Knigge, Jens; Niessen, Anne; Jordan, Anne-Katrin

Erfassung der Kompetenz "Musik wahrnehmen und kontextualisieren" mit Hilfe von Testaufgaben. Aufgabenentwicklung und -analyse im Projekt KoMus

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. 81-107. - (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

JENS KNIGGE, ANNE NIESSEN, ANNE-KATRIN JORDAN

Erfassung der Kompetenz „Musik wahrnehmen und kontextualisieren“ mit Hilfe von Testaufgaben

Aufgabenentwicklung und -analyse im Projekt KoMus

“Historically, task design has been regarded more as an art than a science.”
(Mislevy, Steinberg & Almond 2002, S. 98)

1 Einleitung

‚Kompetenzorientierung’ ist zum pädagogischen und bildungspolitischen Modebegriff avanciert: Schulen entwickeln Programme zur Förderung von Methoden- und Sozialkompetenz; nationale Bildungsstandards reagieren auf die *PISA*-Misere mit der Formulierung von zu erreichenden Kompetenzen, und in der psychologischen und allgemein-pädagogischen Literatur ist die Anzahl der Veröffentlichungen zum Thema Kompetenzen in den letzten 10 Jahren ‚explodiert’ (Klieme & Hartig 2007, S. 13). Im Fach Musik wird die Kompetenzorientierung auf curricularer Ebene breit – oft jedoch leider auch unreflektiert – implementiert (Knigge & Lehmann-Wermser 2008). Der fachdidaktische Diskurs gestaltet sich allerdings sehr heterogen: Partielle Zustimmung erhält das Kompetenzkonzept bspw. im Kontext des „aufbauenden Musikunterrichts“ (Jank 2007), bei anderen Autoren[1] trifft es auf Skepsis bis hin zu kategorischer Ablehnung (z. B. Richter 2008, 2009). Die Hintergründe für die unterschiedlichen Reaktionen zu beleuchten, würde eine eigene Publikation erfordern, aber ein Teil der konträren Einschätzungen hat sicherlich mit der Unsicherheit darüber zu tun, was der Kompetenzbegriff eigentlich bezeichnet: Die Kritiker finden ihn zu kognitionslastig und fürchten angesichts des Booms der Kompetenzmessungen eine Abwendung vom klassischen Bildungsbegriff; die Befürworter betonen gerade die Vieldimensionalität des Kompetenzkonzepts und meinen, dass es zur Steigerung von Unterrichtsqualität und durchaus

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

auch zur Bildung beitragen kann, wenn Schüler ihre Kompetenzen entwickeln und ausbauen (Klieme & Hartig 2007,3 S. 22). Die Kompetenzdiskussion wird jedenfalls mit Leidenschaft geführt – auch in der Musikpädagogik (Knigge & Lehmann-Wermser 2008; Niessen 2009).

Die Unterschiedlichkeit der Einschätzung hängt sicherlich mit der Geschichte und den Verwendungszusammenhängen des Kompetenzbegriffs zusammen, der in Sprachwissenschaft, Psychologie und Erziehungswissenschaften jeweils sehr unterschiedliche Bestimmungen erfahren hat (für eine ausführliche Diskussion siehe z. B. Klieme & Hartig 2007). Klieme & Hartig (2007) fassen „zentrale Bestandteile des Begriffsverständnisses" folgendermaßen zusammen: „Kompetenzen sind Dispositionen, die im Verlauf von Bildungs- und Erziehungsprozessen erworben (erlernt) werden und die Bewältigung von unterschiedlichen Aufgaben bzw. Lebenssituationen ermöglichen. Sie umfassen Wissen und kognitive Fähigkeiten, Komponenten der Selbstregulation und sozial-kommunikative Fähigkeiten wie auch motivationale Orientierungen. Pädagogisches Ziel der Vermittlung von Kompetenzen ist die Befähigung zu selbstständigem und selbstverantwortlichem Handeln und damit zur Mündigkeit" (S. 21). Die Autoren betonen, dass ein solches Begriffsverständnis sehr wohl mit der bekannten Definition von Weinert konform geht, die u. a. das Kompetenzverständnis im Rahmen von Bildungsstandards bestimmt (Weinert 2001; vgl. Klieme et al. 2003). Pädagogisch geprägt ist die Beschreibung von Klieme & Hartig (2007, S. 21) u. a. durch die Betonung der Erlernbarkeit von Kompetenzen und die Zielbestimmung, die das Kompetenzkonzept als nicht nur deskriptives ausweist.

Ein solch facettenreiches und pädagogisch akzentuiertes Verständnis des Begriffs wurde auch der Forschung im Projekt *KoMus* zugrunde gelegt, aus dem im Folgenden einige Ergebnisse der Aufgabenentwicklung und -analyse dargestellt werden. Im Rahmen dieses von der DFG geförderten Projekts[2] wurden Kompetenzen von Schülern der 6. Klasse im Bereich ‚Wahrnehmen und Kontextualisieren von Musik' erforscht. Ausgangspunkt der Überlegun-

2 Das Projekt wurde 2007 bis 2009 an der Universität Bremen durchgeführt. Den Antrag verfassten Andreas Lehmann-Wermser (Universität Bremen), Andreas C. Lehmann (Hochschule für Musik Würzburg) und Anne Niessen (Hochschule für Musik und Tanz Köln), die Mitarbeiter waren Jens Knigge und Anne-Katrin Jordan (beide Universität Bremen). Weitere Informationen zum Projekt sind unter der Adresse http://www.musik.uni-bremen.de/forschung/komus.html erhältlich.

gen war die Frage, über welche Kompetenzen Schüler in Bezug auf ihre Hörwahrnehmung von Musik verfügen und wie sie sie aufbauen. Dabei erschien es sinnvoll, von einer Stufung der entsprechenden Kompetenzen auszugehen, weil Wahrnehmungsprozesse sich beispielsweise im Grad der Differenzierung, im Grad der Vernetzung, der Bewusstheit und der Verknüpfung mit Wissensbeständen unterscheiden. Diese Unterschiede beim ‚Wahrnehmen und Kontextualisieren von Musik’ wurden in einem Kompetenzmodell abgebildet, das auf curricularen Analysen, theoretischen Überlegungen sowie empirischen Forschungsergebnissen basiert (Niessen, Lehmann-Wermser, Knigge & Lehmann 2008; Knigge 2010).

Daran anknüpfend wurden Testaufgaben entwickelt, die zunächst in mehreren Feldtests und nach ihrer Revision im Rahmen einer systematischen Pilotierungsstudie validiert wurden. Die endgültige Auswertung der Pilotierung

Abb. 1: Theoretisches *KoMus*-Kompetenzmodell

steht noch aus;[3] sie wird das in Abbildung 1 dargestellte Modell voraussichtlich noch verändern.[4]

Wie funktioniert aber die Messung von Kompetenzen und die Formulierung von Kompetenzniveaus? Gemäß der Weinertschen Definition, die auch im Rahmen von *KoMus* Verwendung fand, sind Kompetenzen „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). Wenn Problemlösefähigkeiten und -fertigkeiten gemessen werden sollen, muss also das Beobachten von Handeln Rückschlüsse auf die Kompetenzen zulassen und aus diesen Beobachtungen wiederum kann man Testinhalte ableiten (Klieme & Hartig 2007, S. 24).[5] Kompetenzen sind zudem Dispositionen, die sich nicht auf der Grundlage einzelner Beobachtungen bestimmen lassen. Nötig sind viele Beobachtungsergebnisse bei variierenden Aufgaben und Situationen. „Die konsistente Zusammenfassung solcher Einzelbeobachtungen zu einer Aussage über das individuelle Kompetenzniveau ist das, was in der psychometrischen Fachsprache

---

3 Erste Ergebnisse sind bei Jordan, Lehmann & Knigge (2010) dargestellt.

4 Das ursprünglich von Niessen et al. (2008) vorgestellte Modell hat bereits im Zuge der Formulierung des Testkonstrukts (vgl. Abschnitt 2.1) und der daran anschließenden Aufgabenentwicklung verschiedene Modifikationen erfahren. Auf Basis der Ergebnisse der Pilotierungsstudie sind weitere Veränderungen und Präzisierungen, vor allem hinsichtlich der Niveaustruktur zu erwarten.

5 Der Begriff des ‚Problemlösens’ wird z. T. fälschlicherweise missverstanden als das Lösen von lebensweltlich relevanten Problemen. Der Begriff wird von Weinert jedoch in seiner psychologischen Konnotation verwendet, der sich zunächst nicht auf einen speziellen Kontext bezieht. Verwirrend mag in diesem Zusammenhang auch die Vermischung von Kompetenzbegriff und Literacy-Konzept sein (z. B. in den PISA-Studien), welches sich explizit auf die Lösung von (berufs-)alltagsrelevanten Problemen bezieht. Da ästhetisch geprägte Schulfächer nicht auf im Alltag im engeren Sinne notwendige Anforderungen vorbereiten, erscheint eine Verknüpfung von Kompetenz- und Literacy-Konzept in Bezug auf das Fach Musik nicht sinnvoll. Wir schließen uns daher den Kollegen der Mathematikdidaktik an, die für didaktische Zusammenhänge das Begriffsverständnis von ‚Problemlösen’ als „Anforderungen bewältigen“ (Büchter & Leuders 2005, S. 188) vorschlagen (vgl. auch die Ausführungen zum Problemlösen im Musikunterricht von Cvetko & Meyer 2009; Niessen 2008).

als Messung bezeichnet wird. Kompetenzmessung hat also nicht, wie es in manchen erziehungswissenschaftlichen Kommentaren immer noch scheint, mit einer ‚Normierung’ von Gedanken zu tun oder mit einem bloßen Abzählen von Richtigantworten. Messungen können in durchaus komplexen Aussagen resultieren“ (Klieme & Hartig 2007, S. 24).

Ziel des *KoMus*-Projekts war eine möglichst differenzierte Modellierung der Wahrnehmung und Kontextualisierung von Musik. Zur Entwicklung und Validierung eines Kompetenzmodells und eines darauf bezogenen Testinstruments wurde ein Drei-Phasen-Design gewählt: (1) Erstellung eines theoretischen Kompetenzmodells (Niessen et al. 2008), (2) Operationalisierung des Modells in Form von Testaufgaben (Knigge 2010), (3) Empirische Validierung des Modells im Rahmen einer systematischen Pilotierungsstudie (Jordan, Lehmann & Knigge 2010). In vorliegendem Beitrag steht die Phase der Operationalisierung des Modells im Vordergrund. Diese Phase lässt sich in methodischer Hinsicht in zwei Abschnitte unterteilen: Aufgabenentwicklung und Aufgabenanalyse.

2 Theoretische und methodische Grundlagen der Aufgabenentwicklung

2.1 Testkonstrukt: Vom Modell zu den Aufgaben

Im Rahmen des *KoMus*-Projekts wurde zunächst eine Modellskizze entworfen, die auf fachdidaktischem Erfahrungswissen basiert und eine möglichst plausible Dimensionierung und Graduierung der Kompetenz beinhaltet (vgl. Niessen et al. 2008). Um eine Operationalisierung des Modells zu ermöglichen, muss in einem Testkonstrukt möglichst präzise beschrieben werden, was genau unter der Kompetenz ‚Musik wahrnehmen und kontextualisieren’ verstanden wird und welche Aspekte der Kompetenz durch den Test erfasst bzw. nicht erfasst werden. Das Testkonstrukt dient dabei einerseits einer theoretischen und empirischen Fundierung im Rahmen vorhandener (musikpsychologischer) Forschungen, andererseits können in diesem Zusammenhang die im theoretischen Modell noch relativ abstrakt formulierten Kompetenzdimensionen und -facetten ausgeschärft und konkretisiert werden. Denn erst wenn man dies „elaboriert hat […], wird man aus der Konstruktdefinition Verhaltensweisen ableiten können, die bei hohen oder geringen Ausprägungen auf dem Konstrukt beobachtbar sein sollten“ (Köller, 2008, S. 166). Die mit einem Modell bzw. Test anvisierte Schülerschaft bestimmt den Rahmen, innerhalb dessen ein

valides Testinstrument entwickelt werden kann. Die Aufgabenentwicklung orientierte sich in *KoMus* an Schülern der sechsten Jahrgangsstufe im Alter von etwa elf bis zwölf Jahren sowie an deren Erfahrungen und musikbezogenen Entwicklungsstand, und sie musste auf die Lernerfahrungen im Fach Musik ausgelegt sein. Aus diesem Grund basiert das Testkonstrukt sowohl auf musikpsychologischen Befunden als auch auf Curriculaanalysen. Das Testkonstrukt bildet somit die Gelenkstelle zwischen dem theoretischen Modell und dessen empirischer Umsetzung in Form von Testaufgaben.

Zusammenfassend ist die Kompetenz des Wahrnehmens und Kontextualisierens von Musik im Testkonstrukt mehrdimensional definiert.[6] Sie wird als ein Zusammenspiel von Hörwahrnehmungsfähigkeit und dem reflektierten Einsatz musikbezogener Wissensbestände verstanden. Durch die Berücksichtigung musikpsychologischer Befunde (z. B. Gembris 2005; Runfola & Swanwick 2002) ist einerseits sichergestellt, dass das zu entwickelnde Modell bzw. Testinstrument dem Entwicklungsstand der Schüler angepasst ist. Darüber hinaus kann das Verständnis musikbezogener Wahrnehmung, wie sie im *KoMus*-Projekt modelliert wird, auf Basis musikpsychologischer Grundlagenforschung konkretisiert werden (z. B. Bruhn 2005; Kreutz 2005; La Motte-Haber 2005; Lange 2005; Nauck-Börner 1987; Stoffer 2005). Durch Curriculaanalysen ist das Testkonstrukt in der unterrichtlichen Praxis verankert, wodurch eine curricular-inhaltliche Validität des Modells und der darauf bezogenen Testaufgaben angestrebt wird.

2.2 Testaufgaben: Funktionen, Gütekriterien und Formate

In der pädagogisch-psychologischen und fachdidaktischen Literatur werden verschiedene Aufgabensystematiken vorgeschlagen (z. B. hinsichtlich Inhalt, Funktion oder Format der Aufgaben). Für unseren Zusammenhang ist vor allem die funktionale Unterscheidung von Lern- und Testaufgaben wichtig (z. B. Benner 2007; Caspari, Grotjahn & Kleppin 2008). Während bei Lernaufgaben der Anregungsgehalt und das Lernpotenzial im Vordergrund stehen, besteht die Funktion von Testaufgaben darin, Kompetenzen (oder andere Merkmale) einer empirischen Überprüfung zugänglich zu machen. In diesem Sinne werden Testaufgaben zur Leistungsüberprüfung in Schulleistungsstudien, nicht zuletzt aber auch in Klassen- und Abschlussarbeiten eingesetzt. Dieser Praxis liegt die Annahme zugrunde, dass aus dem Lösen von Aufgaben

6 Detaillierte Ausführungen zum Testkonstrukt finden sich bei Knigge (2010) und Jordan et al.(i. Vorb.).

mit einer relativ hohen Sicherheit auf das Vorhandensein bzw. Fehlen der entsprechenden Kompetenzen bei Schülern geschlossen werden kann. Um das leisten zu können, müssen Testaufgaben bestimmten formalen und psychometrischen Kriterien genügen.

Die im *KoMus*-Projekt entwickelten Testaufgaben erfüllen zwei Funktionen: (1) Mittels der Aufgaben soll eine empirische Überprüfung und ggf. notwendige Modifikation des Kompetenzmodells, auf das sich die Aufgaben beziehen, erfolgen; (2) durch Analyseverfahren und Itemselektion (s. Abschnitt 3) soll aus den Aufgaben ein standardisiertes Testinstrument gebildet werden. Hierfür ist es erforderlich, dass die Aufgaben bestimmten Qualitätsansprüchen genügen. Neben den Hauptgütekriterien von Tests (Objektivität, Reliabilität und Validität; s. z. B. Moosbrugger & Kelava 2007) ist auch die Wahl eines geeigneten Aufgabenformats entscheidend für die optimale Erfassung einer anvisierten Kompetenz. Für die Entwicklung von Testaufgaben stehen verschiedene Item-Typen und -Formate zur Verfügung (vgl. auch Knigge 2010, Kap. 3.1). Jedes der Formate hat Vor- und Nachteile. Geschlossene Formate sind sehr ökonomisch in der Bearbeitung und Auswertung bei gleichzeitig maximaler Auswertungsobjektivität. Komplexe und kreative Fähigkeiten können aber oft nur schwer oder gar nicht mit geschlossenen Formaten erfasst werden (Rost 2004, S. 59 ff.). Offene Aufgaben sind hingegen eher geeignet für komplexere Anforderungen und das Antwortverhalten lässt sich leichter auf reale Situationen übertragen. ‚Erkauft’ wird dies jedoch mit einem relativ hohen Zeitaufwand bei der Bearbeitung der Aufgaben und vor allem bei der Auswertung. Es muss daher darauf geachtet werden, dass das gewählte Aufgabenformat und die anvisierte Kompetenz in einem entsprechenden Passungsverhältnis zueinander stehen, um eine ökonomische, aber gleichzeitig möglichst objektive und valide Messung zu gewährleisten.

3 Prozess der Aufgabenentwicklung im Projekt *KoMus*[7]

Auf Basis des Testkonstrukts konnte die Operationalisierung des Modells erfolgen. Ein Entwicklungsteam, bestehend aus sechs Kooperationslehrern und den Wissenschaftlern des *KoMus*-Projekts führte von Februar bis Dezember 2008 zehn Sitzungen zur Entwicklung von Testaufgaben durch. Der Prozess war zirkulär konzipiert: (1) monatliche Sitzung des Entwicklungsteams, (2)

7 Ausführlicher ist der Prozess der Aufgabenentwicklung bei Knigge 2010 (v. a. Kap. 4.3) dargestellt.

Erprobung (Feldtest) der Aufgaben in den Klassen der Kooperationslehrer, (3) Auswertung des Tests, (4) nächste Entwicklungssitzung: Überarbeitung der Aufgaben und Entwicklung neuer Aufgaben. Um eine möglichst effiziente und an den Testgütekriterien orientierte Aufgabenentwicklung sicherzustellen, wurde ein Handbuch mit ausführlichen Hinweisen zur Aufgabenkonstruktion erstellt (Knigge 2008).[8] Darin waren u. a. folgende Vorgaben festgelegt:

- Handlungsleitend für die jeweilige Aufgabenentwicklung sollte die vorab zu treffende Entscheidung sein, welche (Teil-)Kompetenz mit einer Aufgabe erfasst werden soll. Jeder Aufgabenentwurf war dementsprechend mit einer Beschreibung der intendierten Kompetenzmessung (Dimension und Niveau) zu versehen.
- Es sollten ca. 25% offene, 25% halb-offene und 50% geschlossene Items verwendet werden, wobei halb-offene und insbesondere offene Items hauptsächlich für komplexere Anforderungen auf höheren Kompetenzniveaus vorgesehen waren.
- Hörbeispiele sollten ein breites stilistisches Spektrum abdecken und möglichst nicht länger als 20-30 sec. sein.
- Für jede Aufgabe waren die Lösungen zu dokumentieren (wichtig v. a. bei offenen Formaten).

Auf dieser Basis sollten insgesamt mind. 120 Items entwickelt und überprüft werden.[9] Um im Rahmen der Tests eine möglichst objektive Testdurchführung zu gewährleisten, wurden die Aufgabenentwürfe (⌀ ca. 28 Items pro Sitzung) im Anschluss an die Entwicklungssitzungen in ein standardisiertes Testheft übertragen und zusammen mit einer Audio-CD und Anweisungen für die Testdurchführung (Testleiter-Manual) an die Kooperationslehrer zur Erprobung in ihren Klassen versandt. Zusätzlich erhielten Schüler und Lehrer einen Rückmeldebogen, sodass Verständnisschwierigkeiten und sonstige Probleme bei der Testdurchführung sofort festgehalten werden konnten.

---

8 Das Handbuch ist eine weiterentwickelte und speziell auf KoMus abgestimmte Fassung von Köller et al. 2005.

9 Zugrunde gelegt wurde hierbei die hypothetische Struktur des Modells von drei Kompetenzniveaus je Modelldimension (vgl. Niessen et al. 2008). Die avisierte Itemanzahl ergibt sich somit aus zehn Items pro Dimension und Niveau. So eine relativ große Anzahl von Items ist im Speziellen bei einer Neuentwicklung eines Testinstruments vonnöten, da im Zuge der Itemselektion ein gewisser Teil der Items aufgrund statistischer Kriterien verworfen werden muss.

4 Aufgabenanalyse

4.1 Itemselektion

Dank einer relativ hohen Schülerbeteiligung (⌀ pro Test N = 215) konnten bereits im Rahmen der Aufgabenentwicklung umfangreiche statistische Analysen durchgeführt werden. Mittels einer Kombination aus klassischen und probabilistischen[10] statistischen Verfahren, aber auch mithilfe der Anwendung von qualitativen Methoden wurden die entwickelten Aufgaben in *KoMus* systematisch auf ihre Messeigenschaften hin untersucht. Zunächst ging es hierbei im Sinne der Itemselektion darum, die besten Items für die Modellvalidierung bzw. für das zu erstellende Testinstrument zu identifizieren. Verschiedene Analysemethoden stellten die Grundlage dar, auf der Items ausgewählt, überarbeitet oder verworfen wurden. Tabelle 1 fasst die wichtigsten Analysen und die dabei angewandten psychometrischen Kriterien zusammen.[11]

Analyse	Kriterium
(Klassische) Itemschwierigkeit	$95 > P_i > 5$
Distraktoren	Alle Distraktoren ungefähr gleich häufig gewählt; Trennschärfe der Distraktoren < .05
Trennschärfe	$r_{it} \geq .25$
Itemfit (Rasch-Modell)	1.20 > MNSQ > 0.80
Differential Item Functioning (DIF)	Für ein Item liegt kein substantielles DIF vor[12]

Tabelle 1 Durchgeführte Itemanalysen im Rahmen der Aufgabenentwicklung

10 Zur Probabilistischen Testtheorie vgl. z. B. Rost (2004) und Bühner (2006). Die im Rahmen der Aufgabenentwicklung und -analyse eingesetzten probabilistischen Verfahren und Modelle (insbesondere das Rasch-Modell) sind bei Knigge (2010, Kap. 5) beschrieben.

11 Für eine ausführlichere Darstellung aller Analysemethoden und Kriterien s. Knigge 2010, Kap. 5 u. 6.1.

12 Als substantielles ‚Differential Item Functioning’ wurden signifikante DIF-Werte > 0.50 logits definiert (vgl. Wang 2000).

Items mit zufriedenstellenden statistischen Werten und unproblematischer Bearbeitung wurden in eine Item-Datenbank übernommen (‚selektiert'). Die übrigen Items wurden erneut im Rahmen der nächsten Entwicklungssitzung diskutiert und ggf. überarbeitet oder verworfen. Durch diesen zirkulären Prozess der Aufgabenerstellung, -erprobung und -überarbeitung konnte ein psychometrisch hochwertiger Itempool (179 Items) generiert werden, der die Strukturen des theoretischen Kompetenzmodells abbildet, auf unterrichtlicher und curricularer Ebene verankert ist und dabei sowohl eine Differenzierung über das gesamte Fähigkeitsspektrum (von Schülern der sechsten Jahrgangsstufe) ermöglicht als auch den strengen testtheoretischen Annahmen des Rasch-Modells genügt (vgl. Knigge 2010, Kap. 6.1). Die zentralen psychometrischen Kennwerte des Itempools sind in Tabelle 2 im Anhang dargestellt.

Des Weiteren war es ein zentrales Anliegen des Forschungsprojekts, über die statistischen Kennwerte hinaus mehr über die Messeigenschaften der entwickelten Aufgaben zu erfahren, da so ein besseres Verständnis der Strukturen der zu erfassenden Kompetenz erlangt werden kann. Konkret handelt es sich hierbei um die Identifikation sogenannter ‚schwierigkeitsgenerierender Aufgabenmerkmale', die im Folgenden ausführlicher dargestellt werden.

4.2 Identifikation von schwierigkeitsgenerierenden Aufgabenmerkmalen[13]

Als ‚schwierigkeitsgenerierende Aufgabenmerkmale' werden die Eigenschaften einer Testaufgabe bezeichnet, „die mit höheren oder niedrigeren Anforderungen an die getesteten Personen einhergehen und damit die Lösungswahrscheinlichkeiten der Aufgaben beeinflussen" (Hartig & Jude 2007, S. 31). Bei der Identifikation solcher Aufgabenmerkmale geht es folglich darum, die Charakteristika einer Aufgabe zu beschreiben, die in Bezug auf die Itemschwierigkeit relevant sind. Im Rahmen von *KoMus* wurde der Versuch unternommen, schwierigkeitsgenerierende Aufgabenmerkmale zu identifizieren, die sich auf die *Aufgabe*, das *Hörbeispiel*, den *Notentext* (sofern vorhanden), die *Wahrnehmungsanforderungen* und die für die Aufgabenlösung notwendige *Wissensbasis* beziehen. An dieser Stelle sollen zusammenfassend Ergebnisse dieser Analysen vorgestellt werden.

13 Die folgenden Abschnitte sind eine Zusammenfassung der bei Knigge (2010, Kap. 7) durchgeführten Analysen.

4.2.1 Vertiefende Analysen der Items zur Rhythmuswahrnehmung

In den meisten Curricula spielt die Rhythmuswahrnehmung eine wichtige Rolle. Schüler sollen sich u. a. Rhythmen merken können, diese wiedererkennen, aus motivischem Material Rhythmen extrahieren oder Rhythmen einem Notenbild zuordnen. Im Testkonstrukt wurde dementsprechend die Rhythmuswahrnehmung als eine Facette der Wahrnehmungskompetenz definiert und anschließend operationalisiert.

Die durchgeführten Analysen zeigen, dass eine vertiefende inhaltliche Betrachtung der Rhythmus-Items und deren empirisch ermittelter Schwierigkeit sehr aufschlussreich ist in Bezug auf die Aufgabencharakteristika und die darauf bezogenen Lösungsprozesse. Auf Basis der Analysen können die identifizierten schwierigkeitsgenerierenden Merkmale zusammenfassend folgendermaßen beschrieben werden (Knigge 2010, Kap. 7.1):

Die Schwierigkeit eines Rhythmus-Items ist abhängig von

- der Komplexität der klanglichen Struktur, in der ein Rhythmus identifiziert werden muss,
- der Komplexität der rhythmischen Struktur,
- dem notwendigen Wissen in Bezug auf Notation,
- der Anwendungsform von Notationskenntnissen,
- den Anforderungen an das musikalische Gedächtnis.

Zusätzlich können auch Merkmale identifiziert werden, die die Aufgabenlösung erleichtern: Wenn die klangliche oder melodische Struktur die rhythmische Struktur unterstützen und somit einprägsamer und leichter memorierbar machen, wirkt dies der Itemschwierigkeit entgegen.

4.2.2 Vertiefende Analysen der Items zur Formwahrnehmung

Eine weitere Facette des Kompetenzmodells bezieht sich auf die Wahrnehmung von musikalischen Formverläufen. Die Formwahrnehmung gehört, ebenso wie die Rhythmuswahrnehmung, zu den zentralen Fähigkeiten, deren Aufbau auf curricularer Ebene gefordert wird. Dort lässt sich auch ein breiter Konsens in Bezug auf die konkreten Inhalte ausmachen. Häufig wird dabei unterschieden zwischen elementaren Form-/Gestaltungsprinzipien (z. B. Wiederholung, Variation) und Formmodellen (z. B. Rondo). Für die Aufgabenentwicklung wurden nur Formen verwendet, die durch einen Großteil der Curri-

cula abgedeckt sind. Um darüber hinaus die unterrichtliche Relevanz der gewählten Formen noch weiter abzusichern, konnten die Ergebnisse von Schulbuchanalysen verwendet werden. Unter Berücksichtigung beider Quellen erfolgte die Auswahl von ‚Kanon', ‚Rondo' und ‚Liedformen'. Insgesamt wurden 13 Items zur Formwahrnehmung entwickelt, die sich größtenteils explizit auf ein bestimmtes Formmodell beziehen, teilweise aber auch die Wahrnehmung einzelner Formprinzipien fokussieren.

Mittels verschiedener Analysen können auch in Bezug auf die Items zur Formwahrnehmung eine Reihe von schwierigkeitsgenerierenden Aufgabenmerkmalen identifiziert werden (Knigge 2010, Kap. 7.2). Es lassen sich vier Merkmale formulieren, für die von einem Einfluss auf die Itemschwierigkeit auszugehen ist. Demgemäß ist die Schwierigkeit der Items zur Erfassung von Formwahrnehmung abhängig von

- der physischen Markierung von Abschnitten. Es wird angenommen, dass die Itemschwierigkeit steigt, umso weniger deutlich physische Markierungen die Segmentgrenzen zwischen zwei Abschnitten kennzeichnen;
- der notwendigen Nutzung des musikalischen Gedächtnisses. Sofern ein Formabschnitt (oder Teile davon) memoriert werden müssen, ist von einer erhöhten Itemschwierigkeit auszugehen;[14]
- der Länge und Komplexität der Formteile;[15]
- dem für die Aufgabenlösung notwendigen expliziten Wissen.

14 In Bezug auf den KoMus-Itempool kennzeichnet dieses Merkmal den Unterschied zwischen der Erkennung von Abschnitten und dem Vergleich von Abschnitten. Dies gilt selbstverständlich nur für diese Items und die Art, wie dort Formwahrnehmung erfasst wird. Da bei den KoMus-Items zur Abschnitterkennung die Segmentgrenzen immer durch relativ deutliche physische Markierungen gekennzeichnet sind, spielt hierbei das musikalische Gedächtnis keine oder nur eine untergeordnete Rolle. Bei komplexeren Formen (z. B. Fuge, Variation) ist es hingegen häufig nicht möglich die Segmentgrenzen zu bestimmen, ohne z. B. ein Thema, Motiv o. ä. zu memorieren.

15 Ein Einfluss dieses Merkmals wird nur für Aufgaben vermutet, die einen Vergleich von Formteilen verlangen. Sofern dieser nicht notwendig ist und auch zur Bestimmung der Segmentgrenzen kein musikalisches Gedächtnis erforderlich ist, sollte das Merkmal nicht schwierigkeitsrelevant sein.

In diesem und dem vorangegangenem Abschnitt konnten einerseits verschiedene Merkmale beschrieben werden, die in genuinem Zusammenhang mit der Wahrnehmung von Rhythmus und Form stehen. Andererseits erbrachten die Analysen aber auch Merkmale, die sich auf grundlegende Wahrnehmungsvorgänge, Eigenschaften eines Hörbeispiels oder die Art einer Aufgabenstellung beziehen, also vermutlich nicht ausschließlich auf die Rhythmus- und Formwahrnehmung beschränkt sind, sondern auf übergeordneter Ebene anzusiedeln sind. Auf diese Gruppe von schwierigkeitsgenerierenden Merkmalen soll in den folgenden beiden Abschnitten näher eingegangen werden.

4.2.3 Wissensbasierte Aufgabenmerkmale

Zunächst mag es verwundern, dass explizites Fachwissen im Rahmen eines Kompetenzmodells bzw. darauf bezogener Testaufgaben eine größere Rolle spielt. Wird doch in der Diskussion um Input- und Outputsteuerung immer wieder betont, dass mit dem Kompetenzbegriff eine Bewegung weg von der Vermittlung einzelner Inhalte, von der Fokussierung auf Faktenwissen, hin zu einem an Fähigkeiten und Fertigkeiten orientieren Lehr-/Lernkonzept verbunden ist. Betrachtet man jedoch den Kompetenzbegriff genauer (vgl. Abschnitt 1), so wird deutlich, dass dieser keineswegs den Stellenwert von Wissen grundsätzlich infrage stellt. Vielmehr geht es darum, den Erwerb von Wissen in einen größeren und vor allem anwendungsbezogenen Zusammenhang zu stellen: „Kompetenz stellt die Verbindung zwischen Wissen und Können […] her und ist als Befähigung zur Bewältigung von Situationen bzw. von Aufgaben zu sehen“ (Klieme et al., 2003, S. 73). Wissen ist also eine zentrale Facette von Kompetenz, die jedoch keinen Wert an sich hat, sondern eher im Sinne einer Ressource verstanden wird, die Schüler in die Lage versetzt, mit den an sie gestellten Handlungsanforderungen sinnvoll umgehen zu können (Criblez et al. 2009, S. 36).

Sofern ein Kompetenzmodell und ein darauf bezogenes Testverfahren beanspruchen auf der theoretischen Basis des Kompetenzbegriffs konstruiert und curricular valide zu sein, so ist es sinnvoll und notwendig auch explizites Fachwissen bei der Konstruktion von Testaufgaben mit einzubeziehen. Im Anschluss an die vorigen Ausführungen ist hierfür entscheidend, dass es sich dabei nicht um die isolierte Abfrage einzelner Inhalte handelt (z. B. „Wann wurde Mozart geboren?“), sondern das Wissen in konkreten Anforderungssituationen angewandt werden muss. Dieser Grundsatz wurde bei der Konstruktion der *KoMus*-Items eingehalten, sodass der Einsatz von Wissen bei einem Item immer nur ein Aufgabenmerkmal unter anderen ist.

Die durchgeführten Analysen zu wissensbasierten Aufgabenmerkmalen zeigen, dass grundsätzlich von einem schwierigkeitsrelevanten Einfluss auszugehen ist, wenn für die Lösung einer Aufgabe der Einsatz von Fachwissen notwendig ist. Das Merkmal ‚Einsatz von Fachwissen' konnte außerdem noch weiter ausdifferenziert werden (Knigge 2010, Kap. 7.3):

- Die Itemschwierigkeit ist nicht nur davon abhängig, ob das Merkmal grundsätzlich vorliegt, auch die Qualität des Wissens ist von Relevanz. Demgemäß ist von einer steigenden Itemschwierigkeit auszugehen, umso detaillierter und elaborierter das Wissen vorhanden sein muss.
- In inhaltlicher Hinsicht erscheint die Aufschlüsselung des Merkmals nach Wissensdimensionen sinnvoll. Es ergeben sich somit vier wissensbasierte Merkmale, die sich in Bezug auf die inhaltliche Dimension unterscheiden (musiktheoretisches, -historisches und -stilistisches Fachwissen sowie Wissen in Bezug auf kulturelle und soziale Kontexte von Musik).

4.2.4 Merkmalsebene ‚Aufgabe'

In Studien, die den Einfluss von Aufgabenmerkmalen auf die Itemschwierigkeit untersuchen, werden häufig die verwendeten Itemformate als Merkmale beschrieben (z. B. Prenzel et al. 2002). Hinter dieser Vorgehensweise steht die Annahme, dass die Schwierigkeit, beispielsweise einer Mathematik-Aufgabe, nicht nur von den mathematischen Kompetenzen der Schüler abhängt, sondern auch durch das Itemformat beeinflusst wird. Es wird angenommen, dass es Schülern grundsätzlich schwerer fällt, eine Antwort eigenständig zu formulieren (freie Formate: halboffen, offen), als eine vorgegebene Antwortalternative auszuwählen (geschlossene Formate: Multiple-Choice-, Richtig-Falsch-, Zuordnungs-Items). Hierfür sind in der Regel zwei Faktoren ausschlaggebend. Einerseits stellt ein freies Format erhöhte Anforderungen an die sprachlichen Fähigkeiten (Textproduktion/Schreibleistung). Andererseits spielen auch motivationale Aspekte eine Rolle, denn ein freies Aufgabenformat ist in Bezug auf die rein technische Bearbeitung immer aufwendiger zu lösen als ein geschlossenes Format, bei dem z. B. lediglich eine Antwort angekreuzt werden muss. Ein weiteres Merkmal, das auf der Ebene der technischen Oberflächencharakteristika einer Aufgabe anzusiedeln ist, sind die sprachlichen Anforderungen, die durch die textspezifische Beschaffenheit des Itemstamms, aber auch der Antwortalternativen gegeben sind.

Auf Basis verschiedener Itemanalysen werden auf der Merkmalsebene ‚Aufgabe' folgende Merkmale spezifiziert (Knigge 2010, Kap. 7.4):

- Das Merkmal ‚Itemformat’ umfasst drei Ausprägungen: geschlossene, halboffene und offene Formate.
- Das Merkmal ‚Textlänge’ bezieht sich sowohl auf den Itemstamm als auch auf die Antwortalternativen. Die Items werden dabei klassifiziert in Items mit viel und wenig Text. In Bezug auf die KoMus-Items genügt solch eine dichotome Kodierung, da sich die Items in der Regel sehr deutlich hinsichtlich des Merkmals unterscheiden.
- Das Merkmal ‚formalsprachliche Anforderungen’ bezieht sich auf das in einer Aufgabe verwendete Vokabular (hochfrequente Wörter, weniger frequente Wörter, erweiterter Wortschatz) und die grammatikalischen Strukturen (einfache syntaktische Strukturen, komplexere Strukturen). Die Ausprägungen des Merkmals wurden dabei in Anlehnung an Nold & Rossa (2007) formuliert.

Eine Untersuchung des Zusammenhangs von Itemschwierigkeit und der auf Aufgabenebene identifizierten Merkmale ist aufschlussreich und wichtig, da so der Einfluss von diesen allgemeinen und eher technischen Oberflächencharakteristika einer Aufgabe unterschieden werden kann von den Aufgabenmerkmalen, die in genuinem Zusammenhang mit den anvisierten musikspezifischen Kompetenzen stehen. Die Analyse von nicht kompetenzspezifischen Merkmalen dient somit auch der Absicherung der Konstruktvalidität der Items bzw. des Kompetenztests.

4.2.5 Empirische Analyse der Zusammenhänge von Aufgabenmerkmalen und Itemschwierigkeiten

Die durch inhaltliche und vergleichende Analysen identifizierten Aufgabenmerkmale wurden abschließend einer statistischen Überprüfung unterzogen (Knigge 2010, Kap. 7.6). Ein geeignetes Verfahren, um die empirisch ermittelten Itemschwierigkeiten mit den Aufgabenmerkmalen in Beziehung zu setzen, ist die multiple lineare Regressionsanalyse (vgl. Hartig 2007). Hierbei wird untersucht, ob die Unterschiede der Itemschwierigkeiten unter Verwendung der Aufgabenmerkmale erklärt werden können. Das Ausmaß erklärter Unterschiede ist ein Indikator dafür, ob sich die angenommenen schwierigkeitsgenerierenden Merkmale durch die tatsächlichen Aufgabenschwierigkeiten bestätigen lassen. Darüber hinaus kann auf Basis der Regressionsanalysen auch beurteilt werden, ob einzelne Merkmale besonders bedeutsam für die Aufgabenschwierigkeit sind oder aber Merkmale eine eher geringe Erklärungskraft für die Schwierigkeit der Aufgaben besitzen.

Auf Basis der Regressionsanalysen können folgende Befunde berichtet werden:

1. Die Analysen zeigen, dass sich empirisch ein schwierigkeitsgenerierender Effekt der Aufgabenmerkmale auf die Itemschwierigkeiten nachweisen lässt. Die in den Regressionen durchweg hohen Varianzaufklärungen ($.628 \leq R^2_{korr} \leq .914$) sind aufgrund der Datenbasis mit der gebotenen Vorsicht zu interpretieren, deuten jedoch auf eine starke Vorhersagekraft der Merkmale hin.
2. Der Großteil der identifizierten Merkmale konnte in die Regressionsmodelle einbezogen und dadurch validiert werden. Lediglich die Überprüfung des Einflusses der beiden Merkmale ‚formalsprachliche Anforderungen' und ‚Fachwissen – soziale/kulturelle Kontexte' war aufgrund der Datenlage nicht möglich und muss weiterführenden Untersuchungen vorbehalten bleiben.
3. Betrachtet man den Einfluss der einzelnen Merkmale, so ist es aufgrund methodischer Grenzen nur bedingt möglich, diese hinsichtlich ihrer Stärke miteinander zu vergleichen. Tendenziell scheint sich aber anzudeuten, dass die Schwierigkeit eines Items primär durch die Anforderungen an die Hörwahrnehmung und das notwendige Fachwissen beeinflusst wird. Ebenso zeigt sich ein schwierigkeitsgenerierender Einfluss der Anforderungen an das musikalische Gedächtnis, der vermutlich aber weniger stark ist.

4.2.6 Fazit

Die Identifikation, detaillierte Beschreibung und empirische Überprüfung von schwierigkeitsgenerierenden Aufgabenmerkmalen ermöglicht ein tieferes Verständnis der zur Aufgabenbearbeitung notwendigen Prozesse und damit des anvisierten Kompetenzkonstrukts (vgl. auch Hartig & Jude, 2007, S. 31). Dieses Verständnis erlaubt eine präzisere Interpretation von Testdaten sowie eine differenzierte Formulierung eines Kompetenzmodells und kann nicht zuletzt für die zukünftige Konstruktion von Items und Tests genutzt werden.

5 Ausblick: Der Einsatz von ‚Cognitive Labs'

Abschließend soll eine Methode skizziert werden, die die Entwicklung und Auswertung von Testaufgaben unterstützen kann. Bei der Arbeit des *KoMus*-Aufgabenentwicklungsteams wurde immer wieder deutlich, wie wenig eigentlich darüber bekannt ist, auf welche Weise Schülerinnen und Schüler Aufga-

ben lösen. Die Strategien der Schüler zu kennen, wäre aber sowohl für die Konstruktion der Aufgaben wie für die Bestimmung der jeweiligen Kompetenzen wichtig gewesen. Ein qualitatives Verfahren für die Untersuchung von Lösungsstrategien stellt die ‚Cognitive-Lab'-Methode dar (z. B. Ericsson & Simon 1999). ‚Cognitive Labs' waren nicht Bestandteil der systematischen Aufgabenentwicklung im Rahmen von *KoMus*, vielmehr wurde diese Methode nur explorativ in Bezug auf einzelne Items angewendet. Im Ausblick am Ende dieses Beitrags möchten wir an einer Beispielaufgabe demonstrieren, wie auf diese Weise weitere Informationen über die Kompetenzen von Schülern und die Messeigenschaften von Items gewonnen werden können.

5.1 Begründung der Aufgabenwahl

Bei einer der *KoMus*-Aufgaben wurde den Schülerinnen und Schülern ein eintaktiger Rhythmus mit zwei verschiedenen Notenwerten vorgestellt, den sie anschließend unter insgesamt vier Rhythmen wieder heraushören sollten: eine der Aufgaben, die in Bezug auf Lebensnähe und Komplexität nicht zu den gelungensten gehört – auch wenn man dieses Format gelegentlich in interaktiven Kinderspielen und in den Medien vorfindet. Gerade an diesem einfachen Beispiel lässt sich aber gut demonstrieren, welche Fragen sich bei der Testkonstruktion stellen können. Bei der Erstellung dieses Items wurde im Kreis der Fachdidaktiker und Lehrenden deutlich, dass über die Strategien, die Schüler zu dessen Lösung anwenden würden, nur spekuliert werden konnte: So war nicht klar, ob die Probanden sich den Rhythmus eher beispielsweise mit Hilfe von Abzählen zu merken versuchen oder ob sie den auditiven Eindruck memorieren würden. Die gewählte Lösungsstrategie ist aber bedeutsam für die Aufgabenkonstruktion: Wenn die Probanden den Rhythmus im Arbeitsgedächtnis speichern, wird die Reihenfolge der Items wichtig: Dann nimmt die Schwierigkeit deutlich zu, je nachdem ob die richtige Antwort an erster, zweiter, dritter oder gar vierter Stelle erscheint. Wenn die Probanden jedoch die Schläge zählen oder sich die Struktur verbalisiert merken (z. B. „erst zwei kurze Schläge, dann noch drei lange“), spielt die Reihenfolge der Antwortalternativen keine so wichtige Rolle. Die Aufgabe sollte auf dem mittleren Schwierigkeitsniveau angesiedelt werden; es wurde entschieden, die richtige Antwort an die dritte Stelle zu setzen und die Aufgabe im Rahmen eines ‚Cognitive Labs' explorativ zu untersuchen.

5.2 Die Methode der ‚Cognitive Labs'

Lautes Denken als übergeordneter Begriff bezeichnet eine empirische Forschungsmethode, bei der Personen aufgefordert werden, „ihre Gedanken laut auszusprechen, während sie sich einer Aufgabe oder Tätigkeit widmen" (Bilandzic 2005, S. 362). Die Methode stammt als ‚thinking aloud technique' aus der US-amerikanischen Psychologie (u. a. Ericsson & Simon 1984). Aus den Audio- und Videoaufnahmen solcher Erhebungssituationen werden ‚verbal protocols' erstellt, die dann ausgewertet werden können. Ericsson und Simon unterscheiden grundsätzlich zwischen zwei verschiedenen Ergebnisarten des lauten Denkens: Zunächst kann in ‚verbal protocols' Denken beschrieben sein, das ohnehin in Form von Sprache vorliegt, ‚vocalisation' oder ‚talk aloud'. Davon unterscheiden sie ‚verbalisation' oder ‚think aloud', nämlich Denken, das in anderer Form, z. B. als Visualisierung, vorliegt und das erst in Wörter ‚übersetzt' werden muss (Nielsen 2002, S. 105).

In den 1980er Jahren entwickelten Forscher im Zusammenhang mit Testverfahren auf der Grundlage der ‚thinking aloud technique' die spezifischere Form der ‚cognitive labs', weniger häufig auch ‚cognitive interviews' genannt. Sie ist charakterisiert durch den Gegenstandsbezug und daraus folgend durch bestimmte methodische und methodologische Besonderheiten: „A cognitive lab is a method of studying the mental processes one uses when completing a task such as solving a mathematics problem or interpreting a passage of text" (Zucker, Sassmann & Case 2004, S. 2). Besonders geeignet sind ‚Cognitive Labs', um die Gründe für Ungereimtheiten und Probleme bei Testungen aufzudecken. So ist es mit ihrer Hilfe möglich, Items, die unterschiedlich interpretiert werden können, zu identifizieren, zu komplizierte oder zu schwierige Anweisungen auszumachen sowie Gründe für alle Arten von erstaunlichen Ergebnissen zu finden (Zucker, Sassmann & Case 2004, S. 4).

5.3 Der Einsatz der Methode in KoMus

Für die ‚Cognitive Labs', die im Rahmen von *KoMus* in explorativer Absicht durchgeführt wurden, wurde ein Design gewählt, das sich in der Datenerhebung eng an das Verfahren von Zucker et al. (2004) anlehnt:

> *„In a cognitive lab, a student completes test items and verbally reports his or her thoughts related to the item using a combination of 'think-aloud' sessions (concurrent verbal reporting) and interviews with the researcher after each item is completed (retrospec-*

> *tive verbal reporting) ... To obtain a more thorough understanding of the subject's mental processes, verbal reports are often combined with other behavioral data observed during the cognitive lab, such as how a subject uses a pencil and paper to solve a mathematics item“ (S. 4).*

Tatsächlich erwies es sich als wichtiges Element der Untersuchung, die Schülerinnen und Schüler genau zu beobachten. So steckte in einer kleinen Verzögerung der Antwort eine wichtige Aussage und so war es möglich, zunächst unverständlich erscheinendes Vorgehen durch Nachfragen zu erhellen.

Die Auswertung der Daten wurde in entscheidenden Punkten anders vorgenommen, als das von Ericsson und Simon (1984) vorgeschlagen wird. Für die Auswertung wurde ein Vorgehen gewählt, das sich an das der Grounded Theory Methodologie (Strauss 1994) anlehnt, in der Annahme, dass es für jeden Aufgabentyp und bei jedem einzelnen Probanden eine je besondere Art und Weise der Lösung geben könnte. Dieses Vorgehen erwies sich z. B. als sinnvoll in Fällen, in denen Probanden von Lösungsmöglichkeiten berichteten, die sie tatsächlich aber gar nicht angewendet hatten. Zudem wurde, anders, als Simon und Ericsson es vorschlagen, in den Auswertungsprozess Kontextwissen einbezogen und die Umstände der Erhebungssituation reflektiert: Die Frage, ob die Probanden über musikpraktische Erfahrung verfügten, erschien ebenso wichtig wie der Einbezug der Qualität des Ergebnisses. Und: Die Analyse erfolgte nicht anhand eines vorgefertigten Kategoriensystems, wie es von Ericsson und Simon empfohlen wird. Tatsächlich sollten ja gerade unerwartete Lösungsstrategien identifiziert und den Äußerungen der Probanden möglichst unvoreingenommen begegnet werden.

Es wurden ca. 20 Schülerinnen und Schüler im Alter von 10 bis 13 Jahren in ‚Cognitive Labs' befragt, allerdings liegen die Ergebnisse in unterschiedlich gut dokumentierter Form vor: ‚Cognitive Labs' mit etwa 15 Schülern fanden im Rahmen eines Universitätsseminars statt und wurden nur mit Hilfe von Gesprächsprotokollen festgehalten; zudem wurden dort jeweils drei bis vier Probanden in kleineren Gruppen befragt, so dass von einer gegenseitigen Beeinflussung durch die Äußerungen der Mitschüler ausgegangen werden muss. Aus den ‚Cognitive Labs' mit fünf weiteren Probanden liegen wörtliche Transkriptionen vor, wobei zwei davon sehr ausführlich sind. In diesen beiden letzten ‚Cognitive Labs' wurden die Lösungen von nur zwei Items begleitet, was ausführlichere Gespräche über die jeweiligen Items erlaubte. Die Auswertung der Daten erfolgte schließlich mit Hilfe eines mehrstufigen Kodierverfah-

rens, das dem Vorgehen im Rahmen der Grounded Theory ähnelt, ohne dass insgesamt im Sinne dieses Ansatzes geforscht worden wäre: Im ersten Schritt wurde kleinschrittig und möglichst ‚offen' kodiert, im zweiten Schritt wurden die Kodierungen systematisiert, verglichen und teilweise stärker miteinander in Beziehung gesetzt. Im dritten Arbeitsschritt wurden dann die Lösungswege der Schüler und Schülerinnen rekonstruiert. Dabei wurde berücksichtigt, ob die Probanden bei der Aufgabenlösung erfolgreich waren und ob sie über längere musikpraktische Erfahrung[16] verfügten.

Zunächst schien die Suche nach Strategien zur Lösung der Rhythmusaufgabe eine beeindruckende Fülle von Möglichkeiten zu Tage zu fördern. Bei genauerem Hinsehen reduziert sich aber die Vielfalt und lässt sich in folgenden Thesen kondensieren:

- Es gibt nur eine Strategie, die relativ zuverlässig die Lösung der Aufgabe ermöglicht: Sie besteht aus dem Memorieren des Rhythmus in einer nichtverbalisierten Form. Von dieser Strategie berichten im Rahmen der ‚Cognitive Labs' vor allem Schüler, die über musikpraktische Vorbildung verfügen.
- Die kognitiven Strategien, die außerdem beschrieben werden, also das verbalisierte Memorieren von Folgen längerer oder kürzerer Notenwerte, werden eher von Schülern ohne intensive musikpraktische Erfahrungen beschrieben und führen bei ihnen nicht zum Erfolg.

5.4 Ausblick

Für die Interpretation der Ergebnisse der ‚Cognitive Labs' ist es sinnvoll, die Daten der *KoMus*-Pilotierungsstudie ergänzend hinzuzuziehen: Insgesamt 59,8 % der Sechstklässler (N = 508) lösten die Aufgabe erfolgreich. Schüler, die mindestens 2 Jahre Instrumentalunterricht hatten, antworteten signifikant häufiger richtig (65,6 %) als Schüler, die keinen oder weniger als ein Jahr Unterricht hatten (53,5 %). Allerdings kann von einer Signifikanz nur auf der Basis

16 Auch wenn diese Etikettierung nicht unproblematisch ist: Unter „längerer musikpraktischer Erfahrung" wird im Folgenden verstanden, dass sich Schülerinnen und Schüler systematisch und mindestens mehrere Monate mit dem Erlernen eines Musikinstruments beschäftigt haben. Bei der Auswertung der Daten für *KoMus* wurde unterschieden zwischen Schülern, die kein Instrument spielen, solchen, die ein bis zwei Jahre ein Instrument gespielt haben und solche, die mehr als zwei Jahre Instrumentalunterricht hatten.

eines 7%-Signifikanzniveaus gesprochen werden; eine DIF-Analyse der Aufgabe ergab keinen signifikanten Unterschied in der Lösungswahrscheinlichkeit zwischen diesen beiden Gruppen. Darüber hinaus haben wir für die Interpretation auch Ergebnisse musikpsychologischer Forschung herangezogen, die ja schon bei der Erstellung des Testkonstrukts eine Rolle spielten. Im vorliegenden Kontext wurde insbesondere die Theorie des Arbeitsgedächtnisses nach Baddeley bedeutsam. Wichtige Begriffe in diesem Kontext sind das Speichermodul ‚phonetische Schleife’ und so genannte ‚rehearsal-Prozesse’, was das innerliche Wiederholen dessen bezeichnet, was man sich merken möchte. Betrachtet man die Daten zusammen mit dem theoretischen Hintergrund, so lassen sich mit aller gebotenen Vorsicht folgende Aussagen treffen:

- Die Aufgabe ist für Schülerinnen und Schüler (ohne vorherige intensive Schulung in Gehörbildung) vermutlich nur mit Hilfe der phonetischen Schleife und von rehearsal-Prozessen erfolgreich zu lösen. Rehearsal-Prozesse sind nötig, weil die Kapazität der phonetischen Schleife nicht ausreicht, vier verschiedene Rhythmen abzuspeichern. Das beschreiben die Schüler deutlich im Rahmen der ‚Cognitive Labs’.
- Diese Strategie wird vermutlich von Schülern mit und ohne Instrumentalunterricht angewendet. Schülern mit intensiver instrumentaler Vorbildung gelingt erwartungsgemäß die Nutzung dieser Strategie etwas erfolgreicher.
- Andere Strategien, die im Rahmen der ‚Cognitive Labs’ beschrieben wurden, wie beispielsweise das Abzählen von Schlägen, führten nicht zur richtigen Lösung.

Der explorative Einsatz von ‚Cognitive Labs’ im Kontext des Projekts *KoMus* ermöglichte detaillierten Aufschluss über Lösungsstrategien: Für diese Art von Aufgabe stellte sich heraus, dass mit einer Veränderung der Reihenfolge der Antwortalternativen der Schwierigkeitsgrad der Aufgabe deutlich vermindert oder gesteigert werden könnte. Hier passen die Befunde der ‚Cognitive Labs’ zu denen, die bei der Analyse der Items erzielt wurden: Die Anordnung der Distraktoren spielt eine wichtige Rolle (s. Abschnitt 4.2.1). Erkenntnisse wurden auch in Bezug auf die getestete Kompetenz gewonnen: Mit Hilfe dieser Aufgabe wird nicht etwa das rationale Erfassen und verbale Kodieren einer rhythmischen Gestalt getestet, sondern die erfolgreiche Nutzung des Arbeitsgedächtnisses. Die hier gewonnenen Ergebnisse passen exakt zur Beschreibung der mittleren Kompetenzausprägung, die bei der Item-Analyse entwickelt wurde, und präzisieren sie in aufschlussreicher Weise. Auch wenn nur eine relativ isolierte Kompetenzfacette näher beleuchtet wer-

den konnte, lässt sich an diesem Beispiel doch schon erkennen, dass eine genauere Kenntnis der Lösungsstrategien nicht nur Erkenntnisse über Kompetenzausprägungen, sondern auch Hinweise auf gezielte Fördermöglichkeiten bereitstellt.

Als Fazit kann festgehalten werden, dass die Methode der ‚Cognitive Labs' die Validierung eines Kompetenzmodells, wie sie im Rahmen von *KoMus* angestrebt wird, sinnvoll unterstützen kann, dass sie aber auch darüber hinaus, insbesondere wenn sie vor dem Hintergrund eines qualitativen Paradigmas durchgeführt wird, für musikpädagogische Forschung interessante Perspektiven eröffnet.[17]

17 Nur erwähnt werden soll hier die Chance, die der Einsatz von ‚Cognitive Labs' in der Ausbildung von Musiklehrern eröffnet: Studierende erfahren im engen Kontakt mit den Schülerinnen und Schülern von deren Lernproblemen und -schwierigkeiten sowie von Strategien, deren Kenntnis für die Gestaltung des eigenen Unterrichts hilfreich sein kann.

Anhang

Tabelle 2

Zusammenfassung der wichtigsten psychometrischen Kennwerte der selektierten Items (Itempool); entnommen aus: Knigge 2010, Kap. 6.1.3

Testheft Nr.	Itemanzahl insg./selektiert	Schwierigkeit (klassisch)		Schwierigkeit (Rasch)	Itemfit (MNSQ)		Trennschärfe		Reliabilität (EAP/PV)[1]
		Min/Max	MW (SD)	Min/Max	Min/Max	MW (SD)	Min/Max	MW (SD)	
1	20/15	12.71/68.36	39.58 (15.11)	-1.48/2.03	0.88/1.13	1.01 (0.07)	0.32/0.57	0.42 (0.08)	0.693
2	28/15	5.98/91.45	56.35 (24.60)	-2.43/2.10	0.81/1.15	0.99 (0.11)	0.30/0.65	0.45 (0.12)	0.797
3	26/15	23.68/94.24	59.36 (25.61)	-2.42/2.25	0.88/1.10	1.00 (0.06)	0.28/0.57	0.41 (0.10)	0.720
4	32/23	14.88/94.69	62.55 (30.40)	-2.20/3.18	0.83/1.11	1.02 (0.05)	0.26/0.52	0.37 (0.07)	0.738
5	41/29	15.77/90.04	57.61 (22.23)	-2.34/2.57	0.88/1.15	1.00 (0.07)	0.25//0.56	0.39 (0.09)	0.822
6	31/24	6.67/87.50	36.19 (24.26)	-2.46/2.54	0.89/1.08	1.00 (0.05)	0.19/0.61[2]	0.36 (0.10)	0.719
7	27/15	5.23/66.56	29.66 (19.54)	-1.67/2.42	0.94/1.11	1.01 (0.05)	0.28/0.53	0.39 (0.09)	0.574
8	34/24	6.20/90.08	33.39 (22.71)	-3.55/2.49	0.81/1.19	1.00 (0.09)	0.26/0.55	0.42 (0.08)	0.823
9	35/19	7.97/88.05	34.78 (23.00)	-2.83/1.48	0.86/1.07	1.00 (0.05)	0.27/0.53	0.37 (0.09)	0.632
Gesamt	275/179								

Schwierigkeit (klassisch) = Schwierigkeitsindizes Pi; Schwierigkeit (Rasch) = Itemparameter des Rasch-Modells; Itemfit (MNSQ) = ‚weighted mean square’ (Fitmaß des Statistikprogramms ConQuest); MW = Mittelwert; SD = Standardabweichung

1 Das EAP/PV-Reliabilitätsmaß (EAP = expected a posteriori, PV = plausible values) ist dem häufig in der Klassischen Testtheorie verwendeten Cronbachs α vergleichbar und führt meist auch zu sehr ähnlichen Resultaten (vgl. Rost, 2004, S. 382).

2 In Testheft 6 wurde ein Item trotz einer zu niedrigen Trennschärfe (.19) selektiert. Das Item wies in Bezug auf alle weiteren psychometrischen Kriterien gute Werte auf und war aus inhaltlicher Sicht unverzichtbar.

Literatur

Benner, D. (2007): Unterricht - Wissen - Kompetenz. Zur Differenz zwischen didaktischen Aufgaben und Testaufgaben. In: Benner, D. (Hg.): Bildungsstandards. Instrumente zur Qualitätssicherung im Bildungswesen. Kontroversen - Beispiele - Perspektiven. Paderborn: Schöningh, S. 124–138.

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

Büchter, A.; Leuders, T. (2005): Mathematikaufgaben selbst entwickeln. Lernen fördern - Leistung überprüfen. Berlin: Cornelsen Scriptor.

Bühner, M. (2006): Einführung in die Test- und Fragebogenkonstruktion. 2., aktualisierte und erw. Auflage. München: Pearson Studium.

Caspari, D.; Grotjahn, R.; Kleppin, K. (2008): Kompetenzorientierung und Aufgaben. Zur Unterscheidung zwischen Lern- und Testaufgaben. In: Tesch, B.; Leupold, E.; Köller, O. (Hg.): Bildungsstandards Französisch: konkret. Sekundarstufe I: Grundlagen, Aufgabenbeispiele und Unterrichtsanregungen. Berlin: Cornelsen Scriptor, S. 85–87.

Criblez, L.; Oelkers, J.; Reusser, K.; Berner, E.; Halbheer, U.; Huber, C. (2009): Bildungsstandards. Zug: Klett und Balmer (Lehren lernen - Basiswissen für die Lehrerinnen- und Lehrerbildung).

Cvetko, A.; Meyer, D. (2009): Problemlösen im Musikunterricht – Interdisziplinarität als Ausgangspunkt für eine kompetenzorientierte Perspektive. In: Schläbitz, N. (Hg.): Interdisziplinarität als Herausforderung musikpädagogischer Forschung. Essen: Die Blaue Eule (Musikpädagogische Forschung, Bd. 30).

Ericsson, K. A.; Simon, H. A. (1999): Protocol analysis. Verbal reports as data. 3. überarbeitete Auflage. Cambridge: Mit Press.

Gembris, H. (2005): Die Entwicklung musikalischer Fähigkeiten. In: La Motte-Haber, H. de; Rötter, G. (Hg.): Musikpsychologie. Laaber: Laaber (Handbuch der Systematischen Musikwissenschaft, 3), S. 394–456.

Hartig, J. (2007): Skalierung und Definition von Kompetenzniveaus. In: Beck, B.; Klieme, E. (Hg.): Sprachliche Kompetenzen - Konzepte und Messung. DESI-Studie (Deutsch Englisch Schülerleistungen International). Weinheim: Beltz (DESI Ergebnisse, 1), S. 83–99.

Hartig, J.; Jude, N. (2007): Empirische Erfassung von Kompetenzen und psychometrische Kompetenzmodelle. In: Hartig, J.; Klieme, E. (Hg.): Möglichkeiten und Voraussetzungen technologiebasierter Kompetenzdiagnostik. Eine Expertise im Auftrag des Bundesministeriums für Bildung und Forschung. Berlin: BMBF (Bildungsforschung, 20), S. 17–36.

Jank, W. (Hg.) (2007): Musik-Didaktik. Praxishandbuch für die Sekundarstufe I und II. 2. Auflage. Berlin: Cornelsen Scriptor.

Jordan, A.-K., Knigge, J., Lehmann-Wermser, A., Lehmann, A. C. & Niessen, A. (i. Vorb.). Entwicklung und Validierung eines Kompetenzmodells im Fach Musik – Wahrnehmen und Kontextualisieren von Musik.

Jordan, A.-K.; Lehmann, A. C.; Knigge, J. (2010): IRT-basierte Kompetenzmodellierung – Erste Ergebnisse der Validierung eines Kompetenzmodells für den Bereich „Musik wahrnehmen und kontextualisieren“. In: Knolle, N. (Hg.): Evaluationsforschung in der Musikpädagogik. Essen: Die Blaue Eule (Musikpädagogische Forschung, 31).

Klieme, E.; Avenarius, H.; Blum, W., et al. (Hg.) (2003): Zur Entwicklung nationaler Bildungsstandards. Eine Expertise. Berlin: BMBF (Bildungsforschung, 1).

Klieme, E.; Hartig, J. (2007): Kompetenzkonzepte in den Sozialwissenschaften und im erziehungswissenschaftlichen Diskurs. In: Prenzel, M.; Gogolin, I.; Krüger, H. (Hg.): Kompetenzdiagnostik (Zeitschrift für Erziehungswissenschaft Sonderheft, 8). Wiesbaden: VS Verlag für Sozialwissenschaften, S. 11–29.

Knigge, J. (2008): Hinweise zur Erstellung von Testaufgaben für das KoMus-Projekt. (unveröffentlichtes Papier). Universität Bremen, Institut für Musikwissenschaft und Musikpädagogik – Projekt KoMus.

Knigge, J. (2010): Modellbasierte Entwicklung und Analyse von Testaufgaben zur Erfassung der Kompetenz „Musik wahrnehmen und kontextualisieren“. Dissertation. Universität Bremen.

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 http://www.zfkm.org/sonder08-knigge-lehmannwermser.pdf, zuletzt geprüft am 27.08.2009.

Köller, O. (2008): Bildungsstandards – Verfahren und Kriterien bei der Entwicklung von Messinstrumenten. In: Zeitschrift für Pädagogik, Jg. 54, H. 2, S. 163–173.

Köller, O.; Böhme, K.; Winkelmann, H.; Bremerich-Vos, A.; Granzer, D.; Vock, M.; Pöhlmann, C.; Robitzsch, A.; Würfel, K. (2005): Hinweise zur Erstellung von Testaufgaben für das Projekt "Evaluation der Standards Deutsch in der Grundschule" ESDeG (Primarbereich, Jahrgang 4). (unveröffentlichtes Papier). IQB. Berlin.

Kreutz, G. (2005): Melodiewahrnehmung: Funktionen von Arbeitsgedächtnis und Aufmerksamkeit. In: La Motte-Haber, H. de; Rötter, G. (Hg.): Musikpsychologie. Laaber: Laaber (Handbuch der Systematischen Musikwissenschaft, 3), S. 185–207.

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. (Hg.): Musikