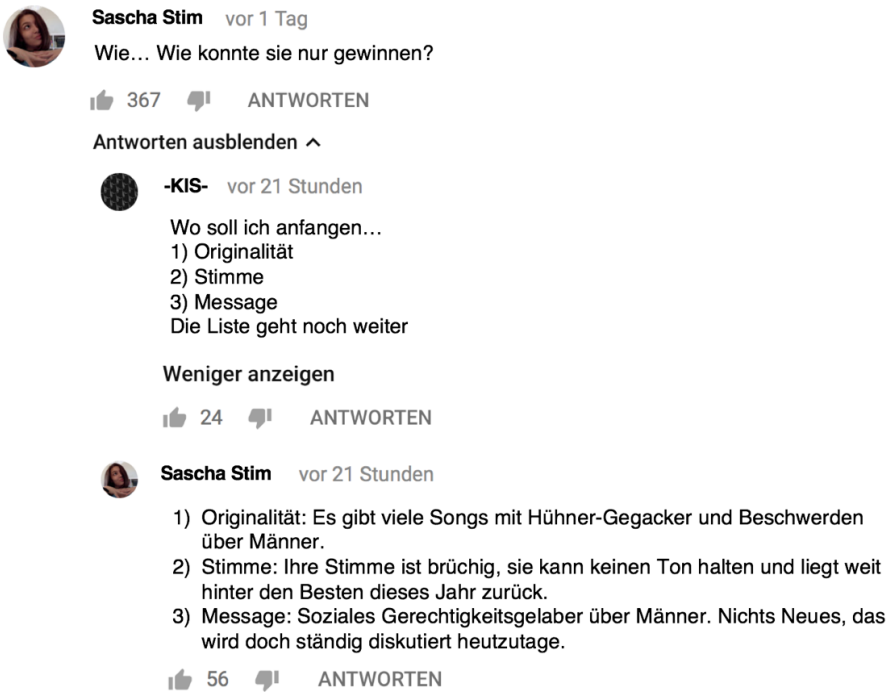
Die israelische Sängerin Netta hat mit ihrem Song „Toy" den Eurovision Song Contest 2018 gewonnen. Der Song wurde in den Medien kontrovers diskutiert.

Im Refrain singt sie „I'm not your toy – You stupid boy" („Ich bin nicht dein Spielzeug – Du dummer Kerl“).

Klicke auf Play, um dir einen Ausschnitt aus Nettas Auftritt anzuschauen.

[Hier steht im Test ein Video-Player]

Unter dem Musikvideo ist auf YouTube die folgende Diskussion entbrannt:



Wie siehst du das? Verfasse einen kritischen Kommentar und beziehe Stellung. Gehe in deiner Antwort auf die Argumente ein, die die anderen beiden ausgetauscht haben.

Anmerkung. In der Aufgabe wurde ein kurzer Ausschnitt aus von Nettas Performance gezeigt (1:03–1:40). Das komplette Video gibt es hier: https://www.youtube.com/watch?v=84LBjXaeKk4. Die Aufgabe ist ebenfalls in Ehninger et al. (2021a, S. 106) abgedruckt.

Viele Testpersonen gaben lediglich die YouTube-Diskussion wieder. Antworten, die Teile des Aufgabenstammes paraphrasierten oder sich auf das subjektive Gefallen bezogen, wurden mit 0 Punkten bewertet. Wenn sich eine Person an der kompletten Diskussion abarbeitete und ggf. ein neues Argument benannte, wurde 1 Punkt vergeben. 2 Punkte gab es, wenn eine Antwort mindestens zwei neue Argumente enthielt und verschiedene Perspektiven abgewogen wurden. Tabelle 3 zeigt

Tabelle 3.
Kodierregeln der Aufgabe „Eurovision Song Contest“

Pkt.	Kodierregel	Beispiele
0	Antworten, die die Teile des Aufgabenstamms paraphrasieren oder/und sich auf das subjektive Gefallen beziehen.	"ich bin mir sicher, dass sie nur wegen dem Songinhalt gewonnen hat- und naja, der Song und ihre Stimme ist nicht schlecht- abgesehen von dem Hühner- Gegacker." (P2_10)
1	Antwort arbeitet sich an der kompletten YouTube-Diskussion ab, bleibt jedoch weitestgehend paraphrasierend. Möglicherweise enthält die Antwort ein neues Argument, das im Aufgabenstamm noch nicht aufgetaucht ist.	„Die Sängerin spricht ein sehr wichtiges und auch aktuelles Thema an. Also die soziale Gleichberechtigung, jedoch finde ich es nicht wirklich passend gewählt es den Mitmenschen rüberzubringen. Der Text wird humorvoll wiedergegeben und somit keinen Sinn erreicht.” (Vp_142)
2	Die Antwort enthält mind. zwei neue Argumente. Verschiedene Perspektiven werden abgewogen (bzw. wird die Argumentation ausführlicher entfaltet).	„Women empowerment ist ein sehr aktuelles Thema, welches wichtig ist. Es ist gut das Künstlerinnen und Künstler ein Zeichen setzen. Teilweise ist der Text eindimensional, da auch Frauen mit Frauen ‘spielen’. Oft ist es jedoch anders herum und schon seit Jahrhunderten der Fall, aufgrund der ungerechten Macht Verteilung, wo Frauen zu kurz kommen. Sie hätte vielleicht lieber singen sollen ‚´I´m not a toy, for no one‘ oder so etwas in der Art, was mehr den Gleichberechtigungs Gedanken hervorbringt. Sie repräsentiert ein starkes Frauenbild, was definitiv Sozialkritisch ist. Durch das sogenannte ‘Hühner-Gegacker’, wie Sascha es bezeichnet, ist das Lied ungewohnt und anders und unterscheidet sich von der gesellschaftlichen Norm, die die Masse beeinflusst, wie Sascha und 367 andere Personen zeigen. Viel Spaß mit eurem Mitläufertum und dem daraus resultiertem Einheitsbrei“ (Vp_89).

Anmerkung. Die Rechtschreibung und Zeichensetzung der Schüler*innen-Beispiele wurde nicht korrigiert. Die Kodierregeln für diese Aufgabe sind ebenfalls in Ehninger et al. (2021a, S. 107) abgedruckt.

die Kodierregeln der Aufgabe „Eurovision Song Contest“ einschließlich Beispielantworten.

Auch bei dieser Aufgabe wurden ca. 15 % der erhobenen Daten von zwei Rater*innen kodiert. Die Interrater-Reliabilität lag bei $\kappa = .94$.[31] Aus ökonomischen Gründen kodierte lediglich eine Person die verbleibenden Daten.

[31] Für die Berechnung der Interrater-Reliabilität wurde Cohens gewichteter Kappa-Koeffizient mit linear weights verwendet (s. a. Kapitel 4.2).

6. Hauptstudie

Publikation D[32]

Die vorliegende Arbeit verfolgte zwei wesentliche Ziele (s. a. Kapitel 3). Eines bestand im Design eines Kompetenztests für musikbezogenes Argumentieren. Dieses Testdesign wurde im vorherigen Kapitel bereits vorgestellt. Ob der entwickelte Test die Kompetenz objektiv, reliabel und valide maß, wird im folgenden Kapitel analysiert. Das zweite Ziel der vorliegenden Arbeit war, musikbezogene Argumentationskompetenz zu messen und Kompetenzniveaus ausgehend von empirischen Daten zu beschreiben. Die Kompetenzmessung sowie die Beschreibung von Kompetenzniveaus sind ebenfalls Gegenstand dieses Kapitels.

Wie bereits beschrieben, wurde der Kompetenztest für musikbezogenes Argumentieren in einer Pilotstudie entwickelt (Kapitel 5). Die teilnehmenden Schüler*innen und Studierenden sollten in den Testaufgaben ihr musikbezogenes Urteil begründen. Zu diesem Zweck produzierten die Versuchsteilnehmenden in offenen Aufgabenstellungen Texte, die anschließend mithilfe von Kodierregeln bewertet wurden (Kapitel 5.2). Ob die Testaufgaben musikbezogene Argumentationskompetenz objektiv, reliabel und valide maßen, wurde im Rahmen der Hauptstudie mit statistischen Verfahren geprüft. Die statistischen Verfahren, die dabei zum Einsatz kamen, sind überwiegend der Item-Reponse-Theorie (IRT) zuzuordnen (Kapitel 4.1). In der IRT wird angenommen, dass die Wahrscheinlichkeit eine Aufgabe zu lösen im Wesentlichen von zwei Dingen abhängt: von der Schwierigkeit der Aufgabe (des Items) und von der Fähigkeit der Person, die die Aufgabe löst. Die Itemschwierigkeit und die Personenfähigkeit werden auf einer gemeinsamen Skala abgebildet. So können Rückschlüsse über die zugrundeliegende Kompetenz gezogen werden und es können Kompetenzniveaus beschrieben werden.

In Kapitel 6.1 beschreibe ich die Datenerhebung und stelle die Stichprobe sowie die verwendeten Erhebungsinstrumente vor. Wie bereits erwähnt, sollten die Testaufgaben musikbezogenes Argumentieren objektiv, reliabel und valide messen. Um dies zu überprüfen, wurden die Testaufgaben mithilfe statistischer Verfahren analysiert (Kapitel 6.2). In Kapitel 6.3 stelle ich das statistische Modell vor, das am besten zu den empirischen Daten passt. Außerdem werden in diesem Kapitel Kompetenzniveaus für musikbezogenes Argumentieren beschrieben. Weiterführende Analysen zu Einflussfaktoren auf die Ausprägung der Kompetenz sind Gegenstand von Kapitel 6.4.

[32] Publikation D: Ehninger, J., Knigge, J., Schurig, M. & Rolle, C. (2021b). A New Measurement Instrument for Music-Related Argumentative Competence: The MARKO Competency Test and Competency Model. *Frontiers in Education*, *6*(191). https://doi.org/10.3389/feduc.2021.668538.
Bei diesem Kapitel handelt es sich um eine Zusammenfassung der Publikation. Zur besseren Lesbarkeit verzichte ich weitestgehend auf die Kennzeichnung von indirekten Zitaten.

6.1. Datenerhebung

6.1.1. Stichprobe[33]

Neun Gymnasien und zwei Hochschulen aus Nordrhein-Westfahlen nahmen 2019 an den Datenerhebungen teil ($N = 440$, 44.5 % weiblich, 52.3 % männlich, 3.2 % fehlende Werte). Die teilnehmenden Studierenden waren in musikpädagogischen Studiengängen eingeschrieben. Drei Schüler*innen der Stichprobe waren von anderen Schulen zu Besuch. Die Datenerhebungen fanden während des regulären Musik- bzw. Seminarunterrichts statt und folgten den Datenschutzrichtlinien des *Schulgesetzes für das Land Nordrhein-Westfahlen* (§ 120 Abs. 4 SchulG). Etwa ein Drittel der Stichprobe besuchte die neunte Klasse (Alter: 14–15), 24.5 % waren in der zehnten Klasse (Alter: 15–16), 28.6 % in der elften Klasse (Alter: 16–17), 5.5 % in der zwölften Klasse (Alter: 17–18) und 7.7 % waren Studierende musikpädagogischer Studiengänge. Das Durchschnittsalter lag bei 16 Jahren ($SD = 2.79$). Bei 24.4 % der Testpersonen waren beide Elternteile und bei 18.8 % war je ein Elternteil im Ausland geboren worden.

6.1.2. Erhebungsinstrumente

Die 90-minütige Datenerhebung begann mit dem *MARKO*-Kompetenztest für musikbezogenes Argumentieren. Bereits in der Testentwicklung hatte sich abgezeichnet, dass die Bearbeitungsdauer des Tests zwischen den Testpersonen stark variierte. Dies konnte vor allem auf die offenen Aufgabenformate zurückgeführt werden. Außerdem unterschieden sich die Texte, die die teilnehmenden Personen schrieben, stark in ihrer Länge. Bei der Datenerhebung sollte die Möglichkeit geschaffen werden, sich ausführlich mit den Aufgaben zu beschäftigen. Den Versuchsteilnehmenden wurde deshalb in der Versuchseinführung mitgeteilt, dass sie nicht alle Items beantworten müssen, sondern sich Zeit lassen können bei der Bearbeitung. Dies hatte zur Folge, dass einige Personen nicht alle 25 Aufgaben bearbeiteten, sondern nur einen Teil. Damit Antworten für möglichst viele Testaufgaben gesammelt werden konnten, kamen drei verschiedene Testhefte zum Einsatz. Diese Testhefte enthielten jeweils dieselben Aufgaben, jedoch in unterschiedlicher Reihenfolge (Tabelle A.1 auf S. 113). Mit diesem Vorgehen konnte gewährleistet werden, dass genügend Daten pro Aufgabe gesammelt wurden. Die drei Testhefte waren etwa gleichmäßig innerhalb der Stichprobe verteilt (Heft I: 33.6 %; Heft II: 33.9 %; Heft III: 32.5 %). 15 Minuten vor Ende der Erhebung sollten die Testpersonen die Bearbeitung des Kompetenztests abbrechen, um den Begleitfragebogen zu bearbeiten. Im Begleitfragebogen gab es demografische Abfragen (Geschlecht, Alter, Migrationsgeschichte, Sprachgebrauch). Außerdem wurden musikspezifische Fragen gestellt, etwa ob die Personen Instrumental- oder Gesangsunterricht nahmen oder ein Instrument spielten. Zudem kam ein Fragebogen zur musikalischen Erfahrenheit zum Einsatz (*Gold-MSI*, Subskala „General Musical Sophistication“; Müllensiefen et al., 2014; Schaal et al., 2014).

[33] Die Stichprobe der Hauptstudie wird ebenfalls in Ehninger et al. (2021b, S. 3) beschrieben. Indirekte Zitate sind für die bessere Lesbarkeit des Abschnitts nicht gekennzeichnet.

Die Aufgaben des *MARKO*-Kompetenztests wurden zunächst statistisch analysiert. Es sollte gewährleistet sein, dass die Testaufgaben die untersuchte Kompetenz objektiv und reliabel erfassten.

6.2. Item-Analysen und Überprüfung der Modellstruktur[34]

Bevor eine Kompetenz auf Basis empirischer Daten beschrieben werden kann, muss zunächst sichergestellt werden, dass die Daten gewisse Kriterien erfüllen. Zu diesen Kriterien gehört beispielsweise, dass die Testaufgaben dieselbe Kompetenz messen müssen. Es muss also sichergestellt werden, dass eine Aufgabe nicht versehentlich etwas anderes misst (z. B. Lesekompetenz statt musikbezogener Argumentationskompetenz). Außerdem sollten die Aufgaben weder zu schwer noch zu leicht sein, da sie sonst wenig Informationen über die untersuchte Kompetenz liefern. Ferner sollten die Testaufgaben fair sein. Dies bedeutet, dass Personen z. B. aufgrund ihres Geschlechts nicht benachteiligt werden dürfen. Wenn Testaufgaben diese Kriterien *nicht* erfüllten, wurden sie aussortiert. Das Modell, das die Kompetenz statistisch beschreibt, wurde also nur mit den Aufgaben geschätzt, die bestimmte Gütekriterien erfüllten. Die Analyseverfahren und die zugrundeliegenden Gütekriterien, die in der vorliegenden Arbeit Anwendung fanden, wurden in Kapitel 4.2 zusammengetragen.

Die Analyse der Daten erfolgte in *R* (Version 3.6.2; R Core Team, 2019). Für die statistischen Schätzungen (IRT-Skalierung) wurden in erster Linie die Pakete *TAM* (Robitzsch et al., 2020) und *eRm* (Mair et al., 2020) verwendet. 27 Personen bearbeiteten weniger als ein Drittel der Testaufgaben und wurden deshalb aus der Analyse ausgeschlossen.[35] Die Analysen wurden schließlich mit $N = 440$ durchgeführt. Fehlende Werte wurden nicht ergänzt bzw. imputiert.

Um zu gewährleisten, dass die Testaufgaben objektiv bewertet wurden, wurde zunächst die Übereinstimmung der Rater*innen (*Interrater-Reliabilität*) sichergestellt. Hierfür kodierten zwei Rater*innen ca. 15 % der Antworten jeder Testaufgabe. Wenn ihre Übereinstimmung in einem guten bis sehr guten Bereich lag, kodierte aus ökonomischen Gründen nur eine Person die restlichen Daten. Die Übereinstimmung der beiden Rater*innen war für alle Testaufgaben gut bis sehr gut. Die Interrater-Reliabilität wird in der vorliegenden Arbeit mit Cohens κ angegeben, welche zwischen .73 und .94 lag (κ Koeffizient mit linear weights) (s. a. Ehninger et al., 2021b, S. 4), wobei Werte $\geq .75$ als sehr gute Übereinstimmung gelten (Wirtz & Caspar, 2002, S. 59). Im Anhang sind die κ-Werte für jede Aufgabe aufgelistet (Tabelle A.2 auf S. 114).

Wie eingangs erwähnt, sollten die Aufgaben weder zu schwer noch zu leicht sein. Wenn beispielsweise alle Personen eine Aufgabe in einem Test lösen, liefert diese Aufgabe keine Informationen über die untersuchte Kompetenz. Dasselbe gilt auch für die Antwortkategorien von Aufgaben.

[34] Dieser Abschnitt behandelt das Unterkapitel „Item Selection“ aus Ehninger et al. (2021b, S. 4-5). Die Darstellung hier ist etwas ausführlicher als in der Publikation. Indirekte Zitate sind für die bessere Lesbarkeit nicht gekennzeichnet.

[35] Die drei Testhefte bestanden aus jeweils drei Aufgabenblöcken, die in unterschiedlicher Reihenfolge angeordnet waren (s. Tabelle A.1 auf S. 113). Es wurden nur Personen in die Analyse aufgenommen, die mindestens einen dieser Aufgabenblöcke bearbeitet hatten. Dies entsprach etwa einem Drittel der Testaufgaben bzw. 8 Aufgaben.

Wenn eine Aufgabe mit 0 bis 3 Punkten bewertet wird, hat diese Aufgabe vier Antwortkategorien (0, 1, 2, 3). Wenn nur sehr wenige Personen 3 Punkte bei dieser Aufgabe erhalten, dann ist diese Antwortkategorie wenig informativ. Deshalb hat es sich als Konvention in der Testtheorie etabliert, dass mindestens 5 % und höchstens 95 % der Testpersonen eine Aufgabe bzw. eine Antwortkategorie lösen sollten (Kelava & Moosbrugger, 2020a, S. 155). Die Lösungshäufigkeit einer Aufgabe bzw. einer Antwortkategorie wird auch *klassische Itemschwierigkeit* genannt (Kapitel 4.2).

Antwortkategorien, die nicht häufig genug vorkamen, wurden mit der darunter liegenden Antwortkategorie zusammengefasst. Beispielsweise konnten bei der Aufgabe „Eurovision Song Contest" (Kapitel 5.2) Versuchspersonen ursprünglich maximal drei Punkte erreichen. Allerdings wurden drei Punkte nur sehr selten vergeben, sodass die relative Häufigkeit dieser Antwortkategorie unter 5 % lag. Die höchste Kategorie (3 Punkte) wurde daher mit der darunter liegenden Kategorie (2 Punkte) zusammengefasst. Alle Personen, die bei dieser Aufgabe ursprünglich drei Punkte bekommen hatten, erhielten nun also nur noch zwei Punkte. Insgesamt musste bei acht Items die höchste Antwortkategorie mit der darunter liegenden Kategorie zusammengefasst werden.[36] So lag letztendlich die relative Häufigkeit der Antwortkategorien der Testaufgaben zwischen 5 % und 79 %.

Ein weiteres Kriterium, das die Aufgaben erfüllen sollten, war, dass es schwieriger sein musste, drei Punkte bei einer Aufgabe zu erreichen, statt beispielsweise zwei Punkte. Es musste also schwieriger sein, eine höhere Antwortkategorie (z. B. 3 Punkte) zu erreichen, statt einer niedrigeren Antwortkategorie (z. B. 2 Punkte). Dies ist eine Voraussetzung für das statistische Modell, das in der vorliegenden Arbeit geschätzt wurde, das *Partial-Credit-Modell* (PCM) (siehe S. 51). Im PCM werden für jede Antwortkategorie Schwierigkeitsparameter geschätzt. Diese müssen richtig geordnet sein. Das bedeutet, dass z. B. die Antwortkategorie ‚2 Punkte' einen niedrigeren Schwierigkeitsparameter haben muss, als die Antwortkategorie ‚3 Punkte'. Für die Schätzung dieser Schwierigkeitsparameter gibt es verschiedene Möglichkeiten. Eine Möglichkeit ist die Verwendung sog. *Thurstonian Thresholds* (Kapitel 4.2). In dieser Arbeit traten die Schwierigkeitsparameter der Antwortkategorien aller Aufgaben in der richtigen Reihenfolge auf (Tabelle A.3 auf S. 115). D. h., dass es bei allen Aufgaben empirisch gesehen schwieriger war beispielsweise 3 Punkte bei einer Aufgabe zu bekommen als 2 Punkte.

In der vorliegenden Arbeit wurde ein statistisches Modell geschätzt, das auf die empirischen Daten ‚passen' sollte. Das statistische Modell sollte also die empirischen Daten möglichst gut beschreiben. Hierfür muss geprüft werden, inwiefern die empirischen Daten zum statistischen Modell passen. Die empirischen Daten – also die Testaufgaben – müssen zu den statistischen Erwartungen passen. Um dies zu überprüfen werden *Itemfit*-Werte herangezogen. Sie geben an, ob die empirisch beobachteten Werte einer Aufgabe (eines Items) zu den erwarteten statistischen Werten passen (Kapitel 4.2). Ein Wert von 1 würde bedeuten, dass die empirisch beobachteten Werte den erwarteten statistischen Werten entsprechen. Die Items zeigten angemessene Infit-Werte (zwischen

[36] Welche Itemkategorien zusammengefasst wurden kann Abschnitt „Item Selection" im *R*-Skript entnommen werden (Link auf S. 17 in der vorliegenden Arbeit).

0.79 und 1.22) und Outfit-Werte (zwischen 0.78 bis 1.36) (Tabelle A.4 auf S. 116).[37]

Der Test sollte für alle Personen der Stichprobe dieselbe Fähigkeit bzw. Eigenschaft messen, was auch *Personenhomogenität* genannt wird (Bühner, 2021, S. 272). Verschiedene Analyseverfahren überprüften in der vorliegenden Arbeit, ob der Test für alle Personen dieselbe Fähigkeit bzw. Eigenschaft maß (Andersen-Test bzw. bedingter Likelihood-Quotienten-Test, Q3-Statistiken, Wald-Test; s. Kapitel 4.2). Mithilfe dieser Verfahren wird auch geprüft, ob ein statistisches Modell auf die empirischen Daten passt. Für die Überprüfung der Modelpassung und der Personenhomogenität wurde die Stichprobe geteilt und es wurde analysiert, ob die Aufgaben für beide Teilstichproben gleich schwierig waren ($n = 220$ in beiden Teilstichproben). Im Idealfall sollten die Items für beide Teilstichproben gleich schwer sein. Die Stichprobe wurde in zwei getrennten Analysen anhand der Variable „Geschlecht“ sowie einer zufälligen Variable in zwei Teilstichproben geteilt. Anschließend wurde überprüft, ob sich die Itemschwierigkeitsparameter innerhalb der beiden Teilstichproben voneinander unterschieden. Der Andersen-Test war signifikant, was bedeutet, dass die Itemschwierigkeitsparameter innerhalb der Teilstichproben voneinander abwichen. Wenn man eine Stichprobe teilt, kann es jedoch passieren, dass einzelne Antwortkategorien innerhalb der Teilstichproben nicht häufig genug auftreten. In diesem Fall kann es zu Problemen bei der statistischen Schätzung kommen, weshalb der Andersen-Test die Personenhomogenität etwas zu streng bewertet. Deshalb wurde zusätzlich eine grafische Modellkontrolle durchgeführt, um die Itemschwierigkeitsparameter der beiden Teilstichproben zu untersuchen. Abbildung 9 zeigt die grafische Modellkontrolle mit einem Zufallskriterium. Der Abbildung kann man entnehmen, dass es keine Ausreißer gab. Die Konfidenzellipsen aller Items liegen in der Nähe der Diagonale oder schneiden sie. Deshalb wurde davon ausgegangen, dass der Test für alle Personen dieselbe Fähigkeit bzw. Eigenschaft misst und die Personenhomogenität gewährleistet war. Die Q3-Statistiken[38] und der Wald-Test[39] zeigten ebenfalls keine größeren Auffälligkeiten.

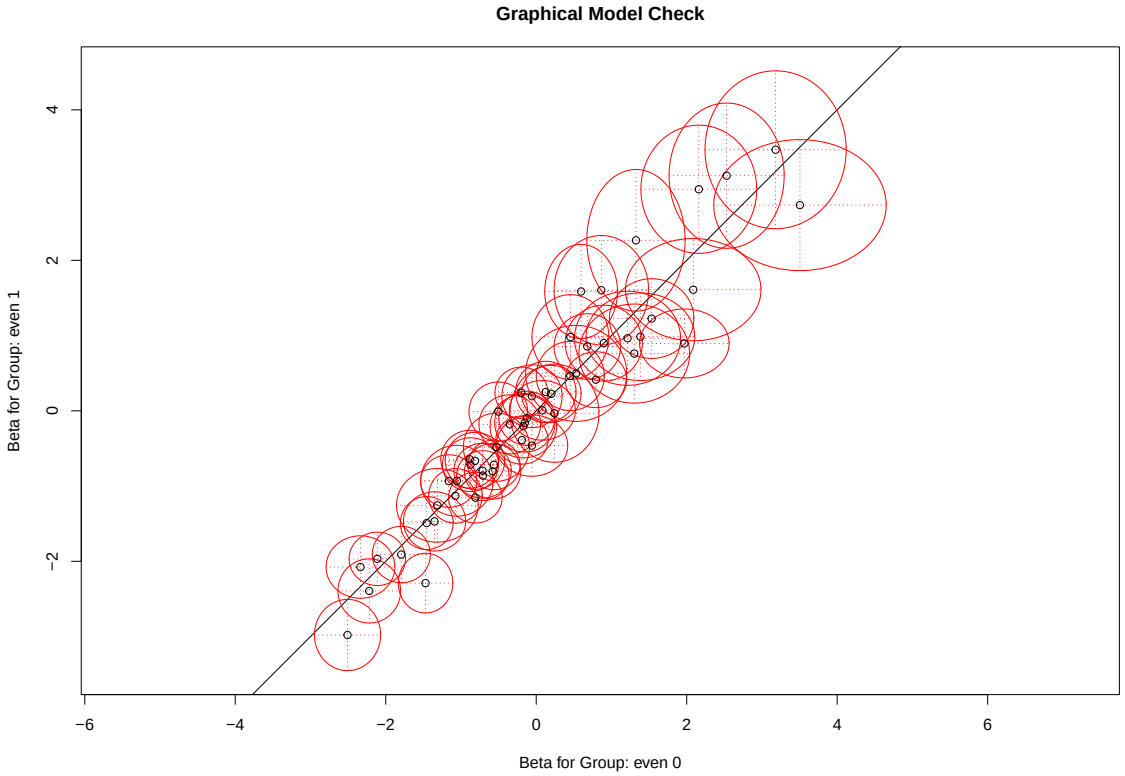
Ein Test sollte für alle Personen gleich fair sein. Beispielsweise sollte weiblichen Personen in einem Test kein Vorteil dadurch entstehen, dass sie weiblich sind. Dasselbe gilt auch z. B. für Personen, die zuhause vorwiegend deutsch sprechen oder Instrumental- oder Gesangsunterricht erhalten. Ob der Test für die jeweiligen Personengruppen fair war, wurde mit sog. *DIF-Analysen* analysiert (Kapitel 4.2). Diese prüfen die Fairness des Tests im Hinblick auf das Geschlecht, den häuslichen Sprachgebrauch sowie hinsichtlich des Sachverhalts, ob eine Testperson Instrumental- oder Gesangsunterricht erhielt. Das genaue Vorgehen für die Durchführung von DIF-Analysen wurde bereits in Kapitel 4.2 beschrieben. Für die DIF-Analysen wurde ein Facetten-Modell mit einem Interaktionsterm (‚item-by-group‘) berechnet. Laut den Richtlinien des Educational Testing Service (ETS) liegt ein vernachlässigbarer DIF bei einer Effektstärke ≤ 0.426 vor und ein

[37] Auf S. 55 in der vorliegenden Arbeit sind die verschiedenen empfohlenen Wertebereiche von Itemfit-Indizes dargestellt.

[38] 2.33 % (Q3) und 0.67 % (aQ3) aller 600 Item-Paare zeigten Werte über dem Cut-off Wert (> 0.2; Chen und Thissen, 1997). Der Mittelwert aller Q3- und aQ3-Werte lag fast bei null (Q3: $M = -0.05$; $SD = 0.07$; aQ3: $M < 0.01$; $SD = 0.07$).

[39] Eine Antwortkategorie in der Analyse mit dem Teilungskriterium Geschlecht zeigte eine Auffälligkeit (Item „Eisenbahn“, zweite Itemkategorie).

Abbildung 9.
Grafische Modellkontrolle mit einem zufälligen Teilungskriterium



Anmerkungen. Die Teilstichprobe wurde nach einem Zufallskriterium geteilt (gerade vs. ungerade Fallzahl). Die Abbildung zeigt ein Streudiagramm mit Konfidenzellipsen der getrennt geschätzten Itemschwierigkeitsparameter. Dieselbe Abbildung ist auch in Ehninger et al. (2021b, S. 5) abgedruckt.

starker DIF ab einer Effektstärke von ≥ 0.638 (Signifikanzniveau von .05) (Trendtel et al., 2016b, S. 127-131). Zwei Items hatten DIF (-0.69 Logits und -0.72 Logits) (Tabellen A.5, A.6, A.7 auf S. 117-119).[40] Beide Items wurden nicht entfernt, da die beobachteten Unregelmäßigkeiten vernachlässigbar waren.

Es konnten also alle 25 Testitems[41] für die Schätzung eines statistischen Modells verwendet werden, das die empirischen Daten beschreiben sollte. Lediglich acht Itemkategorien mussten zusammengefasst werden, da sie nicht häufig genug im Datensatz vorkamen (relative Häufigkeit < 5 %). Insgesamt gab es bei den Daten des *MARKO*-Tests 26.2 % fehlende Werte. Die Schätzung des statistischen Modells (Partial-Credit-Modell) erfolgte mit allen fehlenden Werten (keine Werte wurden imputiert).

[40] Das Item „Eisenbahn“ hatte DIF für männliche Testpersonen (-0.72 Logits) und das Item „America“ für Personen ohne Instrumentalunterricht (-0.69 Logits).
[41] 23 polytome und 2 dichotome Items

6.3. Statistisches Modell, Kompetenzniveaus und Kompetenzausprägung[42]

In diesem Kapitel beschreibe ich zunächst das finale statistische Modell, das am besten auf die erhobenen Daten passte. Dieses Modell wurde mit zwei anderen, theoretisch ableitbaren, statistischen Modellen verglichen. Aus dem finalen statistischen Modell konnten schließlich die Kompetenzniveau-Beschreibungen für musikbezogenes Argumentieren abgeleitet werden. Des Weiteren stellte das finale statistische Modell Informationen über die Kompetenzausprägungen der Versuchsteilnehmenden bereit. Im Laufe dieses Kapitels wird daher die Stichprobe in Hinblick auf die Kompetenzausprägungen (bzw. Personenfähigkeiten) der Teilnehmenden genauer untersucht.

Rolle ging in seinen theoretischen Annahmen davon aus, dass musikbezogene Argumentationskompetenz eine eindimensionale Kompetenz ist. Dies bedeutet, dass sich musikbezogene Argumentationskompetenz nicht in einzelne Teilkompetenzen bzw. einzelne Dimensionen unterteilen lässt. In Hasselhorns (2015) Arbeit zu musikpraktischen Kompetenzen beispielsweise teilte sich musikpraktische Kompetenz in die drei Dimensionen „Gesang“, „Instrumentales Musizieren“ und „Rhythmusproduktion“ auf. Eine solche Aufteilung von musikbezogener Argumentationskompetenz wäre laut Rolles theoretischen Annahmen nicht vorgesehen. So lag auch die Annahme, dass es sich bei musikbezogener Argumentationskompetenz um eine eindimensionale Kompetenz handelt, der Testentwicklung zugrunde. Dementsprechend wurde zunächst ein eindimensionales statistisches Modell geschätzt. Das statistische Modell, das die Daten beschreibt, ist das Partial-Credit-Modell (PCM) (Kapitel 4.1). Das eindimensionale PCM zeigte sehr gute Reliabilitätswerte. Die EAP/PV Reliabilität lag bei 0.91 und die WLE Reliabilität bei 0.90.

Um zu überprüfen, ob die Kompetenz eindimensional ist, wurden zusätzlich zum eindimensionalen Partial-Credit-Modell (Modell A), zwei zweidimensionale Modelle berechnet (Modelle B und C) und miteinander verglichen. Für die Berechnung von Modell B wurden vier Items einer zweiten Dimension zugeordnet, die den sozialen Kontext von Musik behandelten. In Modell C wurden einzelne Antwortkategorien von Items, die den sozialen Kontext von Musik thematisierten, einer zweiten Dimension zugeordnet.[43] Der Vergleich der Modelle A, B und C erfolgte mithilfe informationstheoretischer Maße (Kapitel 4.2). Den rechten fünf Spalten in Tabelle 4 kann man entnehmen, dass die informationstheoretischen Maße mit Ausnahme des *AIC* für Modell A die geringsten Werte aufwiesen. Ein Vergleich der informationstheoretischen Maße sprach daher für Modell A. Außerdem maßen die beiden Dimensionen in Modell B und C im Prinzip dieselbe Kompetenz (d. h. dieselbe latente Variable). Die beiden Dimensionen in Modell B und C korrelierten jeweils mit $r = .96$. Es ist also davon auszugehen, dass die untersuchte Kompetenz eindimensional ist. Dennoch wurden weitere denkbare Modelle mithilfe explorativer Faktorenanalysen untersucht

[42] Dieser Abschnitt ist eine Zusammenfassung von Ehninger et al. (2021b, S. 6). Indirekte Zitate sind aus Gründen der besseren Lesbarkeit nicht gekennzeichnet.

[43] Die genauen Modellspezifikationen können dem Abschnitt „Model Tests“ im *R*-Skipt entnommen werden (Link auf S. 17). Dort findet man auch die Zuordnung der Items zu den beiden Dimensionen.

(ebenfalls mit dem *R*-Paket *TAM*). Die Ergebnisse der explorativen Faktorenanalysen mit zwei bzw. drei Faktoren zeigten, dass die einzelnen Items den jeweiligen Faktoren nicht sinnvoll zugeordnet werden konnten. Diese Ergebnisse der Faktorenanalysen lieferten niedrige bzw. doppelte Ladungen der Items auf die Faktoren und waren daher nicht sinnvoll interpretierbar.[44] Das statistische Modell, das die Daten am besten beschrieb, war daher ein eindimensionales Partial-Credit-Modell (Modell A).

Tabelle 4.
Modellvergleich und Informationskriterien

Model	Loglike	Deviance	Npars	Nobs	AIC	BIC	AIC3	AICc	CAIC
Model A	-6720.02	13440.03	54	440	13548.03	13768.72	13602.03	13563.46	13822.72
Model B	-6641.70	13283.39	129	440	13541.39	14068.59	13670.39	13649.58	14197.59
Model C	-8421.70	16843.40	107	440	17057.40	17494.69	17164.40	17127.02	17601.69

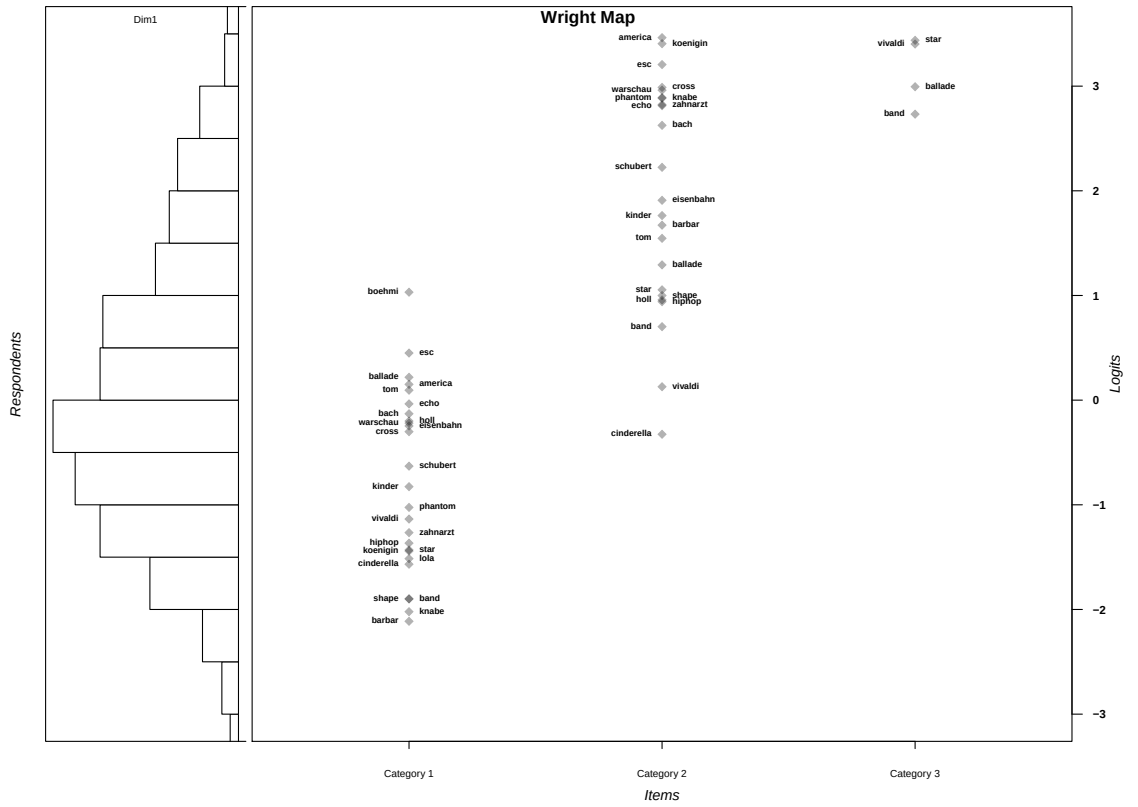
Anmerkungen. Modell A ist ein eindimensionales Partial-Credit-Modell. Modell B und C sind zweidimensionale Modelle. In Modell B wurden vier Items einer zweiten Dimension zugewiesen, in Modell C einzelne Itemkategorien. Dieselbe Tabelle ist in Ehninger et al. (2021b, S. 6) abgebildet.

Nachdem nun das eindimensionale Partial-Credit-Modell als finales statistisches Modell identifiziert worden war, konnten die Kompetenzniveaus beschrieben werden. Diese wurden in Anlehnung an die *Bookmark-Methode* bestimmt. Das genaue Vorgehen ist ausführlicher in Kapitel 4.3 beschrieben. Um die Kompetenzniveaus zu bestimmen, wurden zunächst die einzelnen Antwortkategorien der Testitems nach ihrer Schwierigkeit geordnet. Die Anordnung der einzelnen Itemkategorien nach Schwierigkeit kann Abbildung 10 entnommen werden. Hier sind die einzelnen Antwortkategorien aller Testitems vertikal nach ihrer Schwierigkeit geordnet (unten leicht, oben schwer) (Lösungswahrscheinlichkeit 65 %). Diese Darstellungsform wird auch *Wright Map* oder *Person-Item Map* genannt. Ausgehend von dieser Wright Map diskutierte das Projekt-Team, bestehend aus Christian Rolle, Jens Knigge und mir, welche Kompetenzen benötigt wurden, um eine bestimmte Anzahl von Punkten bei einer bestimmten Aufgabe zu bekommen. Die Kodierregeln der Testaufgaben enthielten bereits viele Informationen darüber, welche Kompetenzen benötigt wurden, um eine Aufgabe zu lösen. Wir besprachen so wiederholt jede einzelne Antwortkategorie einer Aufgabe. Bei welchen Aufgaben zeigte sich eine ‚neue' Kompetenz, die davor noch nicht beherrscht wurde? Auf diese Art und Weise wurden Schwellen zwischen den Kompetenzniveaus gesetzt.

Tabelle 5 zeigt das Ergebnis dieses Prozesses. Personen auf dem niedrigsten Kompetenzniveau A können ihre eigene Meinung zur präsentierten Musik äußern und beziehen sich in ihren Urteilen auf saliente musikalische Merkmale (z. B. „laut", „schnell"). Auf Niveau B verwenden Personen ebenfalls hauptsächlich saliente musikalische Merkmale, um sich auf die Musik zu beziehen. Allerdings können sie in ihrer Argumentation mehrere solcher Merkmale benutzen als Personen auf Niveau A. Außerdem können sie kausale Bezüge zwischen den salienten musikalischen Merk-

[44] Die Faktorenanalysen können im Abschnitt „model tests" im R-Skipt nachvollzogen werden (Link auf S. 17).

Abbildung 10.
Wright Map (Person-Item Map)



Anmerkungen. Die Itemschwierigkeitsparameter (Thurstonian Thresholds) sind auf derselben Skala wie die Personenfähigkeitsparameter (WLE) abgebildet. Die linke Seite der Abbildung zeigt ein Histogramm der Personenfähigkeitswerte und die rechte Seite die Schwierigkeit der einzelnen Itemkategorien (Category 1, 2, 3). Die Lösungswahrscheinlichkeit betrug 65 %. Dieselbe Grafik ist auch in Ehninger et al. (2021b, S. 7) abgedruckt.

malen und der Wirkung bzw. Funktion der Musik herstellen und verschiedene Standpunkte zur Musik wiedergeben. Personen sowohl auf Niveau A als auch auf Niveau B nehmen die Musik als ein ‚großes Ganzes' war. Die Musik wird ganzheitlich mithilfe salienter musikalischer Merkmale beschrieben. Ein deutlicher Bruch erfolgt dann auf Niveau C. Ab diesem Niveau können Personen detailliert auf die Musik Bezug nehmen und spezifische musikalische Merkmale anführen (z. B. die Spielweise von Instrumenten genauer beschreiben oder sich auf einzelne Formteile beziehen). Sie verfügen ebenfalls über grundlegende Kenntnisse über musikalische Normen und Genrekonventionen. Auf Niveau D können Personen schließlich verschiedene Standpunkte diskutieren und den sozialen und kulturellen Kontext der Musik berücksichtigen. Die Verteilung der Kompetenzniveaus innerhalb der Klassenstufen ist in Abbildung 11 dargestellt. Während knapp die Hälfte der Neuntklässler*innen Niveau A erreichte, war dieses Niveau bei den Zwölftklässler*innen sowie Studierenden gar nicht vertreten. Diese erreichten mehrheitlich Niveau C oder D.

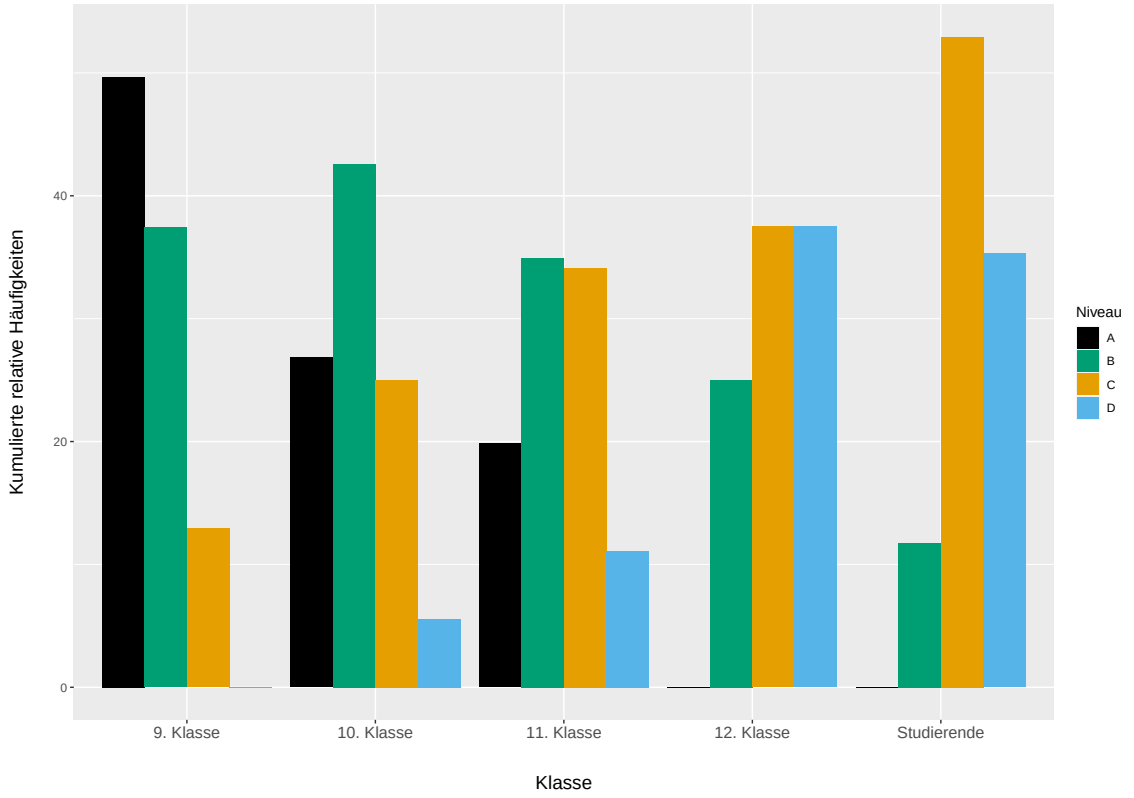
Im Partial-Credit-Modell wurden die Personenfähigkeiten, d h. die Kompetenzausprägungen der Testpersonen, geschätzt. Die Fähigkeit einer Person wird in statistischen Modellen der Item-Re-

Tabelle 5.
Beschreibungen der Kompetenzniveaus für musikbezogenes Argumentieren

Niv.	Logits	Kompetenzbeschreibung	Beispielantworten
D	> 2.22	*Personen verfügen über die Kompetenzen von Niveau A, B sowie C und können...* ... verschiedene Standpunkte diskutieren. ... sich in ihrer Urteilsbegründung auf musikbezogene Normen und Genrekonventionen beziehen. ... in ihrer Urteilsbegründung die sozialen und kulturellen Kontexte der präsentierten Musik berücksichtigen.	*„Women empowerment ist ein sehr aktuelles Thema, welches wichtig ist. Es ist gut das Künstlerinnen und Künstler ein Zeichen setzen. Teilweise ist der Text eindimensional, da auch Frauen mit Frauen "spielen". Oft ist es jedoch anders herum und schon seit Jahrhunderten der Fall, aufgrund der ungerechten Macht Verteilung, wo Frauen zu kurz kommen. Sie hätte vielleicht lieber singen sollen ‚I´m not a toy, for no one' oder so etwas in der Art, was mehr den Gleichberechtigungs Gedanken hervorbringt. Sie repräsentiert ein starkes Frauenbild, was definitiv Sozialkritisch ist. Durch das sogenannte "Hühner-Gegacker", wie Sascha es bezeichnet, ist das Lied ungewohnt und anders und unterscheidet sich von der gesellschaftlichen Norm, die die Masse beeinflusst, wie Sascha und 367 andere Personen zeigen. Viel Spaß mit eurem Mitläufertum und dem daraus resultiertem Einheitsbrei." (VP_89)*
C	≤ 2.22	*Personen verfügen über die Kompetenzen von Niveau A sowie B und können...* ... musikbezogene Urteile begründen unter detaillierter Bezugnahme auf spezifische musikalische Merkmale und diese mit der Wirkung und Funktion der Musik verknüpfen. ... sich in ihrer Urteilsbegründung auf grundlegende Kenntnisse über musikalische Normen und Genrekonventionen beziehen.	*„Ja, ich finde es sehr gelungen.* *die flächenartigen Klänge beschreiben die unendliche Weite des Universums* *die Syntheshizer geben dem Stück das futuristische des Weltalls* *einzelne hohe Töne zur verdeutlichung der Sterne" (VP_589)*
B	≥ 0.45	*Personen verfügen über die Kompetenzen von Niveau A und können...* ... verschiedene Standpunkte zur Musik wiedergeben. ... musikbezogene Urteile begründen unter Bezugnahme auf mehrere saliente musikalische Merkmale (z. B. Tempo, Dynamik, Intonation, Genremerkmale) und diese mit der Wirkung und Funktion der Musik verknüpfen.	*„Die Sängerin spricht ein sehr wichtiges und auch aktuelles Thema an. Also die soziale Gleichberechtigung, jedoch finde ich es nicht wirklich passend gewählt es den Mitmenschen rüberzubringen. Der Text wird humorvoll wiedergegeben und somit keinen Sinn erreicht." (VP_142)*
A	≤ –0,83	*Personen auf diesem Niveau können...* ... die eigene Meinung zur präsentierten Musik äußern. ... musikbezogene Urteile begründen unter Bezugnahme auf saliente musikalische Merkmale (z. B. Tempo, Dynamik, Intonation, Genremerkmale). ... sich in ihrer Urteilsbegründung auf die Wirkung und Funktion der Musik beziehen.	*„Ja, auf Grund der Atmosphäre die im Weltraum herrscht. Dies hat der Komponist sehr gut daargestellt." (VP_661)*

Anmerkungen. Die Tabelle wird von unten nach oben gelesen. A ist das niedrigste Kompetenzniveau, D das höchste. Die Spalte „Logits" zeigt die erforderlichen Personenfähigkeitswerte (WLE; 65 %ige Lösungswahrscheinlichkeit). Die Beispielantworten sind den Aufgaben „Star Wars" und „Eurovision Song Contest" entnommen (Kapitel 5.2.1 und 5.2.2). Die Rechtschreibfehler der Testpersonen wurden nicht verbessert. Dieselbe Tabelle ist auf Englisch in Ehninger et al. (2021b, S. 7) abgebildet.

Abbildung 11.
Relative Häufigkeitsverteilung der Kompetenzniveaus

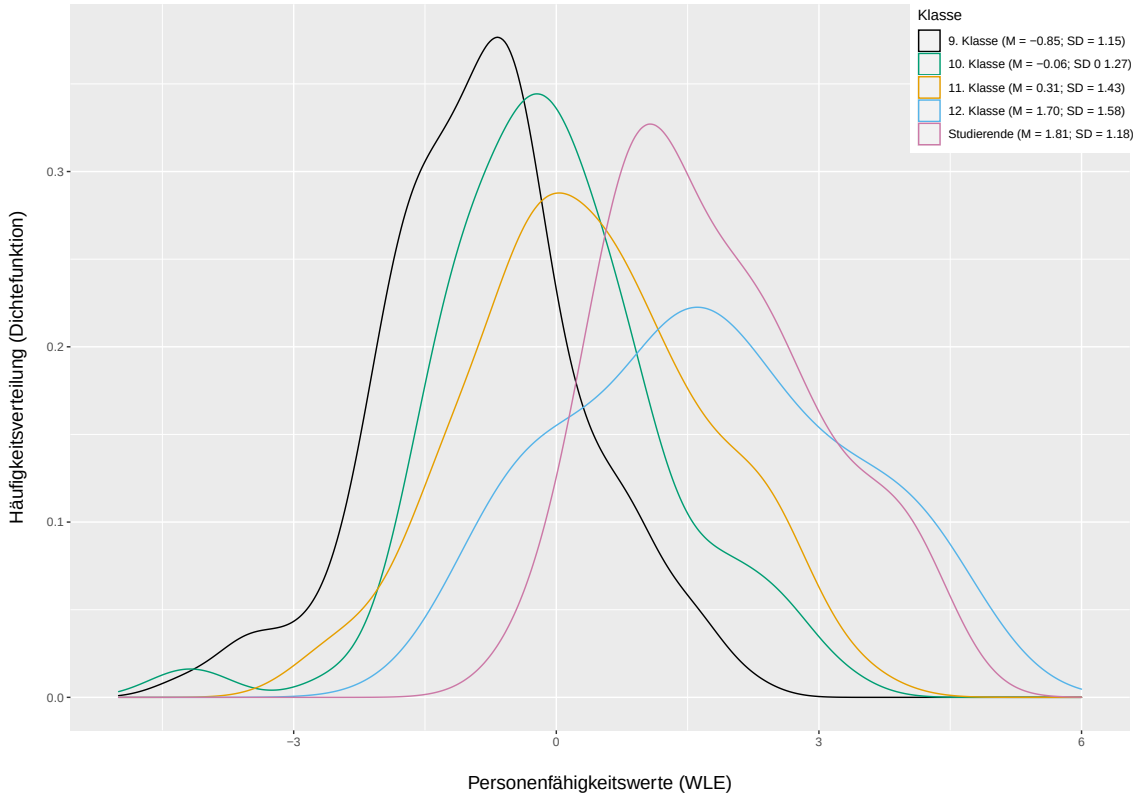


Anmerkungen. Die Abbildung zeigt die relative Häufigkeitsverteilung der Kompetenzniveaus gruppiert nach Klassenstufen ($n = 439$; 9. Klasse: $n = 147$; 10. Klasse: $n = 108$; 11. Klasse: $n = 126$; 12. Klasse: $n = 24$; Studierende: $n = 34$). Dieselbe Grafik ist auch in Ehninger et al. (2021b, S. 8) abgedruckt.

sponse-Theorie (IRT) mit Personenfähigkeitswerten bzw. -parametern (WLE) angegeben (Kapitel 4.1). Die Verteilung dieser Personenfähigkeitswerte innerhalb der einzelnen Klassenstufen ist in Abbildung 12 dargestellt. Höhere Fähigkeitswerte weisen darauf hin, dass Personen fähiger sind als Personen mit niedrigeren Fähigkeitswerten. Innerhalb der vorliegenden Stichprobe sind erwartungsgemäß die Personenfähigkeitswerte der Studierenden am höchsten. Der Abbildung kann man auch entnehmen, dass die Personenfähigkeitswerte der 12. Klasse recht breit verteilt sind. Zwischen den einzelnen Klassenstufen variierten die Personenfähigkeitswerte signifikant, $F(4,434) = 44.85; p < .01; \eta^2 = 0.29$. Post-hoc-Analysen bestätigten die Unterschieden zwischen den Klassenstufen. Abgesehen von zwei Ausnahmen (10. Klasse vs. 11. Klasse; 12. Klasse vs. Studierende) waren die Ergebnisse signifikant mit mittleren bis großen Effektgrößen ($0.66 \leq \delta \geq 2.30$; Tabelle A.8 auf S. 119).

Zudem wurden die Personenfähigkeiten verschiedener Gruppen innerhalb der Stichprobe verglichen (Tabelle 6). Weibliche Personen schnitten bei der Testung besser ab als männliche Teilnehmer, $t(418.58) = -4.21, p < .01, \delta = 0.41$. Personen mit Instrumental- oder Gesangsunterricht erzielten ebenfalls deutlich bessere Ergebnisse im Kompetenztest, $t(319.66) = -6.74, p < .01, \delta = 0.68$. Dasselbe galt für Personen, die ein Musikinstrument spielten, $t(437.79) = -6.99, p < .01, \delta =$

Abbildung 12.
Verteilung der Personenfähigkeitswerte (WLE) zwischen den Klassenstufen



Anmerkungen. Personen mit hohen Fähigkeitswerten verfügen über mehr Fähigkeiten als Personen mit geringen Fähigkeitswerten. Der niedrigste Personenfähigkeitswert in der Stichprobe lag bei -5.40, der höchste bei 4.31. In der Legende der Abbildung sind die Mittelwerte (M) und Standardabweichungen (SD) der Personenfähigkeitswerte angegeben. Die Grafik ist auch in Ehninger et al. (2021b, S. 8) abgedruckt.

0.66. Versuchsteilnehmende, die zuhause immer deutsch sprachen, schnitten ebenfalls besser ab, $t(419.13) = 4.42, p < .01, \delta = 0.42$. Personen, deren Elternteile beide in deutschland geboren worden waren, erzielten ebenfalls bessere Personenfähigkeitswerte, $t(418.25) = 5.47, p < .01, \delta = .52$. Wobei die Migrationsgeschichte der Versuchspersonen mit dem häuslichem Sprachgebrauch korrelierte, $r = .63$, $p < .01$.

6.4. Weiterführende Analysen

(zuvor nicht veröffentlicht)

In diesem Kapitel werden zuvor unveröffentlichte Analysen zu mehreren Variablen vorgestellt, die einen Einfluss auf die Ausprägung musikbezogener Argumentationskompetenz haben. Verschiedene Studien zu musikbezogener Kompetenzforschung zeigten, dass Schüler*innen mit musikalischer Vorerfahrung häufig besser in Kompetenztests abschneiden (s. a. Kapitel 1.2). Dieser Sachverhalt wurde auch in der vorliegenden Studie festgestellt (Kapitel 6.3). Die Analysen im

Tabelle 6.
t-*Tests sowie Mittelwerte und Standardabweichungen der Personenfähigkeitswerte (WLE)*

		n	*M*	*SD*	*t*	*p*	Cohens *d*
Geschlecht	m	230	-0.18	1.48	-4.21	<.01	0.41
	w	196	0.41	1.41			
Instrumental- oder Gesangsunterricht	nein	279	-0.34	1.43	-6.74	< .01	0.68
	ja	161	0.65	1.50			
Musikinstrument spielen	nein	203	-0.49	1.31	-6.99	< .01	0.66
	ja	237	0.46	1.56			
Zuhause immer deutsch sprechen	nein	181	-0.35	1.37	-4.42	< .01	0.42
	ja	259	0.28	1.58			
Beide Elternteile in D geboren	nein	184	-0.39	1.35	5.47	< .01	0.52
	ja	242	0.38	1.58			

Anmerkungen. Werte von Cohens *d* ab 0.20 werden als kleiner Effekt interpretiert, Werte ab 0.50 als mittlerer Effekt und Werte größer 0.80 als großer Effekt (Döring & Bortz, 2016, S. 820).

vorangegangenen Kapitel zeigten ebenfalls, dass weibliche Testpersonen besser abschnitten als männliche. Heß (2018) sowie Fiedler und Hasselhorn (2020) stellten fest, dass Mädchen häufig über ein höheres musikalisches Selbstkonzept verfügen als Jungen. Die Variable Geschlecht könnte demnach durchaus konfundiert sein und mit der musikalischen Erfahrenheit der Versuchspersonen zusammenhängen.

In Kapitel 6.4.1 wird ein Pfadmodell mit den Einflussfaktoren Geschlecht, Klassenstufe, Sprache zuhause und musikalische Erfahrenheit vorgestellt. Der Einfluss von Musikpräferenzen und der Vertrautheit mit einer bestimmten Art von Musik ist in Form explorativer Analysen Gegenstand von Kapitel 6.4.2. Abschließend wird die je nach Kompetenzniveau unterschiedliche Wortanzahl pro Aufgabe sowie die Anzahl der bearbeiteten Aufgaben diskutiert (Kapitel 6.4.3).

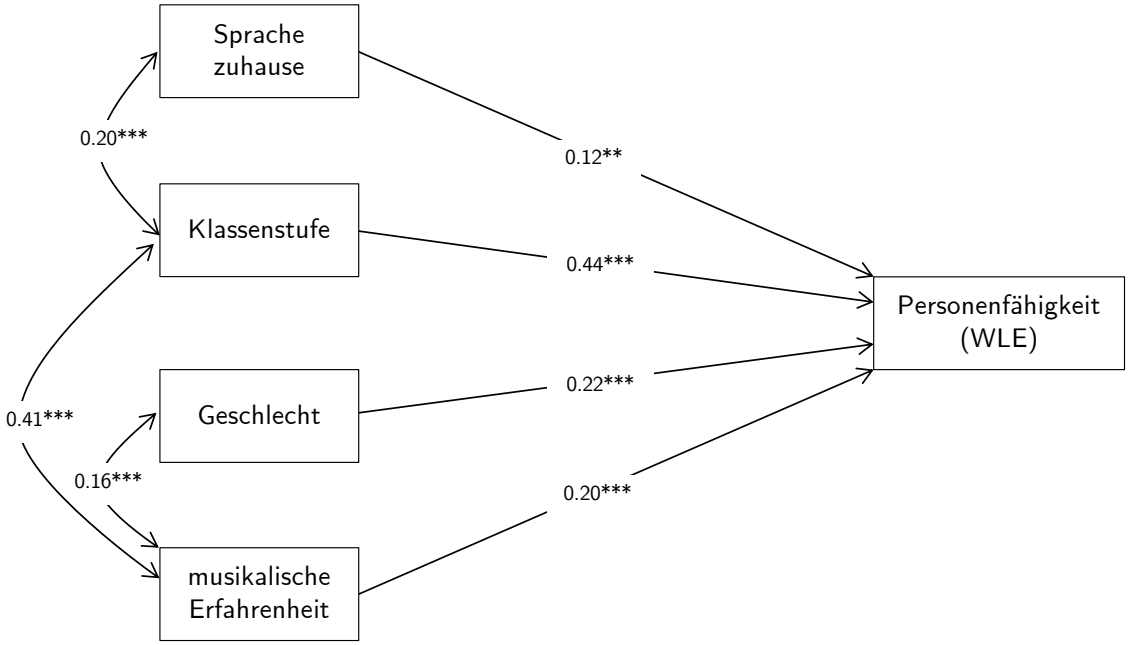
6.4.1. Einflussfaktoren Geschlecht, Klassenstufe, Sprache zuhause und musikalische Erfahrenheit

Im vorherigen Kapitel wurde bereits festgestellt, dass sowohl weibliche Personen als auch Personen mit Instrumental- oder Gesangsunterricht signifikant besser im Test abschnitten. Dasselbe galt für Personen aus höheren Klassenstufen und Personen, die zuhause meist deutsch sprachen. Bisher ist allerdings nicht klar, wie diese Einflussfaktoren zusammenhängen. Möglicherweise waren weibliche Testpersonen musikalisch erfahrener und schnitten deshalb besser im Test ab. Dann würde das positive Testergebnis vermutlich nicht mit dem Geschlecht zusammenhängen, sondern mit der musikalischen Erfahrenheit der Testperson. Um die Zusammenhänge zwischen den Einflussfaktoren zu untersuchen, wurde in einem explorativen Vorgehen ein Pfadmodell mit den Einflussfaktoren Geschlecht, Klassenstufe, Sprache zuhause und musikalische Erfahrenheit geschätzt. Die Schätzung erfolgte mithilfe des *R*-Pakets *lavaan* (Rosseel, 2012). In diesem Pfadmodell wird beschrieben, inwiefern die verschiedenen Einflussfaktoren die Personenfähigkeit der Testpersonen

vorhersagen. Außerdem kann im Pfadmodell analysiert werden, wie die Einflussfaktoren untereinander zusammenhängen.

Die musikalische Erfahrenheit der Testpersonen wurde mit einer Subskala des Gold-MSI erfasst (Müllensiefen et al., 2014; Schaal et al., 2014). Diese Subskala „General Musical Sophistication“ bestand aus 18 Items. Ein Item musste aus der Skala entfernt werden, weil es nicht ausreichend mit dem Gesamtscore der Skala korrelierte.[45] Die verbleibenden 17 Items bildeten die Skala für musikalische Erfahrenheit, $\alpha = .87$. Der Mittelwert der Skala ging als Einflussfaktor „musikalische Erfahrenheit“ in die Schätzung des Pfadmodells mit ein. Einflussfaktoren werden im Folgenden als *Prädiktoren* bezeichnet. Das Pfadmodell mit den vier Prädiktoren „Klassenstufe“, „Geschlecht“, „Sprache zuhause“ und „musikalische Erfahrenheit“ passte gut zu den erhobenen Daten, $RMSEA = 0.023$, $CFI = 0.998$, $TLI = 0.993$, $\chi^2 = 3.645$, $df = 3$, $p = .30$.

Abbildung 13.
Pfaddiagramm mit dem Personenfähigkeitsparameter (WLE) und den Prädiktoren „Klassenstufe“, „Geschlecht“, „Sprache zuhause“ und „musikalische Erfahrenheit“



Anmerkungen. Modellfit und Varianzaufklärung: $RMSEA = 0.023$, $CFI = 0.998$, $TLI = 0.993$, $\chi^2 = 3.645$, $df = 3$, $p = .30$, $R^2 = 0.40$. Im Pfadmodell werden nur signifikante Effekte berichtet. Die Variable „Sprache zuhause“ hatte vier Kategorien (4 = immer zuhause deutsch sprechen, 1 = nie zuhause deutsch sprechen). * $p < .05$, ** $p < .01$, *** $p < .01$.

Abbildung 13 zeigt, dass die Personenfähigkeit (WLE) der Testpersonen erwartungsgemäß durch alle vier Einflussfaktoren vorhergesagt werden konnte, $p < .01$. In der Abbildung sind nur signi-

[45] Das Item 25 „Ich singe nicht gerne in der Öffentlichkeit, weil ich Angst habe, falsche Töne zu treffen“ aus der Subskala „General Musical Sophistication“ korrelierte nur mit $r = .18$ mit dem Gesamtwert der Skala. Die Berechnungen können im *R*-Skript in Abschnitt „Correlation General Musical Sophistication and Proficiency Scores“ nachvollzogen werden (Link auf S. 17).

fikante Zusammenhänge dargestellt. Das Modell klärte einen Teil der beobachteten Varianz auf, $R^2 = 0.40$, wobei Klassenstufe der stärkste Prädiktor war, $\beta = 0.44$, $p < .001$. Die musikalische Erfahrenheit hing dabei deutlich mit der Klassenstufe zusammen, $r = 0.41$, $p < .001$. Personen aus höheren Klassenstufen waren demnach meist musikalisch erfahrener. Die musikalische Erfahrenheit korrelierte ebenfalls mit dem Geschlecht der Testpersonen, $r = 0.16$, $p < .001$. Ein auffälliger Zusammenhang bestand ebenfalls zwischen der Klassenstufe und der Sprache, die die Testpersonen zuhause sprachen, $r = 0.20$, $p < .001$. Demnach sprachen die älteren Schüler*innen bzw. die Studierenden häufiger zuhause deutsch. Der hohe Einfluss der Klassenstufe ging also einher mit der musikalischen Erfahrenheit und der Tatsache, dass eher deutsch zuhause gesprochen wurde. Die Ergebnisse des Pfadmodells verdeutlichen, dass die einzelnen Prädiktoren konfundiert sind. Ältere Schüler*innen waren demnach musikalisch erfahrener und sprachen zuhause eher deutsch. Die Ergebnisse zeigen also in erster Linie ein ‚Stichprobenartefakt' auf.

6.4.2. Musikpräferenzen und Vertrautheit mit Musik

Dieses Kapitel geht der Frage nach, ob das Gefallen und die Vertrautheit mit einer Musik einen Einfluss auf das Testverhalten der Versuchsteilnehmenden hatten. Im Rahmen der Datenerhebungen wurden klingende Musikpräferenzen von sechs Hörbeispielen erhoben, die Gegenstand der Testaufgaben waren. Zusätzlich zu den Musikpräferenzen wurden die Testpersonen gefragt, wie vertraut sie mit dieser Art von Musik waren. Die Musikpräferenzen sowie die Vertrautheit mit einer Art von Musik wurden mithilfe einer 5-stufigen Likert-Skala erhoben (1 = sehr gut, 5 = sehr schlecht). Es wäre mitunter denkbar, dass eine Person, die besonders vertraut mit dem Hörbeispiel der „Star Wars"-Aufgabe ist, bei der Aufgabe besser abschneidet als eine Person, die weniger vertraut mit dem Hörbeispiel ist. Deshalb wurde in den folgenden Analysen untersucht, ob das Vertrautheits- bzw. Präferenzurteil einen Einfluss auf das Testverhalten bei der zugehörigen Aufgabe hatte. Der Zusammenhang wurde mithilfe von Korrelationen und ordinal logistischen Regressionsanalysen überprüft. An dieser Stelle ist hervorzuheben, dass die nachfolgenden Analysen explorativer Natur sind, da Präferenzangaben lediglich für sechs von insgesamt 25 Hörbeispielen vorliegen.

Zunächst wurden die Korrelationen zwischen Itemscore und dem jeweiligen Vertrautheits- bzw. Gefallensurteil betrachtet (Tabelle 7). Es fällt auf, dass bei den beiden Aufgaben mit recht aktuellen Hörbeispielen aus der populären Musik („America" und „Shape")[46] praktisch kein Zusammenhang zwischen Itemscore und Vertrautheits- bzw. Präferenzurteil besteht. Bei den anderen Aufgaben kann ein geringer Zusammenhang beobachtet werden, wobei dieser beim Item „Schubert" am stärksten ausgeprägt war.

Um dem Zusammenhang zwischen dem jeweiligen Itemscore und dem Gefallens- bzw. Vertrauensurteil weiter auf den Grund zu gehen, wurden ordinale Regressionen durchgeführt. In insge-

[46] Bei der Aufgabe „America" erklang „This is America" von Childish Gambino (a.k.a. Donald Glover) und bei der Aufgabe „Shape" „Shape of You" von Ed Sheeran.

Tabelle 7.
Korrelationen Itemscore und Vertrautheits- bzw. Gefallensurteil

	Korrelation mit Itemscore	
Item	**Vertrautheit**	**Präferenz**
America	– 0.03	0.13*
Ballade	0.23**	0.13*
Eisenbahn	0.16**	0.23**
Schubert	0.40**	0.29**
Shape	0.06	0.09
Vivaldi	0.26**	0.21**

Anmerkungen. * $p < .05$,** $p < .01$

samt zwölf Analysen wurde der Itemscore von sechs Items jeweils durch eine Prädiktorvariable (Gefallens- bzw. Vertrauensurteil) vorhergesagt. Die jeweiligen Regressionsmodelle passten jedoch nicht gut zu den beobachteten Daten. Im ordinalen Regressionsmodell wird angenommen, dass die Beziehungen zwischen den ordinalen Stufen der Zielvariable statistisch gleich sind („proportional odds“ oder „equal slopes“) (Schlarmann & Galatsch, 2014). Diese Voraussetzung war nicht gegeben,[47] weshalb ordinale Regressionsanalysen an dieser Stelle keinen weiteren Aufschluss über den Zusammenhang zwischen Itemscore und Präferenz- sowie Vertrautheitsurteil geben konnten.

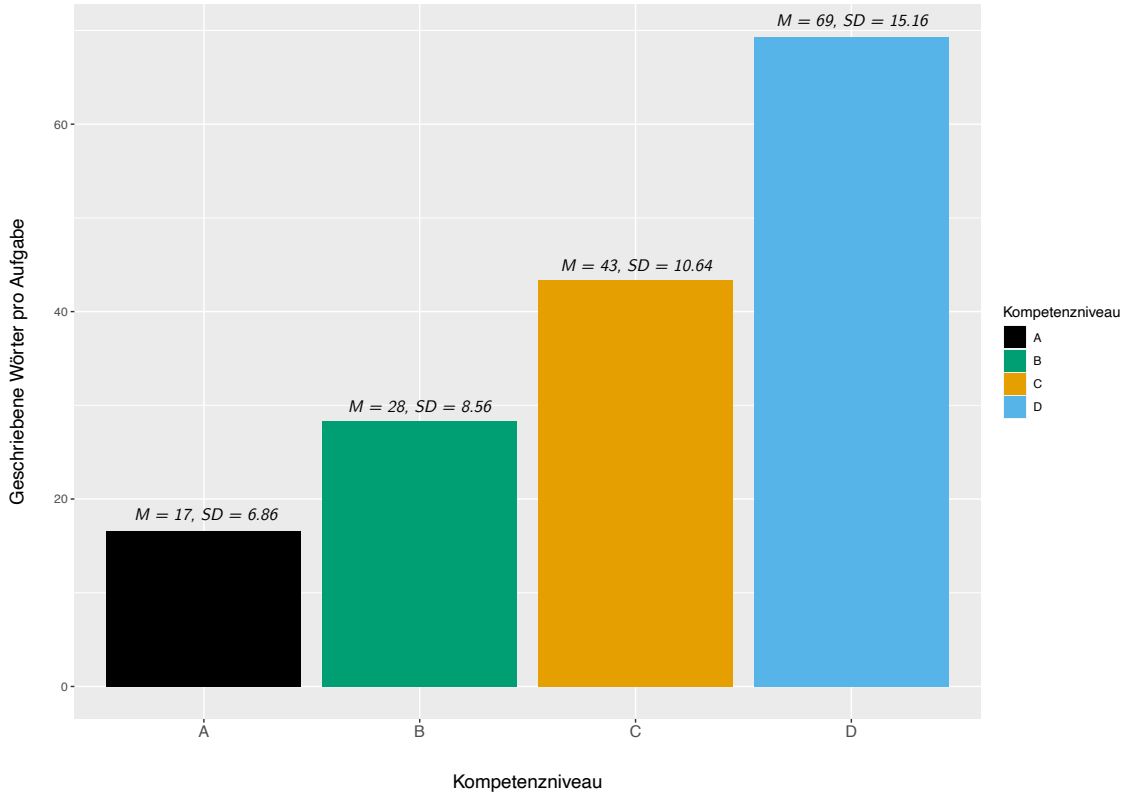
6.4.3. Bearbeitungsdauer und Wortanzahl

Ausgehend von den bisher vorgestellten Beispielantworten kann man bereits erahnen, dass die Testantworten, die die höchste Anzahl an Punkten bekamen, deutlich länger waren als die Antworten, die weniger Punkte bekamen (Tabelle 2 und 3 in Kapitel 5.2). Vermutlich brauchte eine Person mehr Zeit, um eine solche längere Antwort zu produzieren. Diese Vermutung wird dadurch gestützt, dass die Bearbeitungsdauer des Tests innerhalb der Stichprobe stark variierte. Einige Personen bearbeiteten während der Datenerhebung alle 25 Items, während andere Personen sich mit weniger Items beschäftigten.[48] Deshalb wird in diesem Kapitel der Zusammenhang zwischen der beobachteten Personenfähigkeit und der Bearbeitungsdauer des Tests sowie der Anzahl der geschriebenen Wörter und der Anzahl der bearbeiteten Aufgaben genauer untersucht.

Abbildung 14 zeigt, dass sich die geschriebenen Wörter pro Aufgaben zwischen den einzelnen Kompetenzniveaus stark voneinander unterscheiden. Während Personen auf dem höchsten Niveau D im Durchschnitt 69 Wörter schrieben, produzierten Personen auf Niveau A lediglich 17 Wörter. Es konnte ein signifikanter starker Effekt bei dieser Analyse beobachtet werden,

[47] s. Abschnitt „Music Preferences and Familiarity“ im *R*-Skript (Link auf S. 17 in der vorliegenden Arbeit).
[48] Aus rechentechnischen Gründen gingen nur die Daten von Personen in die Analysen ein, die mindestens acht Items bearbeitete hatten (s. a. S. 69 in der vorliegenden Arbeit).

Abbildung 14.
Balkendiagramm zu den geschriebenen Wörtern pro Aufgabe gruppiert nach Kompetenzniveaus

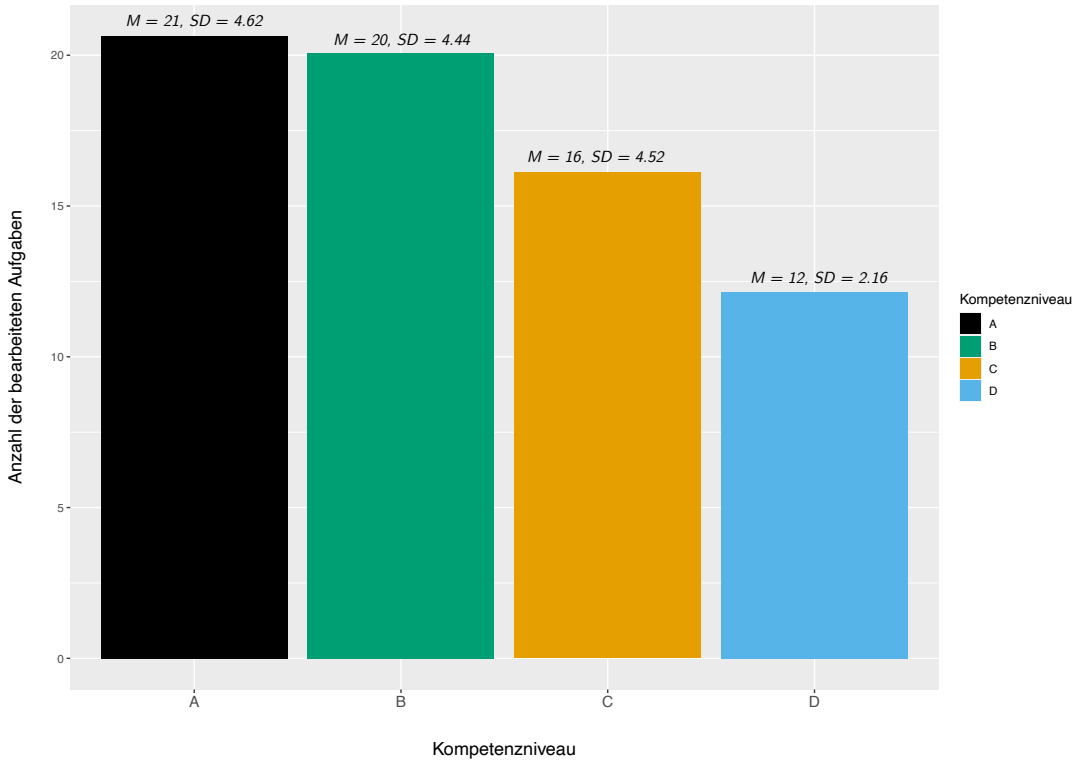


$F(3, 144.27) = 295.045$, $p < .001$, $\eta^2 = 0.73$.[49]

Je nach Kompetenzniveau unterschied sich auch die Anzahl der bearbeiteten Aufgaben im gesamten Test. Während Personen auf Niveau D im Schnitt lediglich 12 Aufgaben bearbeiteten, widmeten sich Personen auf Niveau A 21 Aufgaben. Abbildung 15 zeigt die Anzahl der bearbeiteten Aufgaben gruppiert nach Kompetenzniveaus. Auch hier unterschieden sich die Gruppen signifikant, der beobachtete Effekt war jedoch klein, $F(3, 194.22) = 120.531$, $p < .001$, $\eta^2 = 0.28$.[50] Die Gruppen differierten ebenfalls signifikant in Hinblick auf die durchschnittliche Bearbeitungszeit, $F(3, 151.19) = 77.926$, $p < .004$, $\eta^2 = 0.36$.

[49] Die Varianzen der vier untersuchten Gruppen waren nicht gleich (Levene-Test $p < .01$), weshalb eine Welch-ANOVA gerechnet wurde.
[50] Die Varianzen der vier untersuchten Gruppen waren nicht gleich (Levene-Test $p < .01$), weshalb eine Welch-ANOVA gerechnet wurde.

Abbildung 15.
Balkendiagramm zur Anzahl der bearbeiteten Aufgaben gruppiert nach Kompetenzniveaus



7. Diskussion

Argumentationskompetenz ist als sprachliche Schlüsselkompetenz zentral für den Lernerfolg von Schüler*innen (Quasthoff et al., 2020b). Obwohl die Rolle von sprachlichen und argumentativen Fähigkeiten in den letzten Jahren in der empirischen Bildungsforschung verstärkt untersucht wurde (u. a. Erath et al., 2018; Frederking et al., 2012; Krelle & Willenberg, 2008; Neumann & Lehmann, 2008), gibt es in der Musikpädagogik bisher kaum empirische Arbeiten in diesem Bereich (Bossen, 2017). Die vorliegende Arbeit reagiert auf dieses Desiderat und verfolgte dabei zwei wesentliche Ziele (s. a. Kapitel 3):

(a) Theoriegeleitete Entwicklung eines Kompetenztests für musikbezogenes Argumentieren,

(b) Kompetenzmessung und Beschreibung von Kompetenzniveaus für musikbezogenes Argumentieren auf Basis der empirischen Daten, die im Kompetenztest erhoben wurden.

Im Rahmen einer Pilotstudie wurde der Kompetenztest (Ziel (a)) entwickelt (Kapitel 5). Der Testentwicklung lagen Annahmen eines theoretischen Kompetenzmodells für musikbezogenes Argumentieren zugrunde (Rolle, 2013, 2017; s. a. Kapitel 2.4). Musikbezogene Argumentationskompetenz konnte schließlich zuverlässig gemessen werden und so konnten Kompetenzniveaus für musikbezogenes Argumentieren ausgehend von den empirischen Daten formuliert werden (Ziel (b); Kapitel 6).

Im nachfolgenden Kapitel 7.1 fasse ich die zentralen Ergebnisse der vorliegenden Arbeit zusammen. Anschließend diskutiere ich in Kapitel 7.2 den wissenschaftlichen Erkenntnisgewinn und skizziere Zusammenhänge zum Verhältnis ästhetischer Wahrnehmung und musikbezogener Argumentationskompetenz (Kapitel 7.3). In Kapitel 7.4 werden schließlich die Limitationen der empirischen Studie dargestellt. Zum Schluss ziehe ich ein Fazit und gebe einen Ausblick (Kapitel 7.5).

7.1. Zusammenfassung der Ergebnisse

Eines der wesentlichen Ziele der vorliegenden Dissertation war die Entwicklung des *MARKO*-Kompetenztests für musikbezogenes Argumentieren (Ziel (a)). Als theoretischer Ausgangspunkt für die Entwicklung von Testaufgaben diente Rolles (2013, 2017) Kompetenzmodell. Zunächst wurden in einer Pilotstudie 60 Testaufgaben erprobt ($N = 391$) und an der anschließenden Hauptstudie nahmen 440 Schüler*innen der neunten bis zwölften gymnasialen Jahrgangsstufe teil. Der finale Test bestand aus 25 offenen Items und erfüllte wesentliche psychometrische Gütekriterien. So lagen mitunter die Interrater-Reliabilität sowie Itemfit-Werte in einem guten bis sehr guten Bereich. Es konnte Personenhomogenität für die Stichprobe festgestellt werden und DIF-Analysen zeigten keine größeren Auffälligkeiten (Kapitel 6.2). Ein Modellvergleich ergab, dass ein eindimensionales Partial-Credit-Modell am besten zu den erhobenen Daten passte (EAP/PV Reliabilität = 0.91). Ausgehend von diesem Modell konnten schließlich Personenfähigkeitswerte

sowie Kompetenzniveaus für musikbezogenes Argumentieren identifiziert werden. Mithilfe der Bookmark-Methode wurden vier Kompetenzniveaus bestimmt (Tabelle 5 auf S. 76). Auf dem niedrigsten Niveau A können Personen ihre Meinung zu Musik äußern und sich auf saliente musikalische Merkmale wie beispielsweise „laut“ oder „schnell“ beziehen. Auf Niveau B stellen Personen zusätzlich Bezüge zwischen den salienten musikalischen Merkmalen sowie der Wirkung bzw. Funktion der Musik her und geben verschiedene Standpunkte zur Musik wieder. Ab Niveau C nehmen Personen schließlich mithilfe spezifischer musikalischer Merkmale detailliert auf die Musik Bezug, indem sie z. B. auf die Spielweise von Musikinstrumenten eingehen oder Normen und Genrekonventionen berücksichtigen. Personen auf Niveau D können schließlich verschiedene Standpunkte zur Musik diskutieren und den sozialen wie kulturellen Kontext in ihrer Argumentation berücksichtigen.

In der vorliegenden Stichprobe waren Schüler*innen der 12. Klasse sowie Musikstudierende hauptsächlich auf Niveau C oder D vertreten, während sich Neuntklässler*innen mehrheitlich auf Niveau A und B wiederfanden. Personen aus höheren Klassenstufen sowie Studierende schnitten also besser im Test ab als Personen aus niedrigeren Klassenstufen und die Personenfähigkeitswerte variierten signifikant mit teilweise mittleren bis großen Effektstärken. Weibliche sowie musikalisch erfahrene Versuchsteilnehmende und Personen, die immer Deutsch zuhause sprachen, schnitten ebenfalls besser ab. Personen, die den höheren Kompetenzniveaus zuzuordnen waren, produzierten insgesamt längere Testantworten als Personen niedrigerer Kompetenzniveaus (Kapitel 6.4). Ebenfalls war die Bearbeitungsdauer des gesamten Tests bei kompetenteren Personen länger.

Um der Frage nachzugehen, welche Variablen die Personenfähigkeitswerte beeinflussen, wurde in einem explorativen Vorgehen eine Pfadanalyse mit den Prädiktoren „Geschlecht“, „Klassenstufe“, „Sprache zuhause“ und „musikalische Erfahrenheit“ geschätzt. Die Personenfähigkeit konnte durch alle vier Prädiktoren vorhergesagt werden. Das Pfadmodell passte gut zu den erhobenen Daten und klärte 40 % der Varianz auf (Kapitel 6.4.1). Die Ergebnisse der Pfadanalysen verdeutlichen jedoch in erster Linie, dass die jeweiligen Variablen konfundiert waren. Die Klassenstufe hatte zwar den stärksten Einfluss auf die Ausprägung der Personenfähigkeit, jedoch waren Versuchspersonen aus höheren Klassenstufen musikalisch erfahrener und sprachen zuhause eher deutsch. Ob die Vertrautheit mit einer bestimmten Art von Musik und das Gefallen eines Musikstücks das Testverhalten bei einem bestimmten Item beeinflusste, konnte nicht abschließend geklärt werden (Kapitel 6.4.2).

Die beiden zentralen Ziele des Dissertationsvorhabens konnten somit erreicht werden: Die Entwicklung eines Kompetenztests für musikbezogenes Argumentieren war erfolgreich. Die Kompetenz konnte unter Berücksichtigung psychometrischer Gütekriterien zuverlässig gemessen werden und die Kompetenzniveaubeschreibungen geben Aufschluss über wesentliche Merkmale der Kompetenz.

7.2. Erkenntnisgewinn und Beitrag zum wissenschaftlichen Forschungsstand

7.2.1. Beitrag zur musikbezogenen Kompetenzmodellierung

Die vorliegende Arbeit leistet einen wichtigen Beitrag zur Erforschung musikbezogener Kompetenzen. Auf Basis des *MARKO*-Kompetenztests sowie -modells sind erstmals differenzierte Aussagen über den Leistungsstand von Schüler*innen und Studierenden in Bezug auf musikbezogenes Argumentieren möglich. Damit trägt die Dissertation wesentlich dazu bei, die Modellierung und Messung musikbezogener Kompetenzen weiterzuentwickeln. Die vorliegende Arbeit legt außerdem erstmals empirisch belastbare Befunde zur Struktur musikbezogener Argumentationskompetenz vor. Das Kompetenzmodell beschreibt somit ein Gefüge an Anforderungen, deren Bewältigung beim Argumentieren über Musik erwartet wird. Anhand des Tests und Modells kann der Leistungsstand einer Person erhoben werden und es kann ermittelt werden, welche Kompetenzen dezidiert gefördert werden müssen, um die nächsthöhere Kompetenzstufe zu erreichen. So können die Ergebnisse der vorliegenden Studie einen wichtigen Beitrag zur konzeptionellen Weiterentwicklung von Musikunterricht und schulischen Curricula leisten. Ebenfalls bieten sie einen Ausgangspunkt für mögliche Interventionsstudien zur Förderung musikbezogener Argumentationskompetenz. Zudem könnten mit dem Testinstrument in zukünftigen Studien wichtige Zusammenhänge zu anderen Kompetenzkonstrukten überprüft werden. So könnten bei einer zukünftigen Datenerhebung mehrere Erhebungsinstrumente eingesetzt werden, wie etwa ein Kompetenztest für musikalische Wahrnehmungsfähigkeiten oder ein Test für sprachliche Kompetenzen. Im Anschluss könnte untersucht werden, inwiefern die untersuchten Kompetenzen zusammenhängen.

Die Ergebnisse der vorliegenden Arbeit zeigen, dass musikbezogene Argumentationskompetenz als eindimensionale Kompetenz aufgefasst werden kann. Dies bedeutet, dass sich musikbezogene Argumentationskompetenz nicht in mehrere Teilkompetenzen unterteilen lässt. Zwar wurde musikbezogene Argumentationskompetenz auch in Rolles theoretischem Modell als eindimensionale Kompetenz angenommen, dennoch wäre es denkbar gewesen, dass sich musikbezogene Argumentationskompetenz in mehrere Teilkompetenzen untergliedern lässt, so wie dies in den beiden Forschungsprojekten *KoMus* und *KOPRA-M* der Fall war (Hasselhorn, 2015; Jordan et al., 2012). Im *KoMus*-Modell etwa wurde die Verbalisierung von musikalischer Wahrnehmung und das Kontextwissen zwei verschiedenen Dimensionen zugeordnet (Jordan et al., 2012, S. 514). Eine solche Untergliederung der Kompetenz in mehrere Teilkompetenzen ließ sich in der vorliegenden Arbeit nicht feststellen.

Ein zentrales inhaltliches Merkmal des *MARKO*-Kompetenzmodells ist, dass sich Personen auf den niedrigeren Niveaus A und B auf saliente, also besonders herausstechende musikalische Merkmale beziehen und sie erst ab Niveau C in der Lage sind, auf detailliertere musikalische Aspekte einzugehen. Diese Unterscheidung war in Rolles (2013, 2017) theoretischem Modell nicht vorgesehen, sondern hat sich vielmehr induktiv aus den Daten ergeben, als die Kodierregeln für die Testaufgaben erarbeitet wurden. Hier zeigen sich einige interessante Gemeinsamkeiten mit dem

Kompetenzmodell „Wahrnehmen und Kontextualisieren von Musik“, das im Rahmen der *KoMus*-Studie entwickelt wurde (Jordan et al., 2012). In diesem Kompetenzmodell gibt es ebenfalls die Unterscheidung zwischen „herausstechenden“ bzw. „salienten“ musikalischen Merkmalen auf den unteren Kompetenzniveaus und komplexeren musikalischen Zusammenhängen, die auf den höheren Niveaustufen wahrgenommen und benannt werden können (Jordan et al., 2012, S. 514). Diese Ähnlichkeit zwischen *MARKO*- und *KoMus*-Modell scheint theoretisch begründbar, schließlich können Personen ein Urteil über Musik erst dann fällen, wenn sie diese zuvor wahrgenommen haben (s. a. Kapitel 2.2). Zwischen den beiden Kompetenzmodellen gibt es weitere inhaltliche Gemeinsamkeiten, da die Verbalisierung musikalischer Wahrnehmung eine eigenständige Kompetenzdimension im *KoMus*-Modell ist. Im *KoMus*-Modell ist mitunter die Kompetenz festgehalten, Musik kritisch zu bewerten und das Urteil zu verbalisieren (Jordan et al., 2012, S. 514). Genau diese Kompetenz deckt sich mit der untersuchten Kompetenz im *MARKO*-Test. Ebenfalls findet sich im *KoMus*-Modell die Fähigkeit wieder, musikalische Parameter in Beziehung zum Ausdruck eines Musikstücks zu setzen. Insofern ist es durchaus denkbar, dass beide Kompetenztests in Teilen dasselbe messen. Trotz der beschriebenen Gemeinsamkeiten gibt es einige inhaltliche Unterschiede zwischen beiden Kompetenzkonstrukten, da der Fokus des *KoMus*-Tests auf Wahrnehmungsfähigkeiten und der des *MARKO*-Tests auf argumentativen Fähigkeiten liegt. Wie genau die beiden musikbezogenen Kompetenzen zusammenhängen, die im Rahmen des *KoMus*- und *MARKO*-Projekts modelliert wurden, muss in zukünftigen Untersuchungen geklärt werden.

In Schulleistungsstudien werden häufig Unterschiede zwischen den Leistungen von Jungen und Mädchen diskutiert. Diese geschlechtsspezifischen Unterschiede werden auch auf die unterschiedliche Sozialisation von Jungen und Mädchen zurückgeführt (OECD, 2020). Die Fairness der Testaufgaben im Hinblick auf das Geschlecht war mithilfe von DIF-Analysen sichergestellt worden (Kapitel 6.2). Mädchen wurden demnach nicht bevorteilt aufgrund ihres Geschlechts. In der vorliegenden Studie schnitten weibliche Personen zwar besser im Test ab, waren jedoch auch musikalisch erfahrener. Bisher gibt es nur wenig empirische Genderforschung in der Musikpädagogik. Heß (2018) sowie Fiedler und Hasselhorn (2020) zeigten, dass Mädchen ein höheres musikalisches Selbstkonzept haben als Jungen. In Bezug auf die der vorliegenden Arbeit verwendeten Skala (Gold-MSI) zur Messung musikalischer Erfahrenheit ist bisher noch wenig über genderspezifische Effekte bekannt (Greenberg et al., 2015; Müllensiefen et al., 2014).

7.2.2. Vergleich des theoretischen und empirischen Modells

Der *MARKO*-Kompetenztest wurde ausgehend von Rolles theoretischem Kompetenzmodell zum musikbezogenen Argumentieren entwickelt. Das Kompetenzmodell, das aus den empirischen Daten abgeleitet wurde (Abbildung 5 auf S. 76), bestätigt einige theoretische Annahmen aus Rolles Modell. In beiden Modellen unterscheiden sich Personen in ihrer Art und Weise, musikbezogene Urteile zu begründen. Während das individuelle Gefallen von Musik auf dem jeweils untersten Niveau bei der Urteilsbegründung zentral ist, sind Personen auf den höheren Niveaus in der Lage, verschiedene Standpunkte und Meinungen zur Musik zu diskutieren. In beiden Modellen bezie-

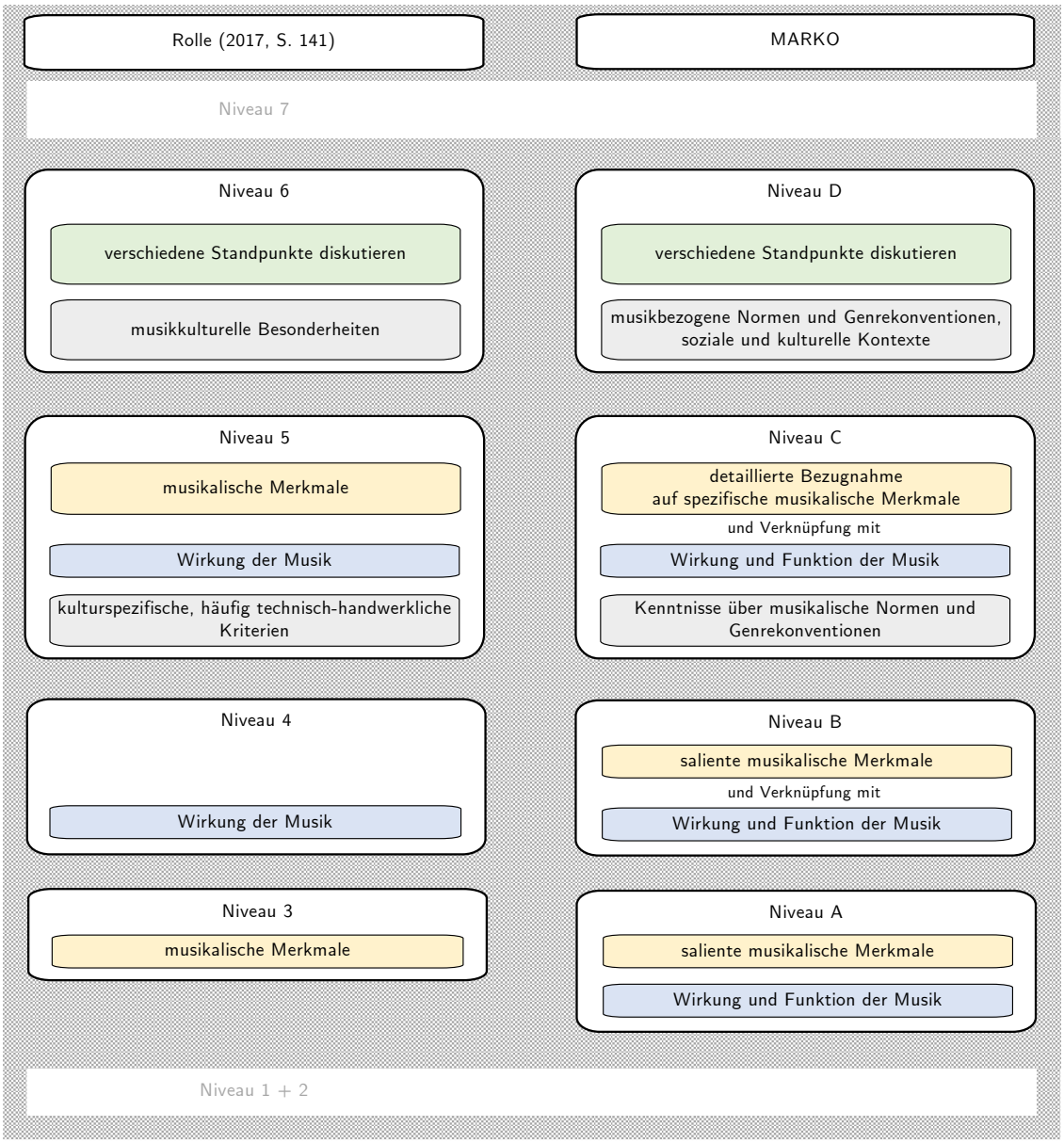
hen sich Personen auf Merkmale der Musik in ihrer Urteilsbegründung und subjektive Sichtweisen spielen ebenfalls eine Rolle, wie etwa die zugeschriebene Wirkung der Musik. Außerdem beziehen sich Personen auf höheren Niveaus häufiger auf musikbezogene Normen. Es gibt jedoch auch einige Unterschiede zwischen theoretischem und empirischem Kompetenzmodell.

Zunächst lässt sich feststellen, dass Rolles Modell sieben Kompetenzniveaus beschreibt und das empirische Modell nur vier. Abbildung 16 zeigt schematisch die Gemeinsamkeiten zwischen Rolles Niveaus 3 bis 6 und den empirischen Niveaus A bis D. Der Abbildung lässt sich ebenfalls entnehmen, dass sich inhaltliche Aspekte der Niveaus 1, 2 und 7 aus Rolles Modell nicht im empirischen Modell wiederfinden. Im theoretischen Modell von Rolle beziehen sich Personen auf dem untersten Niveau 1 ausschließlich auf das persönliche Gefallensurteil. Diese subjektive Art der Bezugnahme ist auch im untersten empirischen Niveau A festgehalten. Jedoch nehmen Personen hier auch Bezug auf musikalische Merkmale. Testantworten, die lediglich ein Gefallensurteil enthielten, wie etwa „schön“ oder „ist schon ok“, wurden mit 0 Punkten kodiert. Diese Vorgehensweise hatte sich bei der Kodierung der Testaufgaben als reliabelste Lösung erwiesen.

Wie bereits erwähnt, konnte Rolles Niveau 2 in dem Modell, das aus den empirischen Daten abgeleitet wurde, inhaltlich nicht abgebildet werden. In Rolles Modell verweisen Personen auf Niveau 2 in ihrer Urteilsbegründung auf Autoritäten. Dieses Niveau wurde weder in der vorliegenden Arbeit noch in der Expertisestudie von Knörzer et al. (2015, 2016) gefunden. Die Annahme, dass sich Personen bei der Urteilsbegründung auf „Autoritäten bzw. Kenntnisse aus zweiter Hand“ beziehen (Rolle, 2017, S. 141), entstammt dem „Reflective Judgment Model“ von King und Kitchener (1994, 2004) und wurde von Rolle auf das musikbezogene Argumentieren übertragen. Inwiefern die Bezugnahme zu Autoritäten beim musikbezogenen Argumentieren ins Gewicht fällt, ist bisher nicht geklärt. Inhaltliche Aspekte, die in Rolles Niveau 7 thematisiert werden, finden sich ebenfalls nicht im empirischen Modell. Eigentlich wurde in den empirischen Daten noch ein höheres Niveau gefunden; gewissermaßen ein Niveau E. Leider gab es zu wenig Daten, um dieses Niveau zuverlässig zu modellieren. In zukünftigen Untersuchungen könnten verstärkt Musikstudierende einbezogen werden, um mehr Daten zu erheben, damit mehr Aufschluss über dieses Niveau gegeben werden kann.

Rolles Niveaus 3, 4, 5 und 6 lassen sich grob auf die empirischen Niveaus A, B, C und D übertragen (Abbildung 16). Jedoch gibt es auch hier einige Unterschiede im Detail. Diese betreffen v. a. den Umgang mit musikalischen Merkmalen und der Wirkung von Musik. Während im theoretischen Modell lediglich allgemein von der Fähigkeit zur Bezugnahme auf die „objektiven Eigenschaften der Musik“ die Rede ist (Rolle, 2017, S. 141), zeigt die empirische Untersuchung, dass das tatsächliche Antwortverhalten komplexer ist. Bei den Testantworten fiel auf, dass Personen häufig mithilfe ‚oberflächlicher‘ musikalischer Merkmale, wie „laut“, „leise“, „hohe Töne“ etc. auf die Musik Bezug nahmen. Diese ‚oberflächlichen‘ Merkmale wurden in der vorliegenden Arbeit *saliente* musikalische Merkmale genannt. Vor allem Personen auf den Niveaus A und B verwenden vorwiegend solche salienten Merkmale, um auf die Musik Bezug zu nehmen. Musikalische Merkmale, die sich auf detailliertere Aspekte der Musik beziehen, wurden im Kompetenzmodell *spezifische* musikalische Merkmale genannt. Hier beziehen sich Personen z. B. auf einzelne

Abbildung 16.
Schematische Gegenüberstellung Rolle (2017) und MARKO*-Modell*



Anmerkungen. Diese Abbildung zeigt Ausschnitte aus beiden Kompetenzmodellen. Zur besseren Übersicht wurden einige Formulierungen aus Rolles (2017) Modell paraphrasiert, die bereits im Zuge der Testentwicklung überarbeitet worden waren (s. a. Kapitel 5). Die Bezeichnungen „objektive Eigenschaften der Musik“ und „musikalische Parameter“ wurden durch das Synonym „musikalische Merkmale“ ersetzt. „Subjektive Eindrücke“ bzw. der „eigene Eindruck des Ausdrucks der Musik“ wurden im Laufe der Testentwicklung mit der Bezugnahme zur „Wirkung der Musik“ gleichgesetzt. Die „eigene [...] Hörweise [...] zu anderen Perspektiven ins Verhältnis setzen“ ist gleichbedeutend mit „verschiedene Standpunkte diskutieren“.

Stellen in der Musik oder verfügen über Kenntnisse gewisser musikalischer Normen. Eine Bezugnahme auf spezifische musikalische Merkmale erfolgt erst ab dem höheren Niveau C. An dieser Stelle differenziert Rolles theoretisches Modell im Vergleich zum empirischen Modell weniger.

Eine weitere inhaltliche Verschiedenheit zwischen theoretischem und empirischem Kompetenzmodell liegt in der Art und Weise, wie auf die Wirkung der Musik Bezug genommen wird. Rolles theoretisches Modell unterscheidet zwischen einer Bezugnahme auf „objektive Eigenschaften“ der Musik (Niveau 3) und dem „Eindruck des Ausdrucks der Musik“ (Niveau 4) (Rolle, 2017, S. 141). Diese beiden Arten der Bezugnahme, die in Rolles Modell auf zwei unterschiedlichen Niveaus festgehalten sind, konnten in den empirischen Daten nur selten getrennt voneinander beobachtet werden. Im empirischen Modell findet sich daher auf den Niveaus A bis C jeweils sowohl die Bezugnahme zu musikalischen Merkmalen als auch zur Wirkung und Funktion der Musik wieder. Auf dem untersten Niveau A können Personen Urteile unter Bezugnahme auf „saliente musikalische Merkmale“ begründen und sich in ihrer Argumentation auf die „Wirkung und Funktion der Musik“ beziehen. Personen dieser Niveaustufe benutzen auch durchaus saliente musikalische Merkmale wie etwa „leise“, um die Wirkung der Musik – im Sinne von „sie wirkt ruhig“ – zu beschreiben. Eine Verknüpfung von musikalischen Merkmalen und der Bezugnahme auf die Wirkung der Musik erfolgt erst ab Niveau B. Eine schematische Darstellung der Unterschiede zwischen theoretischem und empirischem Modell findet sich in Abbildung 16.

Darüber hinaus gibt es einige Aspekte in Rolles Modell, die mit der verwendeten Forschungsmethode nicht operationalisiert werden konnten. Rolle betonte, dass man bei der theoretischen Herleitung eines Kompetenzmodells das dialogische Moment von Argumentation berücksichtigen müsse (Rolle, 2013, S. 56). Auf allen Niveaus in Rolles Modell ist folglich eine Annahme darüber festgehalten, wie Personen beim Argumentieren auf ihr Gegenüber eingehen. Auf Niveau 4 heißt es beispielsweise: „Begründungen, die andere anführen, sind deren Begründungen, die die eigene Interpretation und Einschätzung nicht in Frage stellen können“ (Rolle, 2017, S. 141). Rolle beruft sich in seiner theoretischen Herleitung des Modells auf empirische Arbeiten von Parsons (1987) sowie King und Kitchener (1994, 2004), die mit ihren Proband*innen Interviews führten. Da in der vorliegenden Arbeit keine Interviews geführt wurden, konnte das dialogische Moment von Argumentation nur begrenzt operationalisieret werden. Zwar gibt es einige Testaufgaben, die versuchen, diese Dialogizität abzubilden (wie etwa die Aufgabe „Eurovision Song Contest“; Kapitel 5.2.2), dennoch fehlt in einem Kompetenztest ein Gegenüber, das mit der Testperson unmittelbar argumentiert. Um zu überprüfen, wie sich Personen in argumentativen Gesprächen verhalten, bräuchte man ein anderes Erhebungssetting, wie beispielsweise Gruppendiskussionen oder Interviews.

7.3. Überlegungen zu ästhetischer Wahrnehmung und Argumentation

Der Zusammenhang zwischen ästhetischer Wahrnehmungsfähigkeit und Argumentationskompetenz wurde empirisch nicht untersucht. Allerdings gibt es einige theoretische Arbeiten, die den Zusammenhang diskutieren. Theoretische Arbeiten zu Ästhetik und ästhetischer Argumentation gehen davon aus, dass man einen ästhetischen Gegenstand zunächst wahrnehmen muss, bevor

man ihn bewerten kann (s. a. Kapitel 2.2).

Laut Kleimann (2005, S. 110) ist es jedoch *nicht* der Fall, dass ästhetische Objekte zunächst neutral wahrgenommen werden und die urteilende Person anschließend auf Grundlage bestimmter Kriterien auf den ästhetischen Wert des Objekts schließt, da bei der unmittelbaren ästhetischen Wahrnehmung immer auch subjektive Sichtweisen und Erfahrungen eine Rolle spielen (s. a. Brandstätter, 2008, S. 14-16). Zudem ist die Musik, die man wahrnimmt, immer in einem bestimmten gesellschaftlichen und kulturellen Kontext situiert (Moore & Green, 2018). Dieser Kontext ist für die Bewertung von Musik und die Art und Weise wie wir über Musik kommunizieren zentral. So spielen beispielsweise *korresponsive* Sichtweisen, die das Verhältnis des eigenen Lebens zur Musik betreffen, beim ästhetischen Argumentieren eine Rolle. Wenn wir Musik korresponsiv wahrnehmen, dann stellen wir eine Beziehung zwischen Musik und unserer alltäglichen Wirklichkeit her (Seel, 1993, S. 34). Musik wird also danach beurteilt, ob sie zu unserer aktuellen Lebenssituation passt. Die Bedeutung solcher Sichtweisen im Zusammenhang mit gesellschaftlichen und kulturellen Kontexten sollen in Bezug auf den *MARKO*-Test im Folgenden diskutiert werden.

Musikstücke sind stets eingebettet in einen gesellschaftlichen und kulturellen Kontext. Mit Hip-Hop ist beispielsweise nicht nur eine bestimmte Musik verbunden, die als Klangereignis wahrgenommen wird, sondern auch eine spezielle Art sich zu kleiden, sich zu bewegen und sich auszudrücken. In korresponsiven Urteilen wird demnach immer auch die „Vorstellung gelingenden Lebens“ verhandelt (Seel, 1991, S. 133). Eine Schülerin, die darum gebeten wurde bei der Bearbeitung des *MARKO*-Tests laut zu denken, mochte Deutschrap nicht. Die Gründe hierfür suchte die Schülerin nicht nur im Musikstück selbst, das zur Debatte stand, sondern im Genre, das für sie mit bestimmten Konventionen einherging:

> [Deutschrap] ist immer auf einer Tonhöhe und geht auch relativ schnell die Texte halt. Und für mich, wenn ich dann so an Deutschrap und so denke, denke ich immer so bisschen an [...] so kriminell [...], so ein bisschen assi vielleicht auch. (D2_w14_8)

Wenn diese Schülerin über Deutschrap urteilt, dann bewertet sie also nicht nur die Musik selbst, sondern eine ganze Lebensform, die für sie mit der Musik assoziiert ist.

Rolle und Wallbaum (2011) beschreiben in ihrer Publikation den Fall einer Schülerin, die im Rahmen einer Projektwoche ihre Vorurteile gegenüber Hip-Hop reflektiert. Mit der Zeit konnte sie von einem korresponsiven Urteil absehen und die Musik aus einem anderen Blickwinkel betrachten. Laut eigener Aussage brauchte sie also eine „Kennlernphase“, bis sie die Musik mochte:

> Ich glaube, das hatte damit zu tun, dass ich anfangs das Stück mehr als [G]anzes hörte, und im Gesamteindruck so das Rhythmische des HipHop überwiegt. Später hörte ich dann selektiver, und dabei kamen auch melodiöse Elemente zum Vorschein. (Rolle & Wallbaum, 2011, S. 17)

Obwohl diese Aussage im Rahmen eines zwar verwandten, aber dennoch anderen Forschungskontextes entstanden ist, passt sie sehr gut zur Kompetenzabfolge, die Personen im *MARKO*-Kompetenzmodell durchschreiten. Es klingt geradezu so, als hätte die Schülerin während der Projekt-

woche alle vier Niveaus durchlaufen. Während die Schülerin die Musik zunächst als Ganzes betrachtete und bewertete (Niveau A), konnte sie mit der Zeit die Musik differenzierter wahrnehmen (Niveau B und C). Rolle und Wallbaum stellen fest, dass sich die positive Beschreibung der Schülerin ausschließlich auf klangliche Aspekte des Hip-Hop bezieht (Rolle & Wallbaum, 2011, S. 17). Sie spekulieren darüber, ob es der Schülerin im Laufe der Zeit gelungen ist, kulturtypische Geschlechterrollen und Gesten des Hip-Hop von rein musikalischen Aspekten abzuspalten. Auch diese Überlegungen können auf das *MARKO*-Kompetenzmodell übertragen werden. Auf den unteren Niveaus kann man noch nicht von der eigenen Lebensform absehen und andere Meinungen diskutieren. Dies passiert erst auf dem obersten Niveau D. Auf den unteren Niveaus könnte demnach eine korresponsivere Sichtweite auf die Musik im Mittelpunkt stehen. In einem korresponsiven Wahrnehmungsmodus wird nicht nur die Musik selbst, sondern „eine Lebensform als Ganze[s]; eine Kultur, wenn man so will“ bewertet (Rolle & Wallbaum, 2011, S. 16). Wenn vermeintlich über Musik gestritten wird, könnte es demnach häufig nicht um die Musik selbst, sondern um etwas anderes gehen: um die Vorstellungen einer Lebensform, mit der man sich identifiziert. In diesem Falle würden sich die Gefallensurteile gar nicht auf die Musik als Objekt beziehen, sondern auf das Subjekt, dessen Lebensform zur Debatte steht. Demnach streitet man dann gar nicht über das Objekt (die Musik), sondern über die Vorstellungen, die die eigene Lebenswirklichkeit betreffen.

Ästhetische Werturteile können also stark von einer korresponsiven Sichtweise geprägt sein. Dieser Sachverhalt könnte erklären, weshalb im Pretest bestimmte Aufgabentypen nicht funktionierten. Im ersten Pretest kamen u. a. Aufgaben zum Einsatz, in denen sich Testpersonen mit dem Musikgeschmack anderer auseinandersetzen sollten. Eine dieser Aufgabenstellungen lautete: *„Einer guten Freundin bzw. einem guten Freund gefällt das Stück gar nicht. Was an der Musik könnte ihm/ihr nicht gefallen? Versuche, ihn/sie von Deinem Standpunkt zu überzeugen“*. Viele Personen gingen nicht näher auf die Musik ein und antworteten lediglich, dass verschiedenen Menschen unterschiedliche Musik gefalle (s. a. Ehninger et al., 2021a). Diese Aufgabe bot anscheinend keinen ausreichenden Anreiz, sich mit der Musik zu beschäftigen und so blieb auch die Bereitschaft aus, sich in eine andere Person hineinzuversetzen. Die Versuchspersonen wurden schließlich in überarbeiteten Aufgabenstellungen explizit dazu aufgefordert, auf die Musik Bezug zu nehmen und so unterschiedliche Meinungen nicht bloß als verschiedene „Geschmäcker“ abzutun. Durch die Aufgabenstellung wurden demnach die Schüler*innen dazu angeregt, eine korresponsive Sichtweise zu verlassen und explizit auf die Musik einzugehen. Wenn man von sich selbst absehen und die eigene Lebensweise und Sozialisation reflektieren kann, kommuniziert man möglicherweise sachlicher, als wenn die eigene Lebensweise ebenfalls verhandelt wird.

Wie bereits erwähnt, sind Musikstücke stets eingebunden in kulturelle Praxen. Rolle und Wallbaum (2011, S. 15) gehen davon aus, dass man zwischen verschiedenen Musikkulturen, wie beispielsweise Hip-Hop, Klassik oder Jazz unterscheiden kann. Kleimann (1998, S. 71) stellt fest, dass eine argumentative Rechtfertigung musikbezogener Urteile nur dann funktioniert, wenn urteilende wie adressierte Personen ähnliche Hörgewohnheiten haben und sich in „einem Raum geteilter musikalischer Grunderfahrungen“ bewegen. Argumentiert werden könnte dann also nur,

wenn sich Personen innerhalb einer geteilten (Musik-)Kultur bewegen. In Situationen, in denen die Personen nicht über geteilte Musikkulturen verfügen, müssten verschiedene Sicht- und Hörweisen verhandelt werden, damit es möglich ist, die andere Sichtweise nachzuvollziehen. Dies mag in einem realen Gespräch möglich sein, jedoch nicht im Rahmen eines Schulleistungstests.

Im *MARKO*-Test wird die argumentative Leistung von Personen danach bewertet, wie gut sie über Musik argumentieren können, die ihnen vorgegeben wird. Wenn Testpersonen die kulturelle Praxis, in der das vorgegebene Musikstück steht, nicht teilen, wird es schwierig sein, ein Urteil zu fällen, das nach Maßstäben des Tests gut bewertet werden wird. Im Test kam beispielsweise eine Aufgabe zum Einsatz, in der Testpersonen die Leistung einer Sängerin beurteilen sollten, die die „Rachearie“ der Königin der Nacht aus der „Zauberflöte“ offensichtlich schief sang.[51] In der Kodierregel ist festgehalten, dass Versuchspersonen mindestens erwähnen müssen, dass die Sängerin schief singt, um einen Punkt in der Aufgabe zu bekommen. Nehmen wir an, eine Testperson ist nicht mit klassischer westlicher Musik und dem damit einhergehenden tonalen System vertraut. Dann hätte diese Person auch keine Vorstellung davon, wie das Musikstück klingen sollte. Würde diese Person die Darbietung als schief beurteilen? Möglicherweise nicht. Auch wenn im Test verschiedene musikalische Genres wie Klassik, Barock, Pop, Rock, Jazz und Hip-Hop vorkommen, sind diese Genres alle das Produkt einer westlichen Musiktradition. Außerdem setzt der Test voraus, dass man über Musik verbal kommuniziert. Diese Praxis muss ebenfalls von den Versuchspersonen geteilt werden. Wenn Versuchspersonen nicht damit vertraut sind, über Musik verbal zu kommunizieren, dann werden sie ebenfalls nicht gut im Test abschneiden. Dennoch könnte es durchaus sein, dass sich eine Person nonverbal sehr gut über Musik verständigen kann und diese sehr vielfältig wahrnehmen kann. Diese Art der musikbezogenen Kommunikation kann im Rahmen eines derartigen Kompetenztests nicht erfasst werden. Der Test erfüllt demnach nicht den Anspruch, universell für alle Musikkulturen gültig zu sein. Jedoch bezieht er sich eindeutig auf die in deutschen Schulcurricula verankerte Kompetenz Urteile über Musik begründen zu können. Insofern erhebt der Test lediglich Anspruch darauf, im besten Falle in diesem Bildungskontext die untersuchte Kompetenz messen zu können und somit curricular valide zu sein (s. a. Kapitel 7.4).

7.4. Limitationen der Studie

Wie in der Methodenreflexion dargestellt (Kapitel 2.5; s. a. Ehninger, 2021), ist es unmöglich, mit einer Forschungsmethode alle Aspekte zu untersuchen, die beim musikbezogenen Argumentieren eine Rolle spielen. Dementsprechend konnte auch die vorliegende Studie dies nicht gewährleisten. Argumentieren ist ein interaktives Geschehen und ein Schlagabtausch mit einem echten Gegenüber kann, wie bereits diskutiert, in einem Kompetenztest nur begrenzt dargestellt werden. Obwohl es mehrere Items im finalen *MARKO*-Test gab, die dialogische Situationen imitierten (wie z. B. das Item „Eurovision Song Contest“; Kapitel 5.2), kann ein Kompetenztest niemals so interaktiv

[51] Es handelte sich dabei um die Darbietung von Florence Foster Jenkins.

sein wie ein Gespräch mit einer oder mehreren ‚echten' Personen. Insofern läuft der *MARKO*-Test Gefahr – wie andere schriftliche Erhebungsverfahren zum Argumentieren auch – Argumentationskompetenz monologisch zu konzeptualisieren. Pohl spricht in diesem Zusammenhang auch von der Gefahr eines „monologischen Reduktionismus", da die dialogische Perspektive verloren geht (Pohl, 2014, S. 288).

In der Pilotstudie stellte sich heraus, dass sich offene Items besonders eignen, um musikbezogene Argumentationskompetenz zu erfassen, da hier die Versuchspersonen selbstständig Argumente produzieren können. Die Verwendung von offenen Items bringt jedoch – abgesehen von einem sehr aufwändigen Auswertungsprozess – einige methodische Probleme mit sich. Die Genauigkeit einer Messung in Leistungstests ist immer auch von der Motivation der Versuchspersonen abhängig (Bühner, 2021, S.74-75). Es ist durchaus denkbar, dass Personen mit einer geringeren motivationalen Bereitschaft nur kurze Antworten schrieben und daher schlechter im Test abschnitten. Kompetentere Personen hingegen schrieben nicht nur längere Texte, sie bearbeiteten auch weniger Testaufgaben. Dies könnte darauf zurückzuführen sein, dass sie sich ausführlicher mit den einzelnen Aufgaben auseinandersetzten und deshalb umfassender argumentierten. Die fehlenden Werte sind positiv mit den Personenfähigkeiten der Testpersonen korreliert ($r = -.51$; $p < .01$). Dies bedeutet, dass die Stichprobe systematische fehlende Werte enthält, die die Genauigkeit der Messung beeinträchtigen können.

Der Kompetenztest richtete sich an Schüler*innen der neunten bis zwölften gymnasialen Jahrgangsstufe sowie an Musikstudierende. Die Altersspanne der teilnehmenden Personen war daher sehr groß: Die meisten Schüler*innen waren 14 bis 18 Jahre alt und das Alter der Studierenden lag zwischen 19 und 38. Im Pfadmodell ergab sich, dass die Klassenstufe der einflussreichste Prädiktor für die Ausprägung der Personenfähigkeit ist. Ältere Versuchspersonen sprachen allerdings meist deutsch zuhause und waren musikalisch erfahrener. In zukünftigen Untersuchungen könnte durch Stichproben mit kleineren Altersspannen der Einfluss des Alters besser kontrolliert werden und so könnten ggf. dezidiertere Aussagen über den Zusammenhang zwischen musikbezogener Argumentationskompetenz und anderen Prädiktorvariablen, wie etwa musikalischer Erfahrenheit, getroffen werden.

Es ist bisher noch nicht geklärt, welche Rolle allgemeinsprachliche Kompetenzen beim musikbezogenen Argumentieren spielen. In den 90-minütigen Datenerhebungen konnten aus Zeitgründen keine Testinstrumente eingesetzt werden, die ausschließlich sprachliche Kompetenzen erfassen. Es liegt jedoch auf der Hand, dass selbige beim musikbezogenen Argumentieren eine wesentliche Rolle spielen. Morek et al. (2017) betrachten das Argumentieren als bildungssprachliche Praktik, deren Erwerb und Anwendung einen entscheidenden Einfluss auf schulischen Lernerfolg haben. Bildungssprache wird zudem häufig als Ressource gesehen, zu der Schüler*innen keinen gleichberechtigten Zugang haben (Morek & Heller, 2012; Quasthoff et al., 2020a; Schmölzer-Eibinger, 2013). So betonen Quasthoff et al. (2020b) die Rolle von familialen Ressourcen für den Erwerb von Argumentationskompetenz. Möglicherweise schneiden Versuchspersonen schlechter im *MARKO*-Test ab, die über kein Umfeld mit den entsprechenden Ressourcen verfügen. Ein Kompetenztest, der im schulischen Kontext durchgeführt wird, setzt demnach schon ein bestimmtes

bildungssprachliches Niveau beim Argumentieren voraus, das aufgrund ihrer diversen familialen und sprachlichen Sozialisation nicht bei allen Schüler*innen gleichermaßen vorausgesetzt werden kann. Zwar wurde die Testfairness im Hinblick auf den häuslichen Sprachgebrauch in den DIF-Analysen sichergestellt, dennoch schnitten Personen besser im Test ab, die ausschließlich deutsch zuhause sprachen. Allerdings ist fraglich, ob besonders sprachkompetente Personen per se gut in einem Test für musikbezogene Argumentationskompetenz abschneiden. Ein explorativer Blick in die Daten zeigt, dass es durchaus Personen gibt, die sich zwar eloquent ausdrücken, jedoch kaum Bezug auf die Musik nehmen.[52] Der Zusammenhang zwischen allgemeiner Sprachkompetenz und musikbezogener Argumentationskompetenz sollte unbedingt in zukünftigen Studien untersucht werden (s. dazu auch Kapitel 7.5).

Umfassende Untersuchungen zur Validität der Kompetenzmessung im Rahmen des *MARKO*-Tests stehen noch aus. Dies betrifft vor allem die Konstruktvalidität[53] sowie die curriculare Validität des Tests. Der moderate Zusammenhang zwischen musikalischer Erfahrenheit und den Personenfähigkeitswerten im *MARKO*-Test ist ein erster Hinweis auf die konvergente Validität des Tests, $r = .40$, $p < .001$. In zukünftigen Studien sollte der Zusammenhang zwischen dem *MARKO*-Test und anderen musikbezogenen Kompetenztests wie dem *KoMus*- oder *KOPRA-M*-Test genauer untersucht werden (Hasselhorn, 2015; Jordan et al., 2012). Bei der Testentwicklung wurden zwar Schulbücher und -curricula berücksichtigt, um die curriculare Validität zu gewährleisten, aber eine weitere Überprüfung steht noch aus. Beispielsweise könnten Musiklehrer*innen die einzelnen Aufgaben in Hinblick auf ihre Relevanz und Lehrplanpassung bewerten.

7.5. Fazit und Ausblick

Mit dem *MARKO*-Test und -Kompetenzmodell ist es erstmals möglich, empirisch fundierte Erkenntnisse über den Kompetenzverlauf und -erwerb musikbezogenen Argumentierens bereitzustellen. Das Kompetenzmodell bietet Aufschluss darüber, welche Anforderungen Personen beim musikbezogenen Argumentieren bewältigen müssen. Der Test kann nicht nur zur Lernstandser-

[52] In einer Testaufgabe sollten Versuchspersonen auf einen Kommentar zum Musikvideo „This is America“ von Childish Gambino Stellung beziehen. Diese*r Schüler*in schrieb eine sehr ausführliche Antwort, die hier nur in Auszügen dargestellt ist. Die Testperson bezog sich jedoch nur auf die politische Situation in den USA und nicht auf die Musik (Rechtschreibung nicht korrigiert): „In America lebten früher Siedler welche durch Immigranten verdrängt wurden (Anfang des Liedes). Außerdem leben so viele Verschiedene ethnische Gruppen in Amerika, jedoch ist es eher ein aneinander vorbeileben statt ein miteinander und viele werden verdrängt oder für niedriger gehalten. [...] Insbesondere die Afroamerikaner haben es nicht leicht dort. unter der Macht des ‚neuen Präsidenten Trump‘, hat sich die ganze Situation noch verschärft, da er nur noch ‚Amerikaner‘ in seinem Land haben will [...]. Noch dazu kommt, dass es in den USA täglich schießereien gibt und sehr viele Ammokläufe, da Waffen für jedermann einfach käuflich zu erwerben sind, wogegen nichts getan wird. Oftmals werden gezielt Bürger mit afroamerikanischen Hintergrund grundlos erschossen. Gegen all dies sollte man dringen etwas tun [...]“ (P2_46).

[53] *Konstruktvalidität* bedeutet, dass ein Testwert hypothesenkonform mit anderen theoretischen Konstrukten korreliert (Döring & Bortz, 2016, S. 446). Beispielsweise sollten die Ergebnisse des *MARKO*-Tests mit anderen musikbezogenen Kompetenztests höher korrelieren als mit den Ergebnissen eines Tests für räumliches Vorstellungsvermögen. Die wünschenswerte hohe Korrelation mit verwandten Konstrukten bezeichnet man auch *konvergente* Validität und den Zusammenhang mit weniger verwandten Konstrukten *diskriminante* Validität (Döring & Bortz, 2016, S. 446).

hebung verwendet werden, sondern auch als Lernwerkzeug: Lernfelder können identifiziert und musikbezogene Argumentationskompetenz so gezielt gefördert werden. Die Dissertation eröffnet demnach zahlreiche Möglichkeiten für Anschlussstudien, um dem nach wie vor bestehenden Desiderat in der Erforschung musikbezogener Kompetenzen zu begegnen.

Demnächst wird eine Kurzversion des Tests zur Verfügung stehen, mit der es möglich sein wird, im Rahmen einer ca. 30-minütigen Testung die musikbezogene Argumentationskompetenz zu erheben (*MARKO-S*). Dieser Kurztest könnte auch einen positiven Einfluss auf die motivationale Bereitschaft der Versuchspersonen haben. Zudem sind zusätzliche Analysen in Arbeit, die sich mit der Frage auseinandersetzen, weshalb bestimmte Aufgaben im *MARKO*-Test schwieriger waren als andere (Ehninger et al., i. Vorb.). In dieser geplanten Publikation wird in Regressionsanalysen geprüft, welchen Einfluss beispielsweise das Musikgenre oder die Textlänge einer Aufgabe auf deren Schwierigkeit hat. Diese Analysen können zusätzlichen Aufschluss darüber geben, welche Anforderungen Personen bei der Testbearbeitung bewältigen müssen.

Zukünftige Studien könnten ebenfalls vertieft untersuchen, welche Prädiktoren Einfluss auf die Ausprägung musikbezogener Argumentationskompetenz haben. So scheinen Motivation, das Musikinteresse in der Familie, Musizierpraxis sowie der sozioökonomische Hintergrund eine wichtige Rolle im Musikunterricht zu spielen (Fiedler & Hasselhorn, 2020; Harnischmacher & Knigge, 2017; Jordan, 2014). Ebenfalls gibt es erste Befunde zum Zusammenhang zwischen den Big Five Persönlichkeitsdimensionen und der Ausprägung musikalischer Erfahrenheit (Greenberg et al., 2015; Müllensiefen et al., 2014). Der Zusammenhang mit der Persönlichkeitsdimenison Offenheit für Erfahrung bzw. der Unterkategorie „Openess to Aesthetics“ scheint hierbei besonders bedeutsam zu sein (Greenberg et al., 2015, S. 156). Der Einfluss von Musikpräferenzen und Vertrautheit mit Musik sollte in zukünftigen Studien ebenfalls geklärt werden.

Nach wie vor ist soziale Herkunft für den Bildungserfolg von Schüler*innen in Deutschland zentral (Stanat et al., 2019; Stanat et al., 2017). In Bezug auf den Erwerb von Argumentationskompetenz betonen Quasthoff et al. (2020b) die Rolle der familialen Unterstützung. Dieser Einfluss sollte in zukünftigen Studien ebenfalls untersucht werden, zumal diskursive Sprachhandlungen wie das Argumentieren im Fachunterricht häufig nicht vermittelt werden (Morek, 2016). Um dem Zusammenhang zwischen musikbezogener Argumentationskompetenz und allgemeinsprachlichen Kompetenzen auf den Grund zu gehen, könnte in zukünftigen Studien neben dem *MARKO*-Test ein Kompetenztest für Sprache eingesetzt werden. In einem nächsten Schritt könnte dann überprüft werden, inwiefern die beiden Kompetenzen zusammenhängen. Außerdem wäre es denkbar, die sprachliche Qualität von Testantworten zu untersuchen. Es gibt bereits erste Prototypen für die computergestützte Auswertung von offenen Testantworten (Zehner et al., 2016). Eine solche automatisierte Codierung von Testantworten würde es deutlich leichter machen, entsprechende Daten auszuwerten.

Die vorliegende Dissertation legt einen wichtigen Grundstein für die empirische Erforschung musikbezogener Argumentationskompetenz. Erstmals gibt es empirisch belastbare Befunde zu Anforderungen, die beim musikbezogenen Argumentieren zentral sind. Die Verankerung der Kompetenz

in den deutschen Lehrplänen sowie die Bedeutsamkeit sprachlicher Kompetenzen für fachliches Lernen verdeutlichen die Relevanz von Argumentationskompetenz für schulisches Lernen. Die Messung und Untersuchung ist deshalb bedeutend und birgt großes Potenzial für das Musiklernen. Auch für das praktische Musizieren ist die Fähigkeit, sich verbal über Musik zu verständigen entscheidend. Schließlich müssen beim gemeinsamen Musizieren immer wieder unterschiedliche musikalische Vorstellungen verhandelt werden. Argumentationskompetenz bleibt somit sowohl als Ziel des Unterrichts relevant, als auch als Mittel, um andere Kompetenzen aufzubauen (Budke, 2013, S. 13). Beim Argumentieren können Wissensbestände, Werte, Meinungen, gesellschaftliche Vorstellungen und vieles mehr verhandelt werden. Nebenbei werden die argumentierenden Personen dazu aufgefordert, ihren eigenen Standpunkt zu reflektieren. So können sie lernen, die eigene Sichtweise überzeugend darzustellen. Argumentieren ist somit als wesentliches Werkzeug nicht nur für schulisches Lernen, sondern auch für gesellschaftliche Teilhabe zu betrachten und hat damit sowohl institutionelle als auch gesamtgesellschaftliche Relevanz.

Literaturverzeichnis

Ames, A. J. & Penfield, R. D. (2015). An NCME Instructional Module on Item-Fit Statistics for Item Response Theory Models. *Educational Measurement: Issues and Practice*, *34*(3), 39–48. https://doi.org/10.1111/emip.12067

Ansorge, U. (2021). Salienz. In M. A. Wirtz (Hrsg.), *Dorsch Lexikon der Psychologie*. Hogrefe. https://dorsch.hogrefe.com/stichwort/salienz

Beck, B. & Klieme, E. (2007). *Sprachliche Kompetenzen Konzepte und Messung; DESI-Studie (Deutsch-Englisch-Schülerleistungen International)*. Beltz. https://www.pedocs.de/volltexte/2010/3140/pdf/978_3_407_25398_9_1A_D_A.pdf

Blömeke, S., Kaiser, G. & Lehmann, R. (2010). *TEDS-M 2008. Professionelle Kompetenz und Lerngelegenheiten angehender Mathematiklehrkräfte für die Sekundarstufe I im internationalen Vergleich*. Waxmann. https://www.erziehungswissenschaften.hu-berlin.de/de/institut/abteilungen/didaktik/data/aufsaetze/2010/TEDS-M_Primar_End.pdf

Bond, T. G. & Fox, C. M. (2015). *Applying the Rasch Model. Fundamental Measurement in the Human Sciences* (3rd Edition). Routledge.

Bossen, A. (2017). Sprache als Gegenstand der Musikpädagogischen Forschung und des musikdidaktischen Diskurses im Kontext einer Sprachbildung im Fach. In A. Bossen & B. Jank (Hrsg.), *Sprache im Musikunterricht. Ausgewählte Aspekte sprachbewussten Handelns im Kontext von Inklusion* (S. 21–54). Universitätsverlag.

Bossen, A. (2019). *Sprachbewusster Musikunterricht. Problematisierung sprachdidaktischer Ansätze und Perspektiven einer Sprachbildung im Fach*. Waxmann.

Brandstätter, U. (2008). *Grundfragen der Ästhetik. Bild – Musik – Sprache – Körper*. UTB.

Brenk, M. (2014). Nur noch Etüden? – Kritisch-konstruktive Anmerkungen zur Kompetenzorientierung im Musikunterricht. *Zeitschrift für Kritische Musikpädagogik*. https://www.zfkm.org/14-brenk.pdf

Budke, A. (2013). Stärkung von Argumentationskompetenzen im Geographieunterricht – sinnlos, unnötig und zwecklos? In M. Becker-Mrotzek, K. Schramm, E. Thürmann & H. J. Vollmer (Hrsg.), *Sprache im Fach. Sprachlichkeit und fachliches Lernen* (S. 353–364). Waxmann.

Budke, A. & Meyer, M. (2015). Fachlich argumentieren lernen – Die Bedeutung der Argumentation in den unterschiedlichen Schulfächern. In A. Budke, M. Kuckuck, M. Meyer, F. Schäbitz, K. Schlüter & G. Weiss (Hrsg.), *Fachlich argumentieren lernen. Didaktische Forschungen zur Argumentation* (S. 9–28). Waxmann.

Bühner, M. (2021). *Einführung in die Test- und Fragebogenkonstruktion* (4. Aufl.). Pearson.

Chen, W.-H. & Thissen, D. (1997). Local Dependence Indexes for Item Pairs Using Item Response Theory. *Journal of Educational and Behavioral Statistics*, *22*(3), 265–289. https://doi.org/10.2307/1165285

Chng, G. S., Wild, E., Hollmann, J. & Otterpohl, N. (2014). Children's evaluative skills in informal reasoning: The role of parenting practices and communication patterns. *Learning, Culture and Social Interaction*, *3*, 88–97. http://dx.doi.org/10.1016/j.lcsi.2014.02.003

Detterbeck, M. & Schmidt-Oberländer, G. (2015). *Musix: Das Kursbuch Musik 3 für den Unterricht an allgemeinbildenden Schulen*. Helbling.

Domenech, M., Krah, A. & Hollmann, J. (2017). Entwicklung und Förderung der Argumentationskompetenz in der Sekundarstufe I: Die Relevanz familiärer Ressourcen. *Bildung und Erziehung*, *70*(1), 91–107. https://www.degruyter.com/downloadpdf/j/bue.2017.70.issue-1/bue-2017-0108/bue-2017-0108.pdf

Döring, N. & Bortz, J. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften* (5. Aufl.). Springer.

Eemeren, F. H. v., Garssen, B., Krabbe, E. C. W., Snoeck Henkemans, A. F., Verheij, B. & Wagemans, J. H. M. (2014). *Handbook of Argumentation Theory*. Springer.

Eemeren, F. H. v. & Grootendorst, R. (2003). *A Systematic Theory of Argumentation: The pragma-dialectical approach*. Cambridge University Press. https://doi.org/10.1017/CBO9780511616389

Ehninger, J. (2021). Wie lässt sich musikbezogene Argumentationskompetenz empirisch untersuchen? Über die empirische Erforschung einer facettenreichen Kompetenz. *Beiträge empirischer Musikpädagogik*, *12*. https://www.b-em.info/index.php/ojs/article/view/192

Ehninger, J., Knigge, J. & Rolle, C. (2021a). Musikbezogene Argumentationskompetenz. Ein Werkstattbericht über die Entwicklung von Testaufgaben. In A. Budke & F. Schäbitz (Hrsg.), *Argumentieren und Vergleichen* (S. 93–112). LIT.

Ehninger, J., Knigge, J. & Rolle, C. (i. Vorb.). Why are Certain Items More Difficult than Others in a Competency Test for Music-Related Argumentation?

Ehninger, J., Knigge, J., Schurig, M. & Rolle, C. (2021b). A New Measurement Instrument for Music-Related Argumentative Competence: The MARKO Competency Test and Competency Model. *Frontiers in Education*, *6*(191). https://doi.org/10.3389/feduc.2021.668538

Ehninger, J. & Rolle, C. (2020). Musikbezogenes Argumentieren – Nur Geschmacksache? Über die Entwicklung eines Kompetenztests. In M. Schwarzbauer & K. Steinhauser (Hrsg.), *‚Nur' Geschmacksache? Der Umgang mit kreativen Leistungen im Musik- und Kunstunterricht* (S. 168–182). LIT.

Erath, K., Prediger, S., Quasthoff, U. & Heller, V. (2018). Discourse competence as important part of academic language proficiency in mathematics classrooms: the case of explaining to learn and learning to explain. *Educational Studies in Mathematics*, *99*(2), 161–179. https://doi.org/10.1007/s10649-018-9830-7

Fiedler, D. & Hasselhorn, J. (2020). Zum Zusammenhang von musikalischem Selbstkonzept und Motivation im Musikunterricht. *Beiträge empirischer Musikpädagogik*, *11*. https://www.b-em.info/index.php/ojs/article/view/187

Frederking, V., Henschel, S., Meier, C., Roick, T., Stanat, P. & Dickhäuser, O. (2012). Beyond Functional Aspects of Reading Literacy: Theoretical Structure and Empirical Validity of

Literary Literacy. *L1-Educational Studies in Language and Literature*, *12*, 1–24. https://doi.org/10.17239/L1ESLL-2012.01.02

Frederking, V., Meier, C., Brüggemann, J., Gerner, V. & Friedrich, M. (2011a). Literarästhetische Verstehenskompetenz – theoretische Modellierung und empirische Erforschung. *Zeitschrift für Germanistik*, *21*(1), 131–144. https://doi.org/10.3726/92132_131

Frederking, V., Roick, T. & Steinhauer, L. (2011b). Literarästhetische Urteilskompetenz. Forschungsansatz und Zwischenergebnisse. In H. Bayrhuber, U. Harms, B. Muszynski, B. Ralle, M. Rothgangel, L.-H. Schön, J. H. Vollmer & H. G. Weigand (Hrsg.), *Empirische Fundierung in den Fachdidaktiken* (S. 75–94). Waxmann.

Freunberger, R., Robitzsch, A. & Luger-Bazinger, C. (2016). Statistische Analysen produktiver Kompetenzen. In S. Breit & C. Schreiner (Hrsg.), *Large-Scale Assessment mit R. Methodische Grundlagen der österreichischen Bildungsstandardüberprüfung* (S. 225–258). Facultas.

Gies, S. & Jank, W. (2015). *Music Step by Step 2. Aufbauender Musikunterricht ab Klasse 7. Lehrerhandbuch*. Helbling.

Gottschalk, T. (i. Vorb.). *Diagnosegeleitete Förderung ästhetischer Diskursfähigkeit im Kontext produktionsorientierten Musikunterrichts*. [Dissertation, Universität zu Köln].

Gottschalk, T. & Lehmann-Wermser, A. (2013). Der lange Weg zum Unterrichtsdesign. Zur Begründung und Umsetzung fachdidaktischer Forschungs- und Entwicklungsprogramme. In M. Komorek & S. Prediger (Hrsg.), *Der lange Weg zum Unterrichtsdesign. Zur Begründung und Umsetzung fachdidaktischer Forschungs- und Entwicklungsprogramme* (S. 63–78). Waxmann.

Greenberg, D. M., Müllensiefen, D., Lamb, M. E. & Rentfrow, P. J. (2015). Personality predicts musical sophistication. *Journal of Research in Personality*, *58*, 154–158. https://doi.org/https://doi.org/10.1016/j.jrp.2015.06.002

Gronostay, D. (2019). *Argumentative Lehr-Lern-Prozesse im Politikunterricht. Eine Videostudie*. Springer. https://doi.org/10.1007/978-3-658-25671-5

Haberecht, A. (i. Vorb.). *Musikalisch-ästhetische Diskurfähigkeit von Grundschulkindern. Eine videoanalytische Studie im Kontext sprachlicher Heterogenität*. [Dissertation, Friedrich-Alexander-Universität Erlangen-Nürnberg].

Harnischmacher, C. & Knigge, J. (2017). Motivation, Musizierpraxis und Musikinteresse in der Familie als Prädiktoren der Kompetenz „Musik wahrnehmen und kontextualisieren“ und des Kompetenzerlebens im Musikunterricht. *Beiträge empirischer Musikpädagogik*, *8*. https://www.b-em.info/index.php/ojs/article/view/136

Hasselhorn, J. (2015). *Messbarkeit musikpraktischer Kompetenzen von Schülerinnen und Schülern. Entwicklung und empirische Validierung eines Kompetenzmodells*. Waxmann.

Hasselhorn, J. & Knigge, J. (2018). Kompetenz und Expertise. In M. Dartsch, J. Knigge, A. Niessen, F. Platz & C. Stöger (Hrsg.), *Handbuch Musikpädagogik* (S. 197–207). Waxmann.

Hasselhorn, J. & McElvany, N. (2016). Die Bedeutung außerschulischer Prädiktoren für schulrelevante musikpraktische Kompetenzen. In R. McElvany, R. Strietholt, H. Holtappels & W. Bos (Hrsg.), *Jahrbuch der Schulentwicklung* (S. 168–205). Juventa.

Heß, F. (2018). *Gendersensibler Musikunterricht. Empirische Studien und didaktische Konsequenzen*. Springer.

Hessisches Kultusministerium. (2016). Kerncurriculum gymnasiale Oberstufe. https://kultusministerium.hessen.de/sites/default/files/media/kcgo-mu.pdf

Jahnke, H. N. & Ufer, S. (2015). Argumentieren und Beweisen. In R. Bruder, L. Hefendehl-Hebeker, B. Schmidt-Thieme & H.-G. Weigand (Hrsg.), *Handbuch der Mathematikdidaktik* (S. 331–355). Springer. https://doi.org/10.1007/978-3-642-35119-8_12

Johnson, R. H. & Blair, J. A. (1994). *Logical Self-Defence* (United States Edition). McGraw-Hill.

Johnson, R. H. & Blair, J. A. (2000). Informal Logic: An Overview. *Informal Logic*, *20*(2), 93–107.

Jordan, A.-K. (2014). *Empirische Validierung eines Kompetenzmodells für das Fach Musik – Teilkompetenz „Wahrnehmen und Kontextualisieren von Musik“*. Waxmann.

Jordan, A.-K., Knigge, J., Lehmann, A. C., Niessen, A. & Lehmann-Wermser, A. (2012). Entwicklung und Validierung eines Kompetenzmodells im Fach Musik: Wahrnehmen und Kontextualisieren von Musik. *Zeitschrift für Pädagogik*, *58*(4), 500–521.

Kant, I. (1790). Kritik der Urteilskraft. http://archiv-swv.de/pdf-bank/Kant,%20Immanuel%20-%20Kritik.der.Urteilskraft.pdf

Karantonis, A. & Sireci, S. (2006). The Bookmark Standard-Setting Method. A Literature Review. *Educational Measurement*, *25*, 4–7. https://doi.org/10.1111/j.1745-3992.2006.00047.x

Kelava, A. & Moosbrugger, H. (2020a). Deskriptivstatistische Itemanalyse und Testwertbestimmung. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 143–158). Springer.

Kelava, A. & Moosbrugger, H. (2020b). Einführung in die Item-Response-Theorie (IRT). In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 369–421). Springer.

King, P. M. & Kitchener, K. S. (1994). *Developing Reflective Judgment. Understanding and Promoting Intellectual Growth and Critical Thinking in Adolescents and Adults*. Jossey-Bass.

King, P. M. & Kitchener, K. S. (2004). Reflexive Judgment: Theory and Research on the Development of Epistemic Assumptions Through Adulthood. *Educational Psychologist*, *39*, 5–18.

Kleimann, B. (1998). Erfahrung und Argument. Überlegungen zum Begriff musikalischer Rationalität. In M. Pfeffer, J. Vogt, U. Eckart-Bäcker & E. Nolte (Hrsg.), *Systematische Musikpädagogik* (S. 67–80). Wißner.

Kleimann, B. (2005). Wie sprechen und urteilen wir über Kunst? In C. Jäger & G. Meggle (Hrsg.), *Kunst und Erkenntnis* (S. 95–116). Mentis.

Klemenz, S., König, J. & Schaper, N. (2019). Learning opportunities in teacher education and proficiency levels in general pedagogical knowledge: new insights into the accountability of teacher education programs. *Educational Assessment, Evaluation and Accountability*, *31*(2), 221–249. https://doi.org/10.1007/s11092-019-09296-6

Klieme, E., Avenarius, H., Blum, W., Döbrich, P., Gruber, H., Prenzel, M., Reiss, K., Riquarts, K., Rost, J., Tenorth, H.-E. & Vollmer, H. J. (2003). *Zur Entwicklung nationaler Bildungsstandards. Eine Expertise*. Bundesministerium für Bildung und Forschung.

Klieme, E. & Hartig, J. (2007). Kompetenzkonzepte in den Sozialwissenschaften und im erziehungswissenschaftlichen Diskurs. *Zeitschrift für Erziehungswissenschaft*, *10*(Sonderheft 8), 11–29. https://doi.org/10.1007/978-3-531-90865-6_2

Klieme, E. & Leutner, D. (2006). Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen. Beschreibung eines neu eingerichteten Schwerpunktprogramms der DFG. *Zeitschrift für Pädagogik*, *52*, 876–903.

Knigge, J. (2011). *Modellbasierte Entwicklung und Analyse von Testaufgaben zur Erfassung der Kompetenz "Musik wahrnehmen und kontextualisieren"*. LIT.

Knigge, J. (2014). Der Kompetenzbegriff in der Musikpädagogik: Verwendung, Kritik, Perspektiven. In J. Vogt, F. Hess & M. Brenk (Hrsg.), *(Grund-)Begriffe musikpädagogischen Nachdenkens. Entstehung, Bedeutung, Gebrauch* (S. 105–136). LIT.

Knigge, J. & Lehmann-Wermser, A. (2008). Bildungsstandards für das Fach Musik – eine Zwischenbilanz. *Zeitschrift für kritische Musikpädagogik*, (Sonderedition: Bildungsstandards und Kompetenzmodelle für das Fach Musik?). http://www.zfkm.org/sonder08-knigge-lehmannwermser.pdf

Knörzer, L. (2012). *Aufgabenanalysen im Rahmen der Entwicklung eines Tests zur Messung musikbezogener ästhetischer Argumentationskompetenz*. [Unveröffentlichte Abschlussarbeit, Hochschule für Musik Saar].

Knörzer, L., Rolle, C., Stark, R. & Park, B. (2015). „... er übertreibt und das macht mir seine Version zu nervös“: Einzelfallanalysen musikbezogener Argumentationen. In A. Niessen & J. Knigge (Hrsg.), *Theoretische Rahmung und Theoriebildung in der musikpädagogischen Forschung* (S. 147–162). Waxmann.

Knörzer, L., Stark, R., Park, B. & Rolle, C. (2016). “I like reggae and Bob Marley is already dead”: An empirical study on music-related argumentation. *Psychology of Music*, *44*(5), 1158–1174. https://doi.org/10.1177/0305735615614095

Koller, I., Alexandrowicz, R. & Hatzinger, R. (2012). *Das Rasch-Modell in der Praxis. Eine Einführung mit eRm*. facultas.

König, J. (2020). Kompetenzorientierter Ansatz in der Lehrerinnen- und Lehrerbildung. In C. Cramer, J. König, M. Rothland & S. Blömeke (Hrsg.), *Handbuch Lehrerinnen- und Lehrerbildung* (S. 163–171). UTB. https://www.handbuch-lehrerbildung.net/download/19-kompetenzorientierter-ansatz-in-der-lehrerinnen-und-lehrerbildung

König, J., Darge, K. & Kramer, C. (2020). Kompetenzentwicklung im Praxissemester: Zur Bedeutung schulpraktischer Lerngelegenheiten auf den Erwerb von pädagogischem Wissen bei Lehramtsstudierenden. In I. Ulrich & A. Gröschner (Hrsg.), *Praxissemester im Lehramtsstudium in Deutschland: Wirkungen auf Studierende* (S. 67–95). Springer. https://doi.org/10.1007/978-3-658-24209-1_2

Krelle, M. (2014). *Mündliches Argumentieren in leistungsorientierter Perspektive*. Schneider.

Krelle, M. & Willenberg, H. (2008). Argumentation Deutsch. In DESI-Konsortium (Hrsg.), *Unterricht und Kompetenzerwerb in Deutsch und Englisch. Ergebnisse der DESI-Studie* (S. 81–88). Beltz.

Kuhn, D. (1999). A Developmental Model of Critical Thinking. *Educational Researcher*, *28*(2), 16–46. https://doi.org/10.3102/0013189x028002016

Kultusministerkonferenz. (2005). Beschlüsse der Kultusministerkonferenz: Einheitliche Prüfungsanforderungen in der Abiturprüfung Musik. http://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/1989/1989_12_01-EPA-Musik.pdf

Kultusministerkonferenz. (2021a, August 23). Bildungsstandards der Kultusministerkonferenz. https://www.kmk.org/themen/qualitaetssicherung-in-schulen/bildungsstandards.html

Kultusministerkonferenz. (2021b, August 23). Überprüfung und Umsetzung von Bildungsstandards für die Primarstufe, die Sekundarstufe I und die Allgemeine Hochschulreife. https://www.kmk.org/themen/qualitaetssicherung-in-schulen/bildungsmonitoring/ueberpruefungumsetzung-der-bildungsstandards.html

Kurbacher-Schönborn, F. A. (2007). Urteil, ästhetisches. In J. Ritter, K. Gründer & G. Gabriel (Hrsg.), *Historisches Wörterbuch der Philosophie*. Schwabe.

Leutner, D., Fleischer, J., Grünkorn, J. & Klieme, E. (2017). *Competence assessment in education. Research, models and instruments*. Springer. https://doi.org/10.1007/978-3-319-50030-0

Lewis, D. M., Mitzel, H. C., Mercado, R. L., Patz, R. J. & Schulz, E. M. (2012). The bookmark standard setting procedure. In G. J. Cizek (Hrsg.), *Setting performance standards: Foundations, methods, and innovations* (2nd Edition, S. 225–253). Routledge.

Lill, F., Hasselhorn, J. & Lehmann, A. C. (2019). Der Zusammenhang von musikalischem Fähigkeitsselbstkonzept und musikpraktischen Kompetenzen in der Sekundarstufe I. In V. Weidner & C. Rolle (Hrsg.), *Praxen und Diskurse aus Sicht musikpädagogischer Forschung* (S. 171–187). Waxmann. https://www.pedocs.de/volltexte/2020/20711/

List of music considered the worst. (2021). https://en.wikipedia.org/wiki/List_of_music_considered_the_worst

Luger-Bazinger, C., Freunberger, R. & Itzlinger-Bruneforth, U. (2016). Standard-Setting. In S. Breit & C. Schreiner (Hrsg.), *Large-Scale Assessment mit R. Methodische Grundlagen der österreichischen Bildungsstandardüberprüfung* (S. 83–128). Facultas.

Mair, P., Hatzinger, R., Maier, M. J., Rusch, T. & Debelak, R. (2020). *eRm: Extended Rasch Modeling*. https://cran.r-project.org/package=eRm

Manzel, S. & Weißeno, G. (2017). Modell der politischen Urteilsfähigkeit – eine Dimension der Politikkompetenz. In M. Oberle & G. Weißeno (Hrsg.), *Politikwissenschaft und Politikdidaktik. Theorie und Empirie* (S. 59–86). Springer. https://doi.org/10.1007/978-3-658-07246-9_5

Meyer, T. (i. Vorb.). *Schüler*innen beschreiben Musik – Eine empirische Studie über das musikbezogene Sprechen von Schüler*innen der Sekundarstufe I und die Relevanz eines sprachbewussten Musikunterrichts*. [Dissertation, Hochschule für Musik, Theater und Medien Hannover].

Michalak, M., Lemke, V. & Goeke, M. (2015). *Sprache im Fachunterricht. Eine Einführung in Deutsch als Zweitsprache und sprachbewussten Unterricht*. Narr Francke Attempto.

Ministerium für Kultus, Jugend und Sport Baden-Württemberg. (2016). *Musik. Bildungsplan 2016 – Gemeinsamer Bildungsplan der Sekundarstufe I*. http://www.bildungsplaene-bw.de/site/bildungsplan/get/documents/lsbw/export-pdf/depot-pdf/ALLG/BP2016BW_ALLG_SEK1_MUS.pdf

Ministerium für Schule und Berufsbildung des Landes Schleswig-Holstein. (2015). *Fachanforderungen Musik. Allgemeinbildende Schulen. Sekundarstufe I. Sekundarstufe II*. Ministerium für Schule und Berufsbildung des Landes Schleswig-Holstein. https://lehrplan.lernnetz.de/index.php?DownloadID=1046

Ministerium für Schule und Bildung des Landes Nordrhein-Westfalen. (2019). *Musik. Kernlehrplan für das Gymnasium Sekundarstufe I in Nordrhein-Westfalen*. Ministerium für Schule und Bildung des Landes Nordrhein-Westfalen. https://www.schulentwicklung.nrw.de/lehrplaene/lehrplan/207/g9_mu_klp_%203406_2019_06_23.pdf

Moore, G. & Green, L. (2018). (Musik-)Soziologie. In M. Dartsch, J. Knigge, A. Niessen, F. Platz & C. Stöger (Hrsg.), *Handbuch Musikpädagogik* (S. 64–69). Waxmann.

Moosbrugger, H., Schermelleh-Engel, K., Gäde, J. C. & Kelava, A. (2020). Testtheorien im Überblick. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (3. Aufl., S. 251–273). Springer.

Morek, M. (2016). Formen mündlicher Darstellung in situ: Zur Komplexität von Diskursanforderungen in Unterrichtsgesprächen. In C. Bräuer & I. Winkler (Hrsg.), *Mündliches und schriftliches Handeln im Deutschunterricht* (S. 95–132). Peter Lang.

Morek, M. & Heller, V. (2012). Bildungssprache – Kommunikative, epistemische, soziale und interaktive Aspekte ihres Gebrauchs. *Zeitschrift für Angewandte Linguisik*, 67–101. https://doi.org/10.1515/zfal-2012-0011

Morek, M., Heller, V. & Quasthoff, U. (2017). Erklären und Argumentieren. Modellierungen und empirische Befunde zu Strukturen und Varianzen. In I. Meißner & E. L. Wyss (Hrsg.), *Begründen – Erklären – Argumentieren* (S. 11–46). Stauffenburg.

Müllensiefen, D., Gingras, B., Musil, J. & Stewart, L. (2014). The Musicality of Non-Musicians: An Index for Assessing Musical Sophistication in the General Population. *PLoS ONE*, *9*(2), e89642. https://doi.org/10.1371/journal.pone.0089642

Neumann, A. (2007). *Briefe schreiben in Klasse 9 und 11. Beurteilungskriterien, Messungen, Textstrukturen und Schülerleistungen*. Waxmann.

Neumann, A. & Lehmann, R. H. (2008). Schreiben Deutsch. In E. Klieme (Hrsg.), *Unterricht und Kompetenzerwerb in Deutsch und Englisch. Ergebnisse der DESI-Studie* (S. 89–103). Beltz.

Niessen, A., Lehmann-Wermser, A., Knigge, J. & Lehmann, A. C. (2008). Entwurf eines Kompetenzmodells ‚Musik wahrnehmen und kontextualisieren'. *Zeitschrift für Kritische Musikpädagogik*, *Sonderheft*. http://www.zfkm.org/?page_id=317

OECD. (2020). *Girls' and boys' performance in PISA*. OECD Publishing. https://doi.org/10.1787/f56f8c26-en

Parsons, M. J. (1987). *How we understand art. A cognitive developmental account of aesthetic experience*. Cambridge University Press.

Pitoniak, M. J. & Cizek, G. J. (2016). Standard Setting. In C. S. Wells & M. Faulkner-Bond (Hrsg.), *Educational Measurement. From Foundations to Future* (S. 38–61). Guilford Press.

Pohl, T. (2014). Schriftliches Argumentieren. In H. Feilke & T. Pohl (Hrsg.), *Schriftlicher Sprachgebrauch. Texte verfassen* (S. 287–315). Schneider.

Prediger, S., Erath, K., Quasthoff, U. M., Heller, V. & Vogler, A.-M. (2016). Befähigung zur Teilhabe an Unterrichtsdiskursen. Die Rolle von Diskurskompetenz. In J. Menthe, D. Höttecke, T. Zabka, M. Hammann & M. Rothgangel (Hrsg.), *Befähigung zu gesellschaftlicher Teilhabe. Beiträge der fachdidaktischen Forschung.* (S. 285–300). Waxmann.

Prediger, S. & Hein, K. (2017). Learning to meet language demands in multi-step mathematical argumentations: Design Research on a subject-specific genre. *European Journal of Applied Linguistics*, *5*, 309–335. https://doi.org/1.1515/eujal-2017-0010

Quasthoff, U. (2009). Entwicklung der mündlichen Kommunikationskompetenz. In M. Becker-Mrotzek (Hrsg.), *Mündliche Kommunikation und Gesprächsdidaktik* (S. 84–101). Schneider.

Quasthoff, U. & Domenech, M. (2016). Theoriegeleitete Entwicklung und Überprüfung eines Verfahrens zur Erfassung von Textqualität (TexQu) am Beispiel argumentativer Briefe in der Sekundarstufe I. *Didaktik Deutsch*, *41*, 21–43.

Quasthoff, U., Heller, V. & Morek, M. (2017). On the sequential organization and genre-orientation of discourse units in interaction: An analytic framework. *Discourse Studies*, *19*(1), 84–110. https://doi.org/10.1177/1461445616683596

Quasthoff, U., Heller, V. & Morek, M. (2020a). Diskurskompetenz und diskursive Partizipation als Schlüssel zur Teilhabe an Bildungsprozessen: Grundlegende Konzepte und Untersuchungslinien. In U. Quasthoff, V. Heller & M. Morek (Hrsg.), *Diskurserwerb in Familie, Peergroup und Unterricht: Passungen und Teilhabechancen* (S. 13–34). De Gruyter. https://doi.org/doi:10.1515/9783110707168-005

Quasthoff, U., Wild, E., Domenech, M., Hollmann, J., Kluger, C., Krah, A. & Otterpohl, N. (2020b). Familiale Ressourcen für den Erwerb von Argumentationskompetenz. In U. Quasthoff, V. Heller & M. Morek (Hrsg.), *Diskurserwerb in Familie, Peergroup und Unterricht: Passungen und Teilhabechancen* (S. 79–106). De Gruyter. https://doi.org/doi:10.1515/9783110707168-005

R Core Team. (2019). *R: A language and environment for statistical computing.* (R version 3.6.2). R Foundation for Statistical Computing. https://www.R-project.org/

Rapanta, C., Garcia-Mila, M. & Medina, S. (2013). What Is Meant by Argumentative Competence? An Integrative Review of Methods of Analysis and Assessment in Education. *Review of Educational Research*, *83*, 483–520. https://doi.org/10.3102/0034654313487606

Rauch, D. & Hartig, J. (2020). Interpretation von Testwerten in der Item-Response-Theorie (IRT). In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 411–424). Springer.

Reicher, M. E. (2005). *Einführung in die philosophische Ästhetik*. Wissenschaftliche Buchgesellschaft.

Robitzsch, A., Kiefer, T. & Wu, M. (2020). *TAM: Test analysis modules*. https://CRAN.R-project.org/package=TAM

Rolle, C. (1998). Was heißt „ästhetische Erfahrung"? Annäherungen an einen Grundbegriff der Ästhetik in musikdidaktischer Absicht. In M. Pfeffer, J. Vogt, U. Eckart-Bäcker & E. Nolte (Hrsg.), *Systematische Musikpädagogik* (S. 15–37). Wißner.

Rolle, C. (1999). *Musikalisch-ästhetische Bildung. Über die Bedeutung ästhetischer Erfahrung für musikalische Bildungsprozesse*. Bosse.

Rolle, C. (2008). Argumentationsfähigkeit: eine zentrale Dimension musikalischer Kompetenz? In H.-U. Schäfer-Lembeck (Hrsg.), *Leistung im Musikunterricht. Beiträge der Münchner Tagung 2008.* (S. 70–100). Allitera.

Rolle, C. (2013). Argumentation Skills in the Music Classroom: A Quest for Theory. In A. de Vugt & I. Malmberg (Hrsg.), *Artistry* (S. 137–150). Helbling.

Rolle, C. (2014a). Ästhetischer Streit als Medium des Musikunterrichts. Zur Bedeutung des argumentierenden Sprechens über Musik für ästhetische Bildung. *Art Education Research*, *5*(9). http://iae-journal.zhdk.ch/no-9/

Rolle, C. (2014b). Von Standards und Kompetenzen im Fach Musik. Ästhetische Bildung in Zeiten des Messens und Wiegens. In A. Nitschké & D. Sagrillo (Hrsg.), *Die Musik in der Bildung. Aspekte europäischer Musikerziehung und ihre Anwendung in Luxemburg* (S. 361–374). Margraf.

Rolle, C. (2017). Wie gut können wir über Geschmack streiten? In M. Schwarzbauer & M. Oebelsberger (Hrsg.), *Ästhetische Kompetenz – nur ein Schlagwort? Dokumentation einer Tagung der SOMA an der Universität Mozarteum Salzburg* (S. 127–145). LIT.

Rolle, C. & Wallbaum, C. (2011). Ästhetischer Streit im Musikunterricht. Didaktische und methodische Überlegungen zu Unterrichtsgesprächen über Musik. In J. Kirschenmann, C. Richter & K. H. Spinner (Hrsg.), *Reden über Kunst. Fachdidaktisches Forschungssymposium in Literatur, Kunst und Musik* (S. 507–535). Kopaed.

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2). https://doi.org/10.18637/jss.v048.i02

Salmon, W. C. (1973). *Logik*. Reclam.

Sälzer, C. (2016). *Studienbuch Schulleistungsstudien. Das Rasch-Modell in der Praxis*. Springer. https://doi.org/10.1007/978-3-662-45765-8

Schaal, N. K., Bauer, A.-K. R. & Müllensiefen, D. (2014). Der Gold-MSI: Replikation und Validierung eines Fragebogeninstrumentes zur Messung Musikalischer Erfahrenheit anhand einer deutschen Stichprobe. *Musicae Scientiae*, *18*(4), 423–447. https://doi.org/10.1177/1029864914541851

Schlarmann, J. & Galatsch, M. (2014). Regressionsmodelle für ordinale Zielvariablen. *GMS Medizinische Informatik, Biometrie und Epidemiologie*, *10*. https://doi.org/10.3205/mibe000154

Schmölzer-Eibinger, S. (2013). Sprache als Medium des Lernens im Fach. In M. Becker-Mrotzek, K. Schramm, E. Thürmann & H. J. Vollmer (Hrsg.), *Sprache im Fach. Sprachlichkeit und fachliches Lernen* (S. 25–40). Waxmann.

Seel, M. (1985). *Die Kunst der Entzweiung. Zum Begriff der ästhetischen Rationalität*. Suhrkamp.

Seel, M. (1991). *Eine Ästhetik der Natur*. Suhrkamp.

Seel, M. (1993). Zur ästhetischen Praxis der Kunst. *Deutsche Zeitschrift für Philosophie*, *41*, 31–43.

Sibley, F. (1965). Aesthetic and Nonaesthetic. *The Philosophical Review*, *74*, 135–159. https://www.jstor.org/stable/2183262

Staatsinstitut für Schulqualität und Bildungsforschung München. (2021). Lehrplanauszüge. Fachprofile. Gymnasium: Musik. https://www.lehrplanplus.bayern.de/fachprofil/gymnasium/musik

Stanat, P., Schipolowski, S., Mahler, N., Weirich, S. & Henschel, S. (2019). *IQB-Bildungstrend 2018. Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I im zweiten Ländervergleich*. Waxmann. https://doi.org/10.25656/01:18131

Stanat, P., Schipolowski, S., Rjosk, C., Weirich, S. & Haag, N. (2017). *IQB-Bildungstrend 2016. Kompetenzen in den Fächern Deutsch und Mathematik am Ende der 4. Jahrgangsstufe im zweiten Ländervergleich*. Waxmann. https://doi.org/10.25656/01:15477

Tetens, H. (2004). *Philosophisches Argumentieren. Eine Einführung*. C. H. Beck.

Tillmann, K.-J. (2017). Empirische Bildungsforschung in der Kritik – ein Überblick über Themen und Kontroversen. In J. Baumert & K.-J. Tillmann (Hrsg.), *Empirische Bildungsforschung. Der kritische Blick und die Antwort auf die Kritiker* (S. 5–22). Springer. https://doi.org/10.1007/978-3-658-13785-4_2

Toulmin, S. E. (2003). *The Uses of Argument* (Updated Edition). Cambridge University Press.

Trendtel, M., Pham, G. & Yanagida, T. (2016a). Skalierung und Linking. In S. Breit & C. Schreiner (Hrsg.), *Large-Scale Assessment mit R. Methodische Grundlagen der österreichischen Bildungsstandardüberprüfung* (S. 185–224). Facultas.

Trendtel, M., Schwabe, F. & Fellinger, R. (2016b). Differenzielles Itemfunktionieren in Subgruppen. In S. Breit & C. Schreiner (Hrsg.), *Large-Scale Assessment mit R. Methodische Grundlagen der österreichischen Bildungsstandardüberprüfung* (S. 111–147). Facultas.

Vogt, J. (2008). Musikbezogene Bildungskompetenz – ein hölzernes Eisen? Anmerkungen zu den Theoretischen Überlegungen zu einem Kompetenzmodell für das Fach Musik. *Zeitschrift für Kritische Musikpädagogik*, (Sonderedition: Bildungsstandards und Kompetenzmodelle für das Fach Musik?). https://zfkm.org/sonder08-vogt.pdf

Vollmer, H. J. & Thürmann, E. (2010). Zur Sprachlichkeit des Fachlernens: Modellierung eines Referenzrahmens für Deutsch als Zweitsprache. In B. Ahrenholz (Hrsg.), *Fachunterricht und Deutsch als Zweitsprache* (S. 107–132). Narr. http://www.oesz.at/download/Artikel_Prof.Vollmer.pdf

Wallbaum, C. (2009). *Produktionsdidaktik im Musikunterricht. Perspektiven zur Gestaltung ästhetischer Erfahrungssituationen* (2. Aufl.). Qucosa. https://nbn-resolving.org/urn:nbn:de:bsz:14-qucosa-24736

Warrens, M. (2013). Cohen's weighted kappa with additive weights. *Advances in Data Analysis and Classification*, *7*, 41–55. https://doi.org/10.1007/s11634-013-0123-9

Weinert, F. E. (2001a). Concept of Competence: A Conceptual Clarification. In D. S. Rychen & L. H. Salganik (Hrsg.), *Defining and selecting key competencies* (S. 45–65). Hogrefe.

Weinert, F. E. (2001b). Vergleichende Leistungsmessung in der Schule – eine umstrittene Selbstverständlichkeit. In F. E. Weinert (Hrsg.), *Leistungsmessungen in Schulen* (S. 17–31). Beltz.

Willenberg, H., Gailberger, S. & Krelle, M. (2007). Argumentation. In B. Beck & E. Klieme (Hrsg.), *Sprachliche Kompetenzen Konzepte und Messung; DESI-Studie (Deutsch-Englisch-Schülerleistungen International) Beltz Paedagogik* (S. 118–129). Beltz.

Wirtz, M. & Caspar, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität*. Hogrefe.

Wolf, A. (2016). *„Es hört doch jeder nur, was er versteht". Konstruktion eines kompetenzbasierten Assessments für Gehörbildung*. Wissenschaftlicher Verlag.

Wu, M., Tam, H. P. & Jen, T.-H. (2016). *Educational Measurement for Applied Researchers. Theorie into Practice*. Springer.

Zehner, F., Sälzer, C. & Goldhammer, F. (2016). Automatic Coding of Short Text Responses via Clustering in Educational Assessment. *Educational and Psychological Measurement*, *76*(2), 280–303. https://doi.org/10.1177/0013164415590022

Anhang

A. Ergänzende Tabellen und Abbildungen

Tabelle A.1.
Verteilung der Items auf die Testhefte

	Testheft I	Testheft II	Testheft III
1.	Eisenbahn	Star Wars	Ballade
2.	Cinderella	Kinder	Phantom
3.	Hip-Hop	Echo	Königin
4.	ESC	Zahnarzt	Shape
5.	Band	Vivaldi	Barbar
6.	Böhmi	Warschau	America
7.	Bach	Tom	Schubert
8.	Lola	Knabe	Crossover
9.	Star Wars	Ballade	Holländer
10.	Kinder	Phantom	Eisenbahn
11.	Echo	Königin	Cinderella
12.	Zahnarzt	Shape	Hip-Hop
13.	Vivaldi	Barbar	ESC
14.	Warschau	America	Band
15.	Tom	Schubert	Böhmi
16.	Knabe	Crossover	Bach
17.	Ballade	Holländer	Lola
18.	Phantom	Eisenbahn	Eisenbahn
19.	Königin	Cinderella	Star Wars
20.	Shape	Hip-Hop	Kinder
21.	Barbar	ESC	Echo
22.	America	Band	Zahnarzt
23.	Schubert	Böhmi	Vivaldi
24.	Crossover	Bach	Warschau
25.	Holländer	Lola	Tom

Anmerkung. Alle Testhefte enthielten alle Items nur in unterschiedlicher Reihenfolge. Aus den Items wurden drei Blöcke gebildet, die in den Testheften unterschiedlich angeordnet waren. Das Item „America“ wurde nach dem 12.06.2019 aus dem Test entfernt. Das Video, das in der Aufgabe gezeigt wurde, war jugendschutzkonform. Trotzdem kam es zu einer Elternbeschwerde.

Tabelle A.2.
Interrater-Reliabilität aller Testaufgaben

Item	κ
america	.89
bach	.82
ballade	.80
band	.89
barbar	.79
böhmi	.86
cinderella	.87
cross	.85
echo	.85
eisenbahn	.92
esc	.94
hiphop	.81
holl	.82
kinder	.77
knabe	.85
königin	.92
lola	.87
phantom	.86
schubert	.79
shape	.73
star	.78
tom	.77
vivaldi	.83
warschau	.82
zahnarzt	.94

Anmerkung. Zwei Rater*innen kodierten ca. 15 % der Antworten jeder Testaufgabe. Bei guter bis sehr guter Übereinstimmung kodierte eine Person die restlichen Daten. Die Berechnung von Cohens κ erfolgte mit linear weights, wobei Werte $\geq .75$ als sehr gute Übereinstimmung gelten (Wirtz & Caspar, 2002, S. 59).

Tabelle A.3.
Thurstonian Thresholds (Lösungswahrscheinlichkeit 65 %)

Item	Kategorie 1	Kategorie 2	Kategorie 3
america	0.15	3.46	
bach	-0.13	2.63	
ballade	0.22	1.29	2.99
band	-1.90	0.70	2.73
barbar	-2.11	1.67	
boehmi	1.03		
cinderella	-1.57	-0.33	
cross	-0.30	2.99	
echo	-0.04	2.81	
eisenbahn	-0.25	1.91	
esc	0.45	3.21	
hiphop	-1.37	0.94	
holl	-0.20	0.96	
kinder	-0.83	1.76	
knabe	-2.02	2.89	
koenigin	-1.44	3.41	
lola	-1.51		
phantom	-1.02	2.88	
schubert	-0.63	2.22	
shape	-1.90	1.00	3.86
star	-1.43	1.05	3.44
tom	0.09	1.55	
vivaldi	-1.14	0.13	3.40
warschau	-0.22	2.96	
zahnarzt	-1.27	2.82	

Tabelle A.4.
Itemfit-Werte

Item	Outfit	Outfit_t	Outfit_p	Infit	Infit_t	Infit_p
america	1.15	1.64	0.10	1.15	1.90	0.06
bach	0.88	-1.26	0.21	0.91	-1.15	0.25
ballade	1.13	1.13	0.26	1.11	1.36	0.17
band	1.06	0.77	0.44	1.07	0.90	0.37
barbar	0.96	-0.53	0.60	0.95	-0.69	0.49
boehmi	0.87	-1.06	0.29	0.93	-1.07	0.29
cinderella	1.36	2.19	0.03	1.22	2.72	0.01
cross	0.98	-0.22	0.83	0.99	-0.14	0.89
echo	1.11	1.37	0.17	1.13	1.81	0.07
eisenbahn	1.05	0.67	0.51	1.07	1.14	0.26
esc	1.03	0.36	0.72	1.04	0.62	0.54
hiphop	1.22	2.46	0.01	1.13	1.85	0.06
holl	0.90	-0.82	0.41	0.95	-0.64	0.53
kinder	0.98	-0.18	0.85	1.00	0.02	0.98
knabe	0.91	-0.82	0.41	0.92	-0.84	0.40
koenigin	0.90	-1.06	0.29	0.91	-1.03	0.30
lola	0.85	-0.72	0.47	0.94	-0.59	0.56
phantom	0.90	-1.37	0.17	0.90	-1.45	0.15
schubert	0.89	-1.35	0.18	0.92	-1.04	0.30
shape	1.06	0.81	0.42	1.05	0.74	0.46
star	0.98	-0.20	0.85	0.97	-0.38	0.71
tom	1.15	1.29	0.20	1.11	1.43	0.15
vivaldi	0.96	-0.44	0.66	0.96	-0.58	0.56
warschau	0.78	-2.88	> 0.01	0.79	-2.99	> 0.01
zahnarzt	0.90	-1.28	0.20	0.90	-1.31	0.19

Tabelle A.5.
DIF-Analysen Geschlecht

parameter	facet	xsi	se.xsi	z	effect size
america:gender_male__	item:gender	-0.28	0.08	-3.36	-0.55
bach:gender_male__	item:gender	0.03	0.08	0.36	0.06
ballade:gender_male__	item:gender	-0.09	0.06	-1.38	-0.18
band:gender_male__	item:gender	0.15	0.07	2.16	0.31
barbar:gender_male__	item:gender	0.06	0.08	0.79	0.13
boehmi:gender_male__	item:gender	-0.06	0.09	-0.66	-0.12
cinderella:gender_male__	item:gender	0.08	0.07	1.16	0.17
cross:gender_male__	item:gender	-0.03	0.08	-0.32	-0.05
echo:gender_male__	item:gender	-0.15	0.07	-2.03	-0.30
eisenbahn:gender_male__	item:gender	-0.36	0.07	-5.18	-0.72
esc:gender_male__	item:gender	0.03	0.08	0.37	0.06
hiphop:gender_male__	item:gender	-0.05	0.07	-0.72	-0.11
holl:gender_male__	item:gender	0.00	0.07	0.07	0.01
kinder:gender_male__	item:gender	0.00	0.07	-0.01	0.00
knabe:gender_male__	item:gender	0.16	0.09	1.77	0.31
koenigin:gender_male__	item:gender	0.16	0.08	1.91	0.32
lola:gender_male__	item:gender	0.06	0.09	0.66	0.12
phantom:gender_male__	item:gender	0.03	0.08	0.43	0.07
schubert:gender_male__	item:gender	-0.08	0.08	-1.03	-0.16
shape:gender_male__	item:gender	-0.02	0.07	-0.23	-0.03
star:gender_male__	item:gender	-0.02	0.07	-0.28	-0.04
tom:gender_male__	item:gender	0.09	0.07	1.18	0.17
vivaldi:gender_male__	item:gender	-0.05	0.07	-0.80	-0.11
warschau:gender_male__	item:gender	0.09	0.08	1.12	0.18
zahnarzt:gender_male__	item:gender	0.23	0.38	0.61	0.46

Anmerkung. Es wurde ein Facetten-Modell mit einem Interaktionsterm (‚item-by-group') berechnet (Link zum *R*-Skript auf S. 17). Laut den Richtlinien des Educational Testing Service (ETS) liegt ein starker DIF ab einer Effektstärke von ≥ 0.638 vor (Signifikanzniveau von .05) (Trendtel et al., 2016b, S. 127–131).

Tabelle A.6.
DIF-Analysen Sprache

parameter	facet	xsi	se.xsi	z	effect size
america:language_home1	item:language_home	0.06	0.08	0.76	0.12
bach:language_home1	item:language_home	-0.24	0.08	-2.98	-0.48
ballade:language_home1	item:language_home	-0.01	0.06	-0.21	-0.03
band:language_home1	item:language_home	0.03	0.07	0.48	0.07
barbar:language_home1	item:language_home	-0.13	0.08	-1.65	-0.26
boehmi:language_home1	item:language_home	-0.22	0.09	-2.51	-0.44
cinderella:language_home1	item:language_home	0.09	0.07	1.35	0.19
cross:language_home1	item:language_home	-0.15	0.08	-1.91	-0.30
echo:language_home1	item:language_home	-0.06	0.07	-0.82	-0.12
eisenbahn:language_home1	item:language_home	0.14	0.07	1.99	0.27
esc:language_home1	item:language_home	0.28	0.08	3.68	0.56
hiphop:language_home1	item:language_home	0.09	0.07	1.28	0.18
holl:language_home1	item:language_home	0.00	0.07	0.04	0.00
kinder:language_home1	item:language_home	0.06	0.07	0.89	0.13
knabe:language_home1	item:language_home	0.11	0.08	1.29	0.22
koenigin:language_home1	item:language_home	-0.05	0.08	-0.59	-0.10
lola:language_home1	item:language_home	-0.07	0.09	-0.78	-0.14
phantom:language_home1	item:language_home	-0.20	0.08	-2.58	-0.39
schubert:language_home1	item:language_home	-0.05	0.08	-0.61	-0.09
shape:language_home1	item:language_home	0.15	0.07	2.07	0.29
star:language_home1	item:language_home	0.03	0.07	0.41	0.06
tom:language_home1	item:language_home	0.19	0.07	2.72	0.39
vivaldi:language_home1	item:language_home	0.00	0.07	0.06	0.01
warschau:language_home1	item:language_home	-0.10	0.08	-1.33	-0.21
zahnarzt:language_home1	item:language_home	0.04	0.37	0.10	0.07

Anmerkung. DIF-Analyse für die Gruppenvariable „Häuslicher Sprachgebrauch“. Für die Frage „Wie of sprichst du zuhause Deutsch“ gab es vier Antwortmöglichkeiten: (1) immer, (2) meistens, (3) meistens eine andere Sprache, (4) nie. Für die DIF-Analysen wurden zwei Gruppen gebildet aus den Antwortmöglichkeiten (1) und (2-4). Es wurde ein Facetten-Modell mit einem Interaktionsterm (‚item-by-group‘) berechnet (Link zum *R*-Skript auf S. 17). Laut den Richtlinien des Educational Testing Service (ETS) liegt ein starker DIF ab einer Effektstärke von ≥ 0.638 vor (Signifikanzniveau von .05) (Trendtel et al., 2016b, S. 127–131).

Tabelle A.7.
DIF-Analysen Instrumentalunterricht

parameter	facet	xsi	se.xsi	z	effect size
america:instrument_lessons1	item:instrument_lessons	0.35	0.08	4.31	0.69
bach:instrument_lessons1	item:instrument_lessons	-0.22	0.08	-2.69	-0.44
ballade:instrument_lessons1	item:instrument_lessons	0.08	0.06	1.28	0.16
band:instrument_lessons1	item:instrument_lessons	0.18	0.07	2.66	0.36
barbar:instrument_lessons1	item:instrument_lessons	-0.19	0.08	-2.35	-0.37
boehmi:instrument_lessons1	item:instrument_lessons	-0.11	0.09	-1.26	-0.22
cinderella:instrument_lessons1	item:instrument_lessons	0.19	0.07	2.74	0.38
cross:instrument_lessons1	item:instrument_lessons	0.14	0.08	1.84	0.28
echo:instrument_lessons1	item:instrument_lessons	0.11	0.07	1.49	0.22
eisenbahn:instrument_lessons1	item:instrument_lessons	0.05	0.07	0.68	0.09
esc:instrument_lessons1	item:instrument_lessons	0.22	0.08	2.87	0.44
hiphop:instrument_lessons1	item:instrument_lessons	0.22	0.07	3.09	0.44
holl:instrument_lessons1	item:instrument_lessons	0.06	0.07	0.86	0.12
kinder:instrument_lessons1	item:instrument_lessons	0.04	0.07	0.58	0.08
knabe:instrument_lessons1	item:instrument_lessons	-0.23	0.09	-2.68	-0.46
koenigin:instrument_lessons1	item:instrument_lessons	-0.31	0.08	-3.80	-0.63
lola:instrument_lessons1	item:instrument_lessons	0.30	0.09	3.28	0.60
phantom:instrument_lessons1	item:instrument_lessons	-0.20	0.08	-2.63	-0.40
schubert:instrument_lessons1	item:instrument_lessons	-0.23	0.08	-2.98	-0.46
shape:instrument_lessons1	item:instrument_lessons	-0.21	0.07	-2.96	-0.42
star:instrument_lessons1	item:instrument_lessons	-0.09	0.07	-1.35	-0.18
tom:instrument_lessons1	item:instrument_lessons	0.05	0.07	0.74	0.11
vivaldi:instrument_lessons1	item:instrument_lessons	-0.03	0.07	-0.46	-0.06
warschau:instrument_lessons1	item:instrument_lessons	-0.06	0.08	-0.81	-0.13
zahnarzt:instrument_lessons1	item:instrument_lessons	-0.11	0.37	-0.29	-0.21

Anmerkung. DIF-Analyse für das Item „Bekommst du Instrumental- oder Gesangsunterricht?“. Die Frage konnte mit „ja“ und „nein“ beantwortet werden. Es wurde ein Facetten-Modell mit einem Interaktionsterm (‚item-by-group‘) berechnet (Link zum *R*-Skript auf S. 17). Laut den Richtlinien des Educational Testing Service (ETS) liegt ein starker DIF ab einer Effektstärke von ≥ 0.638 vor (Signifikanzniveau von .05) (Trendtel et al., 2016b, S. 127–131).

Tabelle A.8.
Post-hoc-Tests Personenfähigkeitswerte

Klasse	*M*	*SD*	*n*	Gepaarte *t*-Tests							
				9. Klasse		10. Klasse		11. Klasse		12. Klasse	
				p	Cohen's *d*	*p*	Cohen's *d*	p value	Cohen's *d*	*p*	Cohen's *d*
9. Klasse	-0.85	1.15	147								
10. Klasse	-0.06	1.27	108	< .001	0.655						
11. Klasse	0.31	1.43	126	< .001	0.905	.278					
12. Klasse	1.70	1.58	24	< .001	2.096	< .001	1.324	< .001	0.958		
Studierende	1.81	1.18	34	< .001	2.299	< .001	1.493	< .001	1.085	1.000	

Anmerkung. Bonferroni Korrektur; pooled *SD*

B. Publikationen und Darlegung des eigenen Arbeitsanteils

Publikation A

Ehninger, J. (2021). Wie lässt sich musikbezogene Argumentationskompetenz empirisch untersuchen? Über die empirische Erforschung einer facettenreichen Kompetenz. *Beiträge empirischer Musikpädagogik*, *12*. https://www.b-em.info/index.php/ojs/article/view/192

Julia Ehninger verfasste die Publikation als alleinige Autorin. Die Publikation wurde im Anschluss an ein double-blind peer-review Verfahren von der Autorin überarbeitet und veröffentlicht.

Publikation B

Ehninger, J. & Rolle, C. (2020). Musikbezogenes Argumentieren – Nur Geschmacksache? Über die Entwicklung eines Kompetenztests. In M. Schwarzbauer & K. Steinhauser (Hrsg.), *‚Nur' Geschmacksache? Der Umgang mit kreativen Leistungen im Musik- und Kunstunterricht* (S. 168–182). LIT.

Dieser Artikel behandelt die Entwicklung und das Design des *MARKO*-Tests. Der theoretische Hintergrund der Studie geht auf Arbeiten von Prof. Dr. Christian Rolle zurück. Die Verantwortung für die Organisation, Durchführung und Auswertung der Pilotstudie lag bei Julia Ehninger. Die Aufgaben für den ersten Pretest wurden von Julia Ehninger sowie von Studierenden zweier Seminare entwickelt (Musikhochschule Lübeck und Universität zu Köln, Sommersemester 2017) und im Projektteam, bestehend aus Prof. Dr. Christian Rolle, Prof. Dr. Jens Knigge, Thomas Gottschalk und Julia Ehninger, überarbeitet. Studierende der beiden Seminare erhoben die Daten des ersten Pretests. Julia Ehninger wertete die Daten aus und überarbeitete die Testaufgaben. Lediglich elf Aufgaben konnten in überarbeiteter Fassung für die zweite Phase der Pilotierung übernommen werden, weshalb neue Aufgaben von der Doktorandin entwickelt wurden. Diese wurden in regelmäßigen Abschnitten im Projektteam diskutiert. Julia Ehninger erhob allein die Daten des zweiten Pretests und entwickelte die Kodierregel der in der Publikation abgebildeten Aufgabe. Die Schüler*innen-Antworten der besagten Aufgabe wurden von der Doktorandin sowie von zwei weiteren Kodierer*innen bewertet. Das Manuskript wurde von Julia Ehninger verfasst und in Zusammenarbeit mit Prof. Dr. Christian Rolle überarbeitet. Das eingereichte Manuskript wurde von den Herausgeberinnen kommentiert, unterlag jedoch keinem peer-review.

Publikation C

Ehninger, J., Knigge, J. & Rolle, C. (2021a). Musikbezogene Argumentationskompetenz. Ein Werkstattbericht über die Entwicklung von Testaufgaben. In A. Budke & F. Schäbitz (Hrsg.), *Argumentieren und Vergleichen* (S. 93–112). LIT.

Dieser Artikel behandelt ebenfalls die Entwicklung und das Design des *MARKO*-Tests. Die Verantwortung für die Organisation, Durchführung und Auswertung der Pilotstudie lag bei Julia Ehninger. Alle weiteren Details können der Beschreibung des Arbeitsanteils von Publikation B entnommen werden (s. o.). Die Organisation, Datenerhebung und Datenauswertung der Hauptstudie lag ebenfalls federführend bei Julia Ehninger. Alina Ottinger kodierte gemeinsam mit Julia Ehninger alle Testaufgaben der Hauptstudie und wirkte bei der Überarbeitung der Kodierregeln mit. Prof. Dr. Christian Rolle verfasste Abschnitt 1.1 des Manuskripts, die übrigen Abschnitte wurden von Julia Ehninger erstellt. Im Autorenteam wurde der Artikel intern mehrmals überarbeitet. Die Revisionen im Anschluss an das (nicht anonyme) peer-review Verfahren wurden ebenfalls im Autor*innenteam durchgeführt.

Publikation D

Ehninger, J., Knigge, J., Schurig, M. & Rolle, C. (2021b). A New Measurement Instrument for Music-Related Argumentative Competence: The MARKO Competency Test and Competency Model. *Frontiers in Education*, *6*(191). https://doi.org/10.3389/feduc.2021.668538

In dieser Publikation sind die zentralen Ergebnisse der Hauptstudie dargestellt. Der Arbeitsanteil der Doktorandin bei der Organisation, Durchführung und Datenerhebung von Pilot- und Hauptstudie wurde bereits in den vorangegangenen Publikationen ausgeführt. Die Durchführung der statistischen Analysen wurde von Prof. Dr. Jens Knigge und Dr. Michael Schurig betreut. Julia Ehninger verfasste die erste Version des Manuskripts, welches die anderen Autoren überarbeiteten. Die Revisionen im Anschluss an das double-blind peer-review Verfahren wurden ebenfalls im Autor*innenteam durchgeführt.

C. Erklärung

Ich versichere eidesstattlich, dass ich die von mir vorgelegte Dissertation selbständig und ohne unzulässige Hilfe angefertigt, die benutzten Quellen und Hilfsmittel vollständig angegeben und die Stellen der Arbeit einschließlich Tabellen, Karten und Abbildungen, die anderen Werken im Wortlaut oder dem Sinn nach entnommen sind, in jedem Einzelfall als Entlehnung kenntlich gemacht habe sowie dass diese Dissertation noch keinem anderen Fachbereich zur Prüfung vorgelegen hat. Die Promotionsordnung ist mir bekannt. Die von mir vorgelegte Dissertation ist von Prof. Dr. Christian Rolle und Prof. Dr. Jens Knigge betreut worden.

Köln, im September 2021 Eva-Julia Ehninger