Jordan, Anne-Katrin; Lehmann, Andreas C.; Knigge, Jens

Kompetenzmodellierung mit Methoden der Item-Response-Theorie (IRT). Erste Ergebnisse der Validierung eines Modells für den Bereich "Musik wahrnehmen und kontextualisieren"

formal überarbeitete Version der Originalveröffentlichung in:
formally revised edition of the original source in:
Knolle, Niels [Hrsg.]: Evaluationsforschung in der Musikpädagogik. Essen : Die Blaue Eule 2010, S. 109-129. - (Musikpädagogische Forschung; 31)

Musikpädagogische Forschung

Niels Knolle
(Hrsg.)

Evaluationsforschung in der Musikpädagogik

verlag DIE BLAUE EULE essen

Themenstellung: Evaluationsforschung ist zu einem bedeutsamen Zweig der Bildungsforschung geworden, die Vielfalt der Beiträge zur 31. AMPF-Tagung >Evaluationsforschung in der Musikpädagogik< macht deutlich, dass die musikpädagogische Forschung hierzu einen bedeutsamen Beitrag zu liefern in der Lage ist. So zielen die Beiträge dieses Bands darauf, die Voraussetzungen, Inhalte, Methoden und Resultate von musikunterrichtlichen Reformansätzen und Innovationen im Blick auf die mit ihnen verbundenen Ziele zu überprüfen und zu bewerten, um so zu einer Verbesserung des musikbezogenen Handelns bzw. entsprechender Lehr-Lern-Prozesse zu gelangen.

Der Herausgeber: *Niels Knolle*, geb. 1944. Arbeitsschwerpunkte: Multimedia als Instrument, Werkzeug und Thema des Musikunterrichts; Didaktik der Populären Musik; Bildungsreformen in der Musikpädagogik. Langjährige Arbeit in den Vorständen der BFG Musikpädagogik, des AMPF, der Bundesfachausschüsse >Musikpädagogik< und >Musik und Medien< des Deutschen Musikrats. 1999 - 2003 Mitherausgeber der Zeitschrift >Musik in der Schule<. Von 1996 bis 2010 Universitätsprofessor für Musikpädagogik an der Otto-von-Guericke-Universität Magdeburg.

Inhalt

ANNE-KATRIN JORDAN, ANDREAS C. LEHMANN, JENS KNIGGE

Kompetenzmodellierung mit Methoden der Item-Response-Theorie (IRT)

Erste Ergebnisse der Validierung eines Modells für den Bereich „Musik wahrnehmen und kontextualisieren“

1 Einleitung

Seit etwa 10 Jahren gibt es einen Trend in der empirischen Bildungsforschung, der maßgeblich durch die Diskussionen um die Ergebnisse der PISA-Studien angestoßen wurde. Seither haben Bildungsforscher das Konzept der Kompetenzen zu einem zentralen Thema gemacht und damit eine Abkehr in der Betrachtung vollzogen von einem traditionell am Input orientierten Verständnis von Bildungsvorgängen hin zu einer am Output orientierten Sichtweise. Dieser Paradigmenwechsel hat in vielen Didaktikfächern tief greifende und sichtbare Veränderungen nach sich gezogen. Auch in der Musikpädagogik sind in den letzten Jahren zum Teil kontroverse Diskussionen geführt worden (Bähr 2004, Richter 2007). Diese Auseinandersetzung begreift sich als eher konzeptionell, und es geht ihr weniger um den Anschluss an Bemühungen der empirischen Erforschung, wie sie in den Erziehungswissenschaften derzeit üblich sind.

Im Rahmen der Diskussion um Bildungsstandards und Kompetenzorientierung kommt der Modellierung von Kompetenzen eine besondere Rolle zu. In dem musikpädagogischen Projekt KoMus, von dem hier die Rede sein soll, ging es darum, zunächst einmal auf der Grundlage von Überlegungen aus der Fachdidaktik heraus ein vorläufiges theoretisches Modell zu entwickeln, wie weiter unten näher beschrieben wird. Dabei fanden neben der Sichtung curricularen Materials (Knigge, Lehmann-Wermser 2008) auch Forschungsergebnisse der Musikpädagogik und Musikpsychologie Berücksichtigung (z. B. Bruhn 2005, La Motte-Haber 2005, Lange 2005, Serafine 1988, Stoffer 2005). Das resultierende vorläufige Modell wurde dann als Konstrukt zum Zweck der späteren empirischen Überprüfung mit Hilfe von Testaufgaben operationalisiert. Wie diese Operationalisierung im Projekt KoMus erfolgte, wird an ande-

rer Stelle beschrieben (Knigge, Niessen & Jordan 2010). Nach der aufwändigen Entwicklung von Aufgaben für eine bestimmte Zielgruppe (Schüler der Klassenstufe sechs) konnten diese als Test zusammengefasst einer größeren Anzahl von Schülern vorgelegt werden. Die dabei anfallenden Daten waren die Grundlage für eine Modellierung des Kompetenzmodells, um die es in diesem Beitrag gehen wird.

Während es in den schulischen ‚Kernfächern' Mathematik, Deutsch und Englisch bereits groß angelegte Studien (‚large-scale assessments') zu den von Schülerinnen und Schülern[1] erworbenen Kompetenzen gibt (z. B. TIMSS, DESI, IGLU), liegen für die deutsche Musikpädagogik erst seit etwa drei Jahren Überlegungen zur empirischen Erforschung von Kompetenzen im Fach Musik vor (Niessen et al. 2008). Hierbei wurde mit Methoden der Item-Response-Theorie gearbeitet (IRT, s. u. für weitere Details), während in der Vergangenheit Leistungstests auf der Basis der klassischen Testtheorie (KTT) ausgewertet wurden (z. B. Lohmann 1982, Bähr 2001). Im Vordergrund des vorliegenden Beitrages steht die empirische Validierung des theoretischen Kompetenzmodells durch Überprüfung seiner dimensionalen Struktur. Einleitend soll kurz dargestellt werden, was allgemein unter Kompetenzmodellen verstanden wird und welche spezifischen Formen von Kompetenzmodellen differenziert werden können.

Im Zentrum unserer Ausführungen stehen die empirische Validierung des theoretischen Kompetenzmodells und damit die Überprüfung seiner dimensionalen Struktur. Hierbei wird hinterfragt, ob die von uns im Test angelegten inhaltlichen Dimensionen auch empirisch nachweisbar sind. Bevor wir einige Ergebnisse aus dem KoMus-Projekt berichten, wollen wir jedoch zum besseren Verständnis beschreiben, was allgemein unter Kompetenzmodellen verstanden wird und welche spezifischen Formen von Kompetenzmodellen es gibt.

Kompetenzmodelle

Auch wenn eine ausführliche Diskussion des Kompetenzbegriffs hier nicht erfolgen kann, sei darauf verwiesen, dass dem vorliegenden Projekt der Kompetenzbegriff nach Weinert (2001) zugrunde liegt, wie er auch im aktuellen erziehungswissenschaftlichen und pädagogisch-psychologischen Diskurs Ver-

1 Aus Gründen der leichteren Lesbarkeit verzichten wir im Folgenden auf die Nennung beider Geschlechter.

wendung findet (Klieme et al. 2003, vgl. auch Knigge & Lehmann-Wermser 2009). Kompetenzen sind danach „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 2001, S. 27). Ein Kompetenzmodell unterscheidet „Teildimensionen innerhalb einer Domäne […] und es beschreibt jeweils unterschiedliche Niveaustufen auf solchen Dimensionen. Jede Kompetenzstufe ist durch kognitive Prozesse und Handlungen von bestimmter Qualität spezifiziert, die Schüler auf dieser Stufe bewältigen können, nicht aber Schüler auf niedrigeren Stufen“ (Klieme et al. 2003, S. 24). Jede dieser Teildimensionen, z. B. die Fähigkeit des Heraushörens eines Instruments, kann durch verschiedene Niveaustufen zunehmenden Anspruchs beschrieben werden. Im Hinblick auf das Heraushören eines Instruments hieße das etwa, ein Solo-Instrument an seinem Klang zu erkennen bzw. einzelne Instrumente aus einem Orchesterklang herauszuhören. Um eine solche Beschreibung vornehmen zu können, muss das zunächst theoretische Modell aber operationalisiert werden, d. h., es müssen Aufgaben entwickelt werden, die die Aspekte des Modells konkretisieren (vgl. Knigge et al. 2010).

Ziel eines Kompetenzmodells ist erstens die Beschreibung von Anforderungen, „deren Bewältigung von Schülerinnen und Schülern erwartet wird“ und zweitens die Entwicklung begründeter Vorstellungen darüber, „welche Abstufungen eine Kompetenz annehmen kann bzw. welche Grade oder Niveaustufen sich bei den einzelnen Schülerinnen und Schülern feststellen lassen“ (Klieme et al. 2003, S. 74). „Das Erreichen einer Kompetenzstufe sagt etwas darüber aus, welche Handlungen und mentalen Operationen mit hoher Wahrscheinlichkeit korrekt ausgeführt werden können“ (Klieme et al. 2003, S. 23). Kompetenzmodelle sind also nicht gleichzusetzen mit Leistungstests im traditionellen Sinn; vielmehr sind sie wissenschaftliche Konstrukte.

Zusammenfassend stellen Kompetenzmodelle eine wichtige vermittelnde Schnittstelle zwischen verallgemeinerten Zielen von Unterricht und Testaufgaben dar. Sie geben Hinweise auf die Aufgabenkonstruktion, „indem sie kognitive Leistungen mit unterschiedlichem Schwierigkeitsniveau spezifizieren. Umgekehrt ermöglicht erst die Einordnung in ein Kompetenzmodell, zu verstehen, was das Lösen oder Nichtlösen einer Aufgabe bedeutet […]. Modellvorstellungen, die den Aufbau von Kompetenzen über mehrere Stufen hinweg charakterisieren, sind demnach wichtige Orientierungen für die Unter-

richtspraxis und die Bewertung von Lernergebnissen“ (Klieme et al. 2003, S. 24).

Formen von Kompetenzmodellen

In der empirischen Bildungsforschung unterscheidet man zwei Formen von Kompetenzmodellen: *Kompetenzstruktur-* und *Kompetenzniveaumodelle* (Hartig, Klieme 2006; vgl. auch Jordan, Knigge i. Druck). Erstere beschreiben die Grundstruktur bzw. den Aufbau einer Kompetenz ausgehend von den statistischen Zusammenhängen zwischen den auf ein Modell bezogenen Skalen und Tests (Hartig, Klieme 2006; s. auch Abb. 5). Mithilfe faktorenanalytischer Verfahren werden Aufgaben, die hohe Zusammenhänge untereinander zeigen, zu Dimensionen zusammengefasst. Man geht somit davon aus, dass die Aufgaben dasselbe Merkmal erfassen (Hartig, Klieme 2006). Bei der Entscheidung für ein bestimmtes Strukturmodell (Beispiel 3-dimensionales Modell s. linker Kasten Abb. 1), also der Frage, wie viele Dimensionen man differenzieren möchte und wie viele sinnvoll sind, muss jeweils eine Abwägung pragmatischer und theoretischer Aspekte vorgenommen werden. Kompetenzstrukturmodelle können sich auch, statt mit der Kompetenz als solcher, mit der Binnenstruktur einzelner Kompetenzbereiche befassen, d. h. mit den zugrunde liegenden Teilkompetenzen und ihren Zusammenhängen (Hartig, Klieme 2007; s. auch rechter Kasten in Abb. 1).

Dieser Betrachtung liegt die Einsicht zugrunde, dass Kompetenzstrukturmodelle eine ein- oder mehrdimensionale Struktur aufweisen können. Ein eindimensionales Modell beschreibt die Unterschiede in der zu erfassenden Kompetenz auf einem einzelnen Kontinuum. Um differenziertere Aussagen über Kompetenzen und Teilkompetenzen der untersuchten Schüler machen zu können und um komplexere Kompetenzstrukturen abzubilden, können auch mehrdimensionale Item-Response-Modelle verwendet werden. Nach der Identifizierung der Kompetenzstruktur kann mithilfe von *Kompetenzniveaumodellen* erfasst werden, welche fachbezogenen Leistungsanforderungen bzw. welche spezifischen Kompetenzen Schüler bewältigen können und welche nicht.

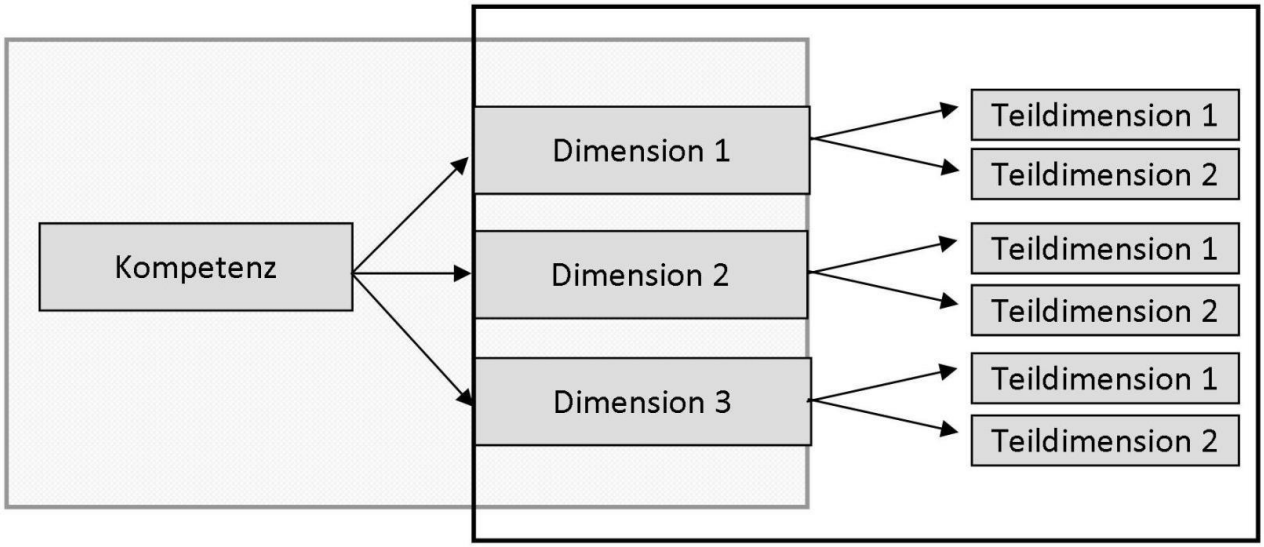

Abb. 1: Schematische Darstellung eines Kompetenzstrukturmodells (Erläuterung s. Text)

Betrachten wir zunächst ein eindimensionales Kompetenzkonstrukt, wobei alle Leistungen auf einem Kontinuum abgebildet werden können, z. B. verschiedene Leistungsniveaus der Wahrnehmungsfähigkeit von Schülern. Dieses Kontinuum wird in Abschnitte unterteilt, welche als *Kompetenzniveaus* oder auch *Kompetenzstufen* bezeichnet werden. In Anlehnung an Hartig und Klieme (2006) soll im Folgenden die Bezeichnung Kompetenzniveaus verwendet werden. Für diese Kompetenzniveaus wird dann eine kriteriumsorientierte Beschreibung der erfassten Kompetenzen durchgeführt, d. h. es wird ein vorab definiertes *Kriterium* der (quantifizierten) Leistungswerte herangezogen. Beispielsweise wäre das einzelne Erkennen aus einem Ensembleklang Niveau II und das gleichzeitige Verfolgen mehrerer Stimmen in einem Musikstück Niveau III (vgl. Abb. 2).

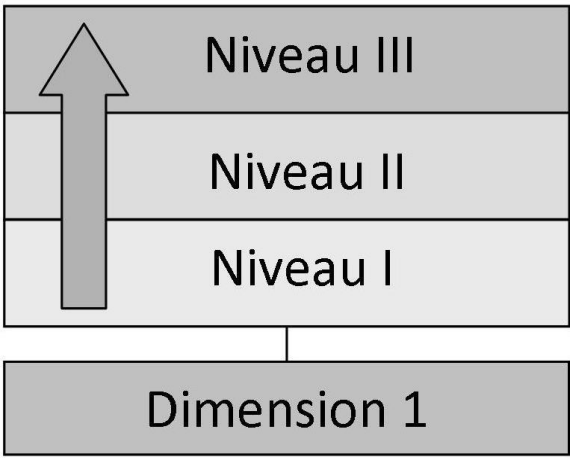

Abb. 2: Schematische Darstellung eines Kompetenzniveaumodells mit einer Dimension

Methoden der Item-Response-Theorie (IRT) (z. B. Rost 2004; Hambleton, Swaminathan 2000) ermöglichen das Bilden einer gemeinsamen Skala, auf der sowohl die Fähigkeiten der Schüler als auch die Schwierigkeiten der Aufgaben dargestellt werden können. Somit können Aussagen über Schüler mit unterschiedlich ausgeprägten Kompetenzen gemacht werden. Gleichzeitig kann festgestellt werden, welche Aufgaben diese Schüler mit einer bestimmten Wahrscheinlichkeit lösen können und welche nicht. Im Gegensatz zur Klassischen Testtheorie besteht bei der Item-Response-Theorie die Möglichkeit, Personenfähigkeit und Itemschwierigkeit auf einer gemeinsamen Skala darzustellen. Es kann gezeigt werden, wie sich die Schwierigkeit einzelner Items über das gesamte Modell verteilt. Man kann daher z. B. sagen: „Der Schüler hat eine Fähigkeit, die es ihm erlaubt, die Aufgaben einer ganz bestimmten Schwierigkeitsstufe mit hoher Wahrscheinlichkeit zu lösen, z. B. solche, bei denen sowohl Violin- als auch Bassschlüssel verwendet wird.“ Interessant wird es, wenn man dann auch unterschiedliche Kompetenzbereiche (z. B. Notenlesen, Melodien merken) auf der gleichen Metrik abbilden kann. Kompetenzniveau- und Kompetenzstrukturmodelle schließen sich keineswegs aus, sondern ergänzen sich idealerweise.

Theoretisch entwickeltes Kompetenzmodell für den Bereich „Musik wahrnehmen und kontextualisieren“

Bevor eine Modellierung der Kompetenzen vorgenommen werden kann, müssen auf Basis fachdidaktischen Wissens sowohl ein theoretisches Kompetenzmodell als auch damit verbundene Testaufgaben erstellt werden. Das im Rahmen von KoMus entwickelte Modell bzw. die darauf basierenden Testaufgaben beziehen sich auf vier Dimensionen: reine Wahrnehmung (1), Wahrnehmung in Verbindung mit Fachterminologie (2), in Verbindung mit Notation (3) und in Verbindung mit der Anwendung von Kontextwissen (4) (s. Abb. 3). Eine ausführliche konzeptionelle und inhaltliche Erläuterung der einzelnen Dimensionen ist bei Niessen et al. 2008 nachzulesen. Nach der Darlegung der theoretischen Grundlagen von Kompetenzmodellierung wenden wir uns dem von uns erhobenen Datensatz zu.

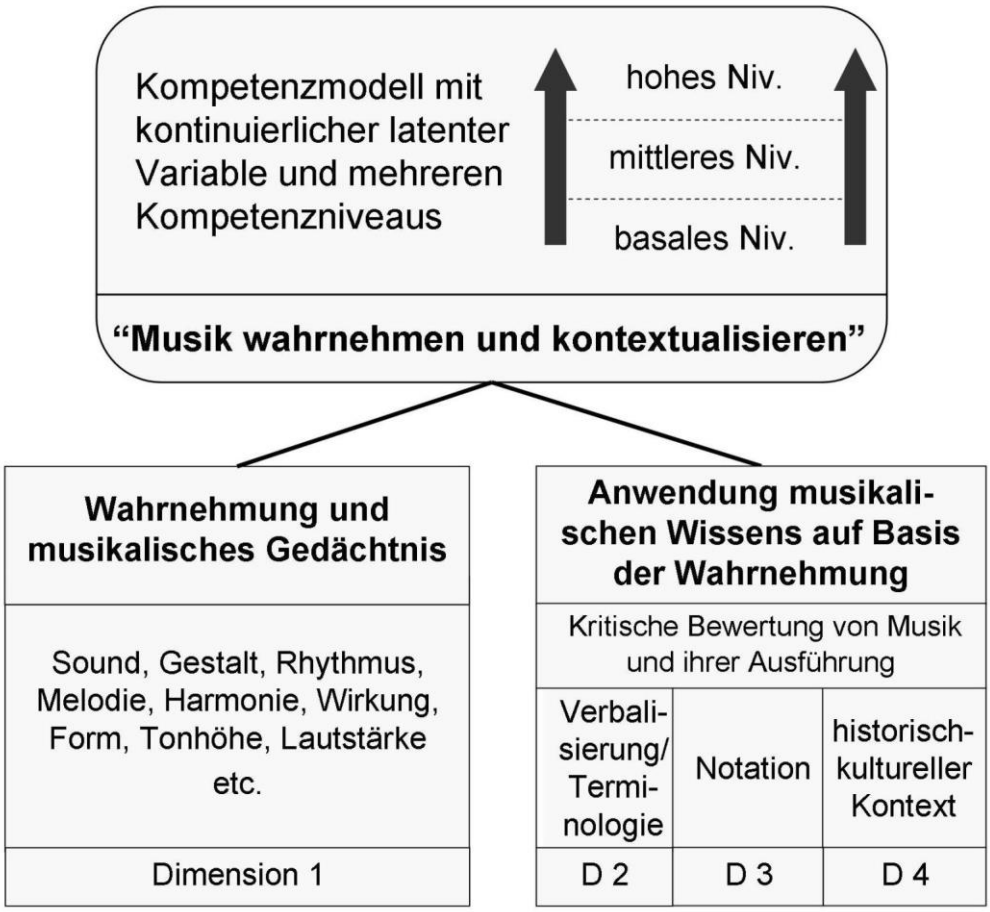

Abb. 3: Kompetenzmodell von Niessen et al. (2008) am Ende der Aufgabenentwicklung[2].

2 Forschungsdesign und Datengrundlage

Um den Forschungsprozess vom theoretisch entwickelten Kompetenzmodell bis zur empirischen Überprüfung nachvollziehbar machen zu können, werden wir zunächst das Forschungsdesign der Pilotierungsstudie vorstellen. Der Prozess der Aufgabenentwicklung und die damit verbundenen statistischen Analysen werden an anderer Stelle thematisiert (Knigge, Niessen & Jordan 2010).

Da im Fach Musik bisher noch keine Kompetenzmodelle entwickelt wurden, ist es in einem ersten Schritt notwendig, die inhaltlich angelegten und

2 Abbildung 3 zeigt eine überarbeitete Fassung des bei Niessen et al. 2008 vorgestellten Modells. In vorliegender Form diente das Modell als Grundlage für die Operationalisierung in Form von Testaufgaben. Die endgültige Formulierung des Modells ist erst nach Abschluss der Analysen der Pilotierungsstudie vorgesehen, sodass das abgebildete Modell weiterhin einen vorläufigen Status besitzt.

vermuteten Dimensionen des Modells empirisch zu validieren. Erst im Anschluss daran kann eine Gesamtmodellierung aller Dimensionen vorgenommen werden. Modellierung heißt in diesem Zusammenhang die Berechnung und Beurteilung der Modellstruktur anhand statistischer Kennwerte.

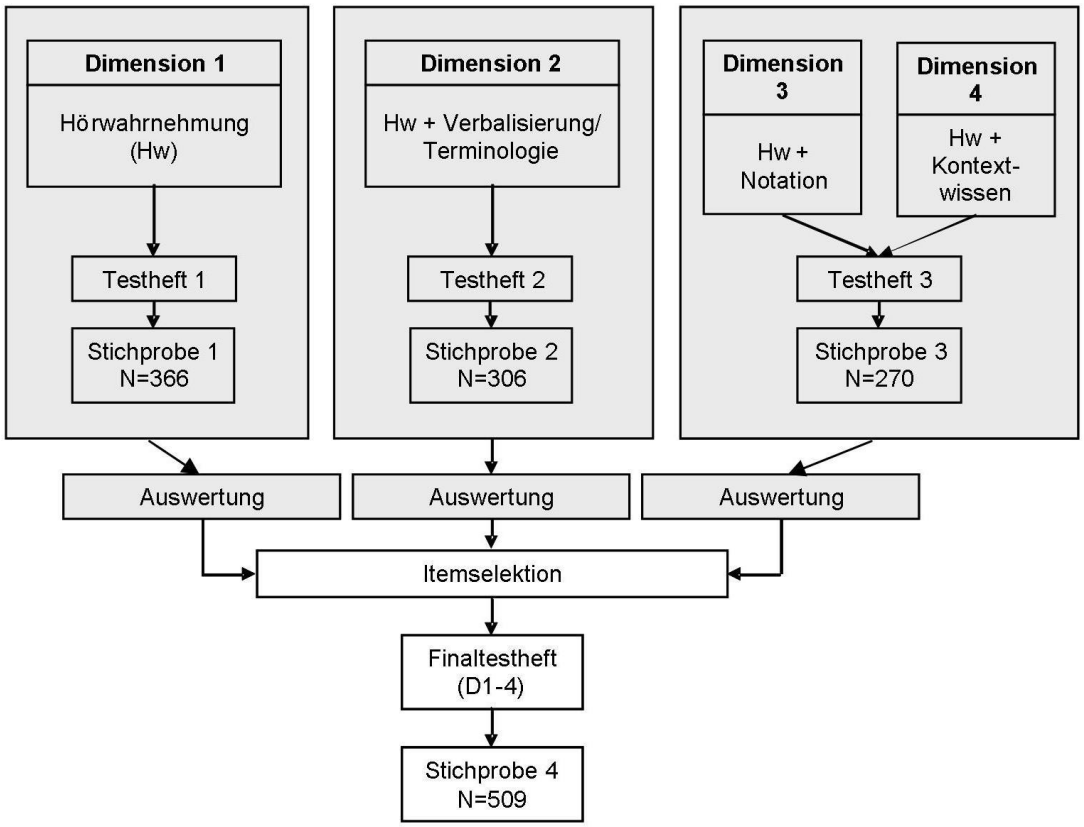

Abb. 4: Testdesign der Pilotierungsstudie

Wie bereits erläutert wurde ein zweistufiges Pilotierungsdesign gewählt. Zunächst wurde pro Dimension ein Testheft mit Aufgaben zusammengestellt und mit einer Schülerstichprobe bestehend aus Schülern von sechsten Klassen in Bremen und Niedersachsen ($N_{(Dim\ 1)}$=366, $N_{(Dim\ 2)}$=306, $N_{(Dim\ 3\text{-}4)}$[3]=270) getestet. Innerhalb des Klassenverbandes bearbeitete jeder Schüler die Testaufgaben an einem eigenen Laptop und konnte die Hörbeispiele individuell mittels Kopfhörer abspielen. Dabei wurde ein so genanntes Ankeritemdesign verwendet. Eine bestimmte Anzahl von gemeinsamen Items, sog. Ankeritems, wurde von allen Schülern bearbeitet. Diese gemeinsamen Items verbinden („verankern“) die Testhefte. Zusätzlich bearbeiteten die Schüler jeweils weite-

3 Aufgrund der geringer Itemanzahl in Dimension 3 und 4 wurden die Aufgaben dieser beiden Dimensionen zusammen in einem Testheft bearbeitet.

re Items. Der Vorteil ist hierbei, dass so eine größere Anzahl von Items einbezogen werden kann, als dies möglich wäre, wenn jede Testperson die gleichen Testitems vorgelegt bekäme. Über die Ankeritems kann später mithilfe der Item-Response-Theorie eine Skalierung für alle Aufgaben vorgenommen werden, auch wenn diese nicht von allen Schülern bearbeitet wurden.

Nach der Auswertung der Testhefte 1 bis 3 wurden Aufgaben für das Finaltestheft ausgewählt. Bei der Auswahl wurde neben der Berücksichtigung statistischer Kriterien (Trennschärfe, Itemschwierigkeit, Modellfit, etc.) auch darauf geachtet, dass die Aufgaben hinsichtlich ihrer Schwierigkeit gut über das gesamte Kontinuum verteilt und alle inhaltlichen Facetten des Modells vertreten waren. Im Finaltestheft wurden also qualitativ hochwertige Items aus allen Dimensionen einbezogen. Da aufgrund mangelnder Testzeit nicht alle Aufgaben der Dimensionen 1 bis 4 in den Finaltest übernommen werden konnten, mussten für jede Dimension je neun Ankeritems für das Finaltestheft ausgewählt werden. Neben diesen 36 Items wurden 71 weitere Items ausgewählt, die aber nicht von allen Schülern bearbeitet wurden (Erläuterung s. o.). Das Finaltestheft wurde mit 508 Schülern verschiedener Schulformen aus Bremen und Niedersachsen getestet und ermöglicht nun die Überprüfung von Zusammenhängen zwischen den Dimensionen. Insgesamt wurden somit 1451 Schüler an 27 Schulen verschiedener Schulformen[4] in die Pilotierung einbezogen.

3 Forschungsleitende Fragestellungen

Mit Hilfe der Daten der Pilotierungsstudie sollen zwei Fragestellungen bearbeitet werden: In einem ersten Schritt wird überprüft, ob das Raschmodell (das einfachste Testmodell der Item-Response-Theorie) gültig ist und auf die Daten angewendet werden kann. Anders ausgedrückt wird untersucht, ob sich die musikalische Kompetenz als latente Variable abbilden lässt. Nach verschiedenen standardmäßigen Modellgültigkeitsüberprüfungen mit Hilfe der Software *ConQuest* kann die Struktur der Daten genauer untersucht werden. Unter der Voraussetzung, dass das Raschmodell gilt, kann mit so genannten Dimensionsanalysen im zweiten Schritt geklärt werden, ob eine eindimensionale oder eine mehrdimensionale Struktur vorliegt. Dabei können verschiedene Modellstrukturen (z. B. zwei- vs. vierdimensionales Modell) einander gegenüberge-

4 12 Gymnasien, 11 Haupt- und Realschulen, 2 Gesamtschulen, 2 Sekundarschulen.

stellt werden und somit das Modell identifiziert werden, das die Daten am besten abbildet.

Nach Aufstellung der Modellstruktur beschäftigen sich die Analysen mit den Zusammenhängen zwischen den gefundenen Dimensionen. Es wird dabei überprüft, ob die Teildimensionen tatsächlich eigenständige Dimensionen bilden oder ob die Korrelationen zwischen den Teildimensionen so hoch sind, dass sie nur in der Zusammenfassung zu einer Dimension sinnvoll interpretierbar sind. Weiterhin wird die interne Konsistenz (Reliabilität) jeder einzelnen Dimension berechnet, die einen Kennwert für ihre Stabilität darstellt.

4 Methoden

Modellierung der Kompetenz „Musik wahrnehmen und kontextualisieren" als latente Variable - Vorgehensweise

Um die Gültigkeit des Raschmodells zu überprüfen, gibt es eine Reihe von Modellgültigkeitstests, die hier nicht im Einzelnen aufgeführt werden können (s. dazu Rost 2004, Bühner 2008, Moosbrugger et al. 2007). Eine zentrale Annahme des Raschmodells bildet die Annahme der Personenhomogenität, die besagt, dass „alle getesteten Personen den Test aufgrund derselben Eigenschaft oder Fähigkeit bearbeiten" (Rost 2004, S. 347). Diese Annahme kann mit dem Likelihood-Quotienten-Test (auch Andersen-Test) überprüft werden (Andersen 1973, Bühner 2008). Dabei werden die Itemparameter für verschiedene Untergruppen geschätzt und auf Unterschiede geprüft. Die Untergruppen können nach verschiedenen Teilungskriterien, z. B. Testscore, Geschlecht oder Schulform, gebildet werden. Im Hinblick auf die musikalische Kompetenz ist es darüber hinaus naheliegend, die Stichprobe nach dem Kriterium der vorhandenen bzw. nicht vorhandenen Instrumentalerfahrung zu teilen. Zur besseren Anschaulichkeit können die Itemparameter der verschiedenen Untergruppen in einem Streudiagramm einander gegenübergestellt werden (s. Abb. 6). Bei Gültigkeit des Raschmodells müssten die Itemparameter eine Gerade mit Steigung 1 bilden, die durch den Nullpunkt verläuft, da es keine großen Abweichungen zwischen den Itemparametern der verschiedenen Gruppen (z. B. Schüler mit/ohne Instrumentalerfahrung) geben dürfte.

Unter der Voraussetzung, dass das Raschmodell gilt, schließen sich weitere Analysen an. Es gilt zu hinterfragen, ob die Daten eine ein- oder mehrdimensionale Struktur aufweisen. Zur Untersuchung des Kompetenzstrukturmodells

(s. Abschnitt 1) kann ein exploratives ebenso wie ein konfirmatorisches Vorgehen genutzt werden, vergleichbar mit der Unterscheidung bei der Faktorenanalyse aus der Klassischen Testtheorie (Bühner 2008). Man nimmt an, dass die Interkorrelationen zwischen den Einzelitems über ein gemeinsames latentes Konstrukt bzw. mehrere Konstrukte erklärt werden können (vgl. Abbildung 5). Der Unterschied zur Faktorenanalyse besteht darin, dass für alle Items die gleiche Ladung vorausgesetzt wird (d. h. alle Items haben die gleiche Trennschärfe). Im einfachsten Fall nimmt man an, dass ein einzelnes latentes Konstrukt – etwa musikalische Kompetenz – die Ergebnisse erklären kann (vgl. Abbildung 5).

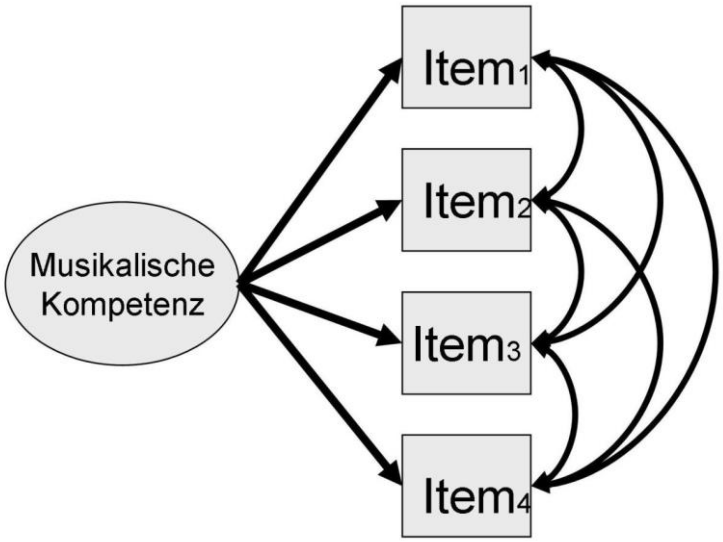

Abb. 5: Darstellung der latenten Variablen „Musikalische Kompetenz“, die hier durch vier untereinander korrelierende Items illustriert wird

Im Folgenden soll zunächst ein theoriegeleitetes, also konfirmatorisches Vorgehen vorgestellt werden. Dabei werden verschiedene theoretische Annahmen einander gegenübergestellt und anhand statistischer Kriterien verglichen, was einen Vergleich zwischen eindimensionalen und z. B. zwei-, vier-, fünf- oder sechsdimensionalen Modellen bedeutet. Die Items werden dazu *a priori* den Dimensionen zugeordnet. In einem weiteren Modell soll getestet werden, ob es Items gibt, die nicht eindeutig einer Dimension zugeordnet werden können, sondern z. B. neben der Wahrnehmungsfähigkeit auch Notations- und Fachterminologiekenntnisse erfordern (*between/within-Modelle*, u. a. Adams, Wilson, Wang 1997; Hartig, Höhler 2008).

Um entscheiden zu können, welches Modell die Struktur der Daten am besten abbildet, werden statistische Modellgütekriterien herangezogen. Allerdings sind Modellprüfungen immer relativ, denn gesucht wird nicht nach ei-

nem allgemeingültigen, überdauernden Modell, sondern nach einer Entscheidung zwischen alternativen Modellen hinsichtlich einer optimalen Anpassung an die Daten. Ein zusammenfassendes Modellgütekriterium bildet der so genannte CAIC-Index[5] (Bozdogan 1987). Dieser berücksichtigt neben der Parameteranzahl und der Final Deviance (Likelihood)[6] auch die Stichprobengröße. Das Modell mit dem kleinsten CAIC-Wert bildet die Daten am besten ab. Im Anschluss an die Identifizierung der Modellstruktur wird zur Überprüfung der Stabilität die Reliabilität der einzelnen Dimensionen bestimmt. Die Reliabilität (hier *EAP/PV Reliabilität*) gibt an, inwieweit eine Gruppe von Test-Items als Messung einer latenten Variablen betrachtet werden kann.

Zusammenhänge zwischen den Dimensionen

Nach Aufstellung der Modellstruktur sollen die Zusammenhänge zwischen den gefundenen Dimensionen genauer beleuchtet werden. Korrelationsanalysen sollen zeigen, ob die Zusammenhänge zwischen den Dimensionen so hoch sind, dass sie eine eindimensionale Modellstruktur nahe legen, oder ob die Dimensionen nur gering zusammenhängen und daher einzeln in einem mehrdimensionalen Modell angemessener abgebildet werden können. Im Falle einer Eindimensionalität wäre die Gesamtkompetenz auf einem singulären Kontinuum von niedriger bis hoher Kompetenz angeordnet. Bei einem mehrdimensionalen Modell würden die einzelnen Kompetenzdimensionen durch jeweils eine separate latente Dimension modelliert.

5 Ergebnisse

Modellierung der Kompetenz „Musik wahrnehmen und kontextualisieren“: Überprüfung der Modellgültigkeit und der Modellstruktur

Um die Modellgültigkeit mithilfe des Likelihood-Quotienten-Tests (Andersen-Test) überprüfen zu können, wurde die Stichprobe in zwei Untergruppen geteilt. Als Teilungskriterium wurde hier beispielhaft die Variable „Spielst du zur Zeit oder hast du früher ein Instrument gespielt?“ eingesetzt. Es wurde untersucht, ob die Abweichungen in den Itemantworten von Schülern mit und

5 Consistent Akaike's Information Criterion. CAIC = -2 * ln Likelihood + (ln Stichprobengröße + 1) * Anzahl Parameter.

6 Die Deviance ist ein Wert welcher angibt, wie gut das Modell die Daten abbildet.

ohne Instrumentalerfahrung so stark voneinander abweichen, dass die Modellgültigkeit in Frage gestellt werden muss. Zur Überprüfung dieser und weiterer Fragestellungen wurde die Software *ConQuest* (Wu et al. 2007) eingesetzt.

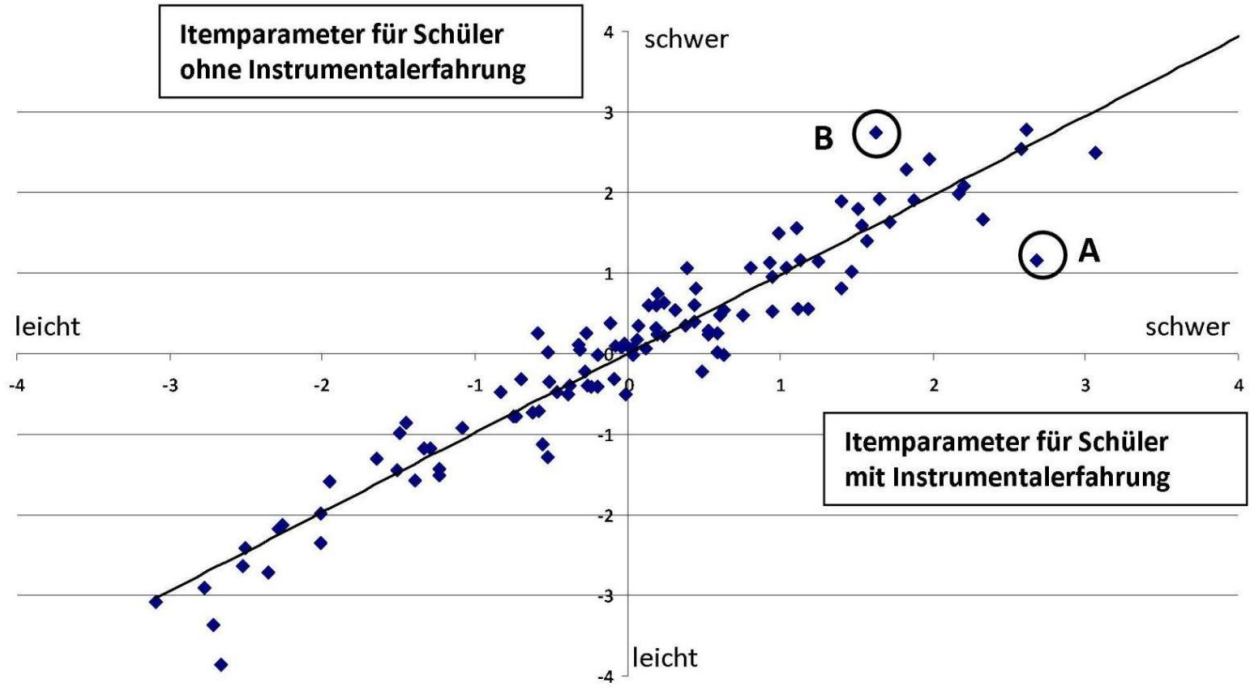

Abb. 6: Streudiagramm der Itemparameter gruppiert nach Schülern mit und ohne Instrumentalerfahrungen.

Legende

A: Dieses Item ist für Schüler mit Instrumentalerfahrung schwerer (2,68) als für Schüler ohne Instrumentalerfahrung (1,16).

B: Dieses Item ist für Schüler mit Instrumentalerfahrung leichter (1,63) als für Schüler ohne Instrumentalerfahrung (2,74).

Abbildung 6 verdeutlicht, dass die Itemparameter einzelner Items nicht auf der Geraden liegen und es somit Unterschiede zwischen den Gruppen gibt. Manche Items sind für Schüler mit Instrumentalerfahrung einfacher, was nicht sehr überraschend ist (s. Erklärung Abb. 6). Es muss aber überprüft werden, ob diese Unterschiede zu einer statistischen Verschlechterung des Modells führen. Tatsächlich mussten einige Items aus dem Finaltestheft für weitere Berechnungen ausgeschlossen werden. Außerdem mussten einige Items aufgrund des Teilungskriteriums „Geschlecht“ aus den weiteren Berechnungen gelöscht

werden. Nach Entfernung der problematischen Items kann von einer Modellgültigkeit ausgegangen werden.

Im Folgenden soll die Modellstruktur genauer betrachtet werden. Es wird überprüft, ob alle Items auf *einer* latenten Variable abgebildet werden können (vgl. Abbildung 5), oder ob sich die Items auf mehrere Dimensionen verteilen. Nach theoretischen Überlegungen werden, in Anlehnung an das erstellte Kompetenzmodell (s. Abb. 3), verschiedene Modellstrukturen gegenübergestellt. Bei dem 2-dimensionalen Modell werden der ersten Dimension alle Items zugeordnet, die den Bereich „Wahrnehmung und musikalisches Gedächtnis“ umfassen (vgl. Dimension 1 in Abbildung 3). Zu der zweiten Dimension gehören dann alle restlichen Items des Bereichs „Anwendung musikalischen Wissens auf Basis der Wahrnehmung“. Eine ähnliche Struktur suggeriert das abgebildete Kompetenzmodell (s. Abb. 3). Die 4-dimensionale Struktur ergibt sich aus den vier intendierten Dimensionen des Kompetenzmodells. Die kritische Bewertung von Musik ist in dem bisherigen Kompetenzmodell noch keiner Dimension zugeordnet. Demnach ist es nahe liegend, die Items, die sich mit der kritischen Bewertung von Musik beschäftigen, einer eigenen und somit der fünften Dimension zuzuordnen. Wie bereits beschrieben besteht darüber hinaus die Möglichkeit, Items mehreren Dimension gleichzeitig zuzuordnen (Within-Item-Modelle). Jedes Item wird somit daraufhin überprüft, ob es mehreren Dimensionen zugeordnet werden kann. Die Basis bildet das 4-dimensionale Modell, ein Item kann somit maximal vier Dimensionen gleichzeitig zugeordnet werden.

Mit der Software *ConQuest* wurden die Modellparameter (z. B. Final Deviance, Anzahl der Parameter) für die verschiedenen mehrdimensionalen Modelle berechnet. Aus der Stichprobengröße, der Anzahl der Parameter und der Likelihood (=Final Deviance) ergibt sich der CAIC-Index (Erläuterung der Formel s. o.). Somit ist das 4-dimensionale Modell im Vergleich mit dem 1-, 2-, 5-dimensionalen bzw. dem Within-Item-Modell das Modell mit dem kleinsten CAIC-Wert und erklärt die Daten am besten. Der Bereich der kritischen Bewertung von Musik bildet also keine eigene Dimension und muss noch einmal genauer betrachtet werden.

	1-dimensionales Modell	2-dimensionales Modell	4-dimensionales Modell	5-dimensionales Modell	4-dimensionales Within-Item-Modell
N	508	508	508	508	508
Anzahl Parameter	102	104	111	116	111
Final Deviance	32229,20	32183,20	32044,76	32027,78	32126,01
CAIC-Index[7]	32966,71	32935,17	32847,34	32866,52	32928,60

Tabelle 1: Modellvergleich zwischen einem 1-, 2-, 4- und 5-dimensionalen sowie einem Within-Item Modell.

Betrachtet man die Stabilität der vier Dimensionen, können folgende Reliabilitäten berichtet werden (Erwartungswert 1): $\alpha_{\text{Dimension 1}} = .81$, $\alpha_{\text{Dimension 2}} = .82$, $\alpha_{\text{Dimension 3}} = .78$ und $\alpha_{\text{Dimension 4}} = .70$. Anschließend wurden die Zusammenhänge zwischen den vier Dimensionen des Modells untersucht, um erneut die Sinnhaftigkeit eines 4-dimensionalen Modells zu überprüfen.

Zusammenhänge zwischen den Dimensionen

Wie in Tabelle 2 erkennbar, ergab sich der höchste Zusammenhang zwischen dem Bereich „Wahrnehmung und musikalisches Gedächtnis“ und der „Verbalisierung auf Basis der Wahrnehmung“. Auch zwischen der Dimension „Verbalisierung/Fachterminologie“ und „Notation“ (jeweils auf Basis der Wahrnehmung) zeigte sich ein hoher Zusammenhang, d. h. Schüler, die eine hohe Fähigkeit im Bereich der „Verbalisierung/Fachterminologie“ bewiesen, zeigten ebenfalls eine hohe Fähigkeit im Bereich „Notation“. Der geringste Zusammenhang bestand zwischen der Dimension „Notation“ und der Dimension „Historisch-kultureller Kontext“. Trotz einiger relativ hohen Korrelationen (s.

7 Das Modell mit dem kleinsten CAIC-Index bildet die Daten am besten ab.

Tab. 2), besonders zwischen Dimension 1 und 2, wurde inhaltlich und aufgrund des CAIC-Index die 4-dimensionale Struktur favorisiert.

	Dimension 1 (Reine Wahrnehmung)	Dimension 2 (Verbalisierung/ Fachterminologie in Verbindung mit Wahrnehmung)	Dimension 3 (Notation in Verbindung mit Wahrnehmung)	Dimension 4 (Historischer und kultureller Kontext in Verbindung mit Wahrnehmung)
Dimension 1	-			
Dimension 2	**0.83**	-		
Dimension 3	**0.80**	**0.84**	-	
Dimension 4	**0.71**	**0.67**	**0.56**	-

Tabelle 2: Latente messfehlerbereinigte Korrelationen zwischen den vier Dimensionen

6 Diskussion

Die Überprüfung der Modellgültigkeit ergab, dass nicht alle Items in das Raschmodell passten, d.h. dass sie die Anforderungen nicht erfüllten. Damit wird die Wichtigkeit des zweistufigen Pilotierungsdesigns (s. Abb. 4) sowie der sorgfältigen Analyse und Entwicklung der Items bestätigt.

Hinsichtlich der Modellstruktur konnte eine 4-dimensionale Struktur nachgewiesen werden, wobei die einzelnen Dimensionen hinreichend hohe Reliabilitäten besaßen. Die interne Konsistenz der vier Dimensionen war damit belegt. Lediglich die Reliabilität der vierten Dimensionen fällt mit $\alpha = .70$ vergleichsweise niedrig aus. Aufgaben dieser Dimension können somit nicht so deutlich trennscharf zwischen Schülern mit hoher und niedriger Fähigkeit unterscheiden.

Bei einem mehrdimensionalen Modell muss geklärt werden, wie hoch die einzelnen Dimensionen zusammenhängen, damit tatsächlich eine Eindimensionalität ausgeschlossen werden kann. Tabelle 2 zeigt, dass die höchste Korre-

lation (r= 0.84) zwischen den Dimensionen 2 (Verbalisierung/Fachterminologie) und Dimension 3 (Notation) zu finden ist. Deshalb ist es sehr wahrscheinlich, dass ein Schüler mit einer hohen Fähigkeit in Dimension 2 ebenfalls über eine hohe Fähigkeit in Dimension 3 verfügt. Zwischen den Dimensionen 3 und 4 zeigt sich die geringste Korrelation (r= 0.56), was aufgrund der völlig unterschiedlichen inhaltlichen Ausrichtung auch plausibel erscheint: Während es in Dimension 3 darum ging, Aufgaben im Bereich Notation auf Basis der Wahrnehmung zu lösen (z. B. ein gehörtes Musikstück einem Ausschnitt einer Partitur zuordnen), sollten die Schüler bei Aufgaben der Dimension 4 historische und kulturelle Aspekte der Musik berücksichtigen (z. B. ein gehörtes Musikstück dem Herkunftsland zuordnen und erläutern, warum diese Entscheidung getroffen wurde). Ein Vergleich unserer Ergebnisse mit denen anderer Studien (z. B. Bos, Lankes 2003) bestätigt die Entscheidung für Mehrdimensionalität. Bei einer Korrelation ab ca. r= 0.90 zwischen zwei Dimensionen müsste hingegen geprüft werden, ob es sich wirklich um ein mehrdimensionales Konstrukt handelt oder ob die beiden Dimension nicht vielmehr eine einzige latente Variable abbilden (vgl. Wu, Adams 2006).

Ein weiterer Vorteil des mehrdimensionalen gegenüber einem eindimensionalen Modell ist die Möglichkeit differenzierter Rückmeldungen der Ergebnisse an die Lehrkräfte. So kann eine Klasse beispielsweise im Bereich der Verbindung von Wahrnehmung mit Notation über sehr hohe und im Bereich der Verbindung von Wahrnehmung mit Verbalisierung/Fachterminologie hingegen nur über sehr geringe Fähigkeiten verfügen. Diese differenzierte Rückmeldung können die Lehrkräfte in ihre weitere Unterrichtsplanung einfließen lassen. Das eindimensionale Modell würde nur die Rückmeldung einer Gesamtkompetenz „Musik wahrnehmen und kontextualisieren“ mit entsprechend weniger diagnostischen und didaktischen Perspektiven ermöglichen.

7 Ausblick

In diesem Beitrag wurde die Überprüfung der Testgütekriterien nur am Rande erwähnt (z. B. Moosbrugger 2007, S. 8 ff.). Noch existieren z. B. keine standardisierten Tests in diesem Bereich, anhand derer die Validität des KoMus-Tests gemessen werden könnte. Der KoMus-Test wurde aber auf Basis eines theoretischen Modells erstellt, welches wiederum auf der Grundlage umfangreicher Curriculaanalysen (s. Knigge, Lehmann-Wermser 2008) entstanden ist. Dies stellt bereits ein Validitätskriterium dar, nämlich das der „Inhalts-“ bzw. „curricularen Validität“. Um die curriculare Validität möglichst objektiv abzu-

sichern, haben die Lehrkräfte der teilnehmenden Klassen im Rahmen der Pilotierungsstudie die Testaufgaben hinsichtlich Schwierigkeit, Bekanntheitsgrad, Relevanz und Lehrplanpassung beurteilt.

Wie eingangs beschrieben, gibt es neben dem vorgestellten Kompetenzstrukturmodell ein so genanntes Kompetenzniveaumodell. Hier werden für jede Dimension Kompetenzniveaus bestimmt. Dabei ist es besonders wichtig, wo die Übergänge von einem zum nächsten Niveau festgelegt werden; also ab welcher Fähigkeit ein Schüler das nächsthöhere Niveau erreicht hat. Die festgelegten Niveaus werden anschließend mit inhaltlichen Kriterien beschrieben, sodass deutlich wird, über welche Kompetenzen Schüler auf einem bestimmten Niveau verfügen, nicht aber auf dem nächst niedrigeren (vgl. Klieme et al. 2003, S. 24, Jordan (i.Vorb.)).

Bereits in den Modellgültigkeitsüberprüfungen (s. Abschnitt 5) wurde die Relevanz der Instrumentalerfahrung der Schüler deutlich. In weiteren Analysen sollte der Einfluss dieser und weiterer Hintergrundvariablen auf die Kompetenz im Bereich „Musik wahrnehmen und kontextualisieren“ differenziert untersucht werden. Neben den Instrumentalerfahrungen der Schüler könnten u. a. das Musikinteresse der Familie, die Anzahl der Jahre an Musikunterricht in der Schule und allgemeine Hintergrundvariablen wie der sozio-ökonomische Hintergrund oder das Geschlecht einbezogen werden.

Literatur

Adams, R. J.; Wilson, M.; Wang, W.-C. (1997): The Multidimensional Random Coefficients Multinomial Logit Model. In: Applied Psychological Measurement, Jg. 21, H. 1, S. 1-23.

Andersen, E. B. (1973): A goodness of fit test for the Rasch model. In: Psychometrika, Jg. 38, S. 123-140.

Bähr, J. (2001): Zur Entwicklung musikalischer Fähigkeiten von Zehn- bis Zwölfjährigen. Evaluation eines Modellversuchs zur Kooperation von Schule und Musikschule. Göttingen: Cuvillier.

Bähr, J. (2004): Bildungsstandards für den Musikunterricht? In: Ansohn, Meinhard; Terhag, Jürgen (Hg.): Musikkulturen - fremd und vertraut. Oldershausen: Lugert (Musikunterricht heute), S. 404-419.

Bos, W.; Lankes, E.-M. (2003): Erste Ergebnisse aus IGLU. Schülerleistungen am Ende der vierten Jahrgangsstufe im internationalen Vergleich. Münster: Waxmann.

Bozdogan, H. (1987): Model selection an Akaikes information criterion (AIC): The general theory and its analytical extension. In: Psychometrika, Jg. 52, H. 3, S. 345-370.

Bruhn, H. (2005): Wissen und Gedächtnis. In: Oerter, R./Stoffer, T. H. (Hrsg.): Allgemeine Musikpsychologie. Göttingen: Hogrefe (Enzyklopädie der Psychologie, Serie VII, Bd. 1), S. 537-590.

Bühner, M. (2008): Einführung in die Test- und Fragebogenkonstruktion. 2., aktualisierte und erw. Aufl., [Nachdr.]. München: Pearson Studium (psMethoden/Diagnostik).

Hambleton, R. K.; Swaminathan, H. (2000): Item response theory. Principles and applications. 10. printing. Boston: Kluwer-Nijhoff (Evaluation in education and human services series).

Hartig, J.; Höhler, J. (2008): Representation of Competencies in Multidimensional IRT Models with Within-Item and Between-Item Multidimensionality. In: Zeitschrift für Psychologie, Jg. 216, H. 2, S. 89-101. Online verfügbar unter http://psycontent.metapress.com/content/w264233j2t034874/fulltext.pdf, zuletzt geprüft am 10.09.2008.

Hartig, J.; Klieme, E. (2006): Kompetenz und Kompetenzdiagnostik. In: Schweizer, K. (Hg.): Leistung und Leistungsdiagnostik. Mit 18 Tabellen. Heidelberg: Springer Medizin , S. 127-143.

Hartig, J.; Klieme, E. (Hg.) (2007): Möglichkeiten und Voraussetzungen technologiebasierter Kompetenzdiagnostik. Eine Expertise im Auftrag des Bundesministeriums für Bildung und Forschung. Berlin: BMBF (Bildungsforschung, 20).

Jank, W. (2001): Ist Musiklernen wie Sprechenlernen? Musikalische Grundkompetenzen: Die Musikdidaktik muss von der Lerntheorie lernen. In: Musik & Bildung, H. 3, S. 31-39.

Jordan, A.-K.; Knigge, J. (i. Druck): The development of competency models: An IRT-based approach to competency assessment in general music education. In: Brophy, T. S. (Hg.): The Practice of Assessment in Music Education: Frameworks, Models, and Designs. Chicago: GIA .

Jordan, A.-K. (i. Vorb.): Empirische Validierung eine Kompetenzmodells für das Fach Musik – Teilkompetenz „Wahrnehmen und Kontextualisieren von Musik“ (Arbeitstitel).

Kaiser, H. J. (2006): Bildungsoffensive Musikunterricht? Das Grundsatzpapier der Konrad-Adenauer-Stiftung in der Diskussion. Regensburg: ConBrio.

Klieme, E.; Avenarius, H.; Blum, W.; Döbrich, P.; Gruber, H.; Prenzel, M. et al. (2003): Zur Entwicklung nationaler Bildungsstandards. Eine Expertise. Herausgegeben von Bundesministerium für Bildung und Forschung. (Bildungsforschung, 1). Online www.bmbf.de/pub/zur_entwicklung_nationaler_bildungsstandards.pdf, zuletzt geprüft am 10.07.2008.

Knigge, J.; Lehmann-Wermser, A. (2008): Bildungsstandards für das Fach Musik: Eine Zwischenbilanz. In: Zeitschrift für Kritische Musikpädagogik, Sonderedition: Bildungsstandards und Kompetenzmodelle für das Fach Musik?, S. 60–98. Online verfügbar unter www.zfkm.org/sonder08-knigge-lehmannwermser.pdf, zuletzt geprüft am 27.08.2009.

Knigge, J.; Lehmann-Wermser, A. (2009): Kompetenzorientierung im Musikunterricht. In: Musik und Unterricht, H. 94, S. 56–60.

Knigge, J.; Niessen, A.; Jordan, A.-K. (2010): Erfassung der Kompetenz „Musik wahrnehmen und kontextualisieren“ mit Hilfe von Testaufgaben Aufgabenentwicklung und -analyse im Projekt KoMus In: Knolle, N. (Hg.): Evaluationsforschung in der Musikpädagogik. Essen: Die Blaue Eule (Musikpädagogische Forschung, 31).

La Motte-Haber, H. de (2005): Modelle der musikalischen Wahrnehmung. Psychophysik - Gestalt - Invarianten - Mustererkennen - Neuronale Netze - Sprachmetapher. In: La Motte-Haber, H. de; Rötter, G. (Hg.): Musikpsychologie. Laaber: Laaber (Handbuch der Systematischen Musikwissenschaft, 3), S. 55–73.

Lange, E. B. (2005): Musikpsychologische Forschung im Kontext allgemeinpsychologischer Gedächtnismodelle. In: La Motte-Haber, H. de/Rötter, G. (Hrsg.): Musikpsychologie. Laaber: Laaber (Handbuch der Systematischen Musikwissenschaft), S. 74–100.

Lohmann, W. (1982): Ansätze zu einer objektiven Bewertung von Leistungen im Musikunterricht. Wolfenbüttel: Möseler.

Moosbrugger, H.; Kelava, A. (Hg.) (2007): Testtheorie und Fragebogenkonstruktion. Mit 43 Tabellen. Heidelberg: Springer Medizin (Springer-Lehrbuch).

Niessen, A.; Lehmann-Wermser, A.; Knigge, J.; Lehmann, A. C. (2008): Entwurf eines Kompetenzmodells 'Musik wahrnehmen und kontextualisieren'. In: Zeitschrift für Kritische Musikpädagogik, Sonderedition: Bildungsstandards und Kompetenzmodelle für das Fach Musik?, S. 3–33. Online verfügbar unter http://www.zfkm.org/sonder08-niessenetal.pdf, zuletzt geprüft am 20.03.2010.

Richter, C. (2007): Bildungsstandards und Kompetenzformulierungen im Fach Musik. In: Labudde, Peter (Hg.): Bildungsstandards am Gymnasium. Korsett oder Katalysator? Bern: h.e.p. , S. 273–281.

Rost, J. (2004): Lehrbuch Testtheorie - Testkonstruktion. 2., vollst. überarb. und erw. Aufl. Bern: Huber (Psychologie Lehrbuch).

Serafine, M. L. (1988): Music as cognition. The development of thought in sound. New York: Columbia University Press.

Stoffer, T. H. (2005): Aufmerksamkeitsprozesse beim Musikhören: Wissensunabhängige und wissensabhängige Selektionsprozesse. In: Oerter, R./Stoffer, T. H. (Hrsg.): Allgemeine Musikpsychologie. Göttingen: Hogrefe (Enzyklopädie der Psychologie, Serie VII, Bd. 1), S. 591–656.

van der Linden, W. J.; Hambleton, R. K. (1997): Handbook of modern item response theory. New York, NY: Springer.

Weinert, Franz E. (Hg.) (2001): Leistungsmessungen in Schulen. Weinheim: Beltz.

Wu, M.; Adams, R. J. (2006): Modelling Mathematics Problem Solving Item Responses Using a Multidimensional IRT Model. In: Mathematics Education Research Journal, Jg. 18, H. 2.

Wu, M.; Adams, R. J.; Wilson, M.; Haldane, S. A. (2007): ConQuest - Generalised item response modelling software. Camberwell: Australian Council for Educational Research.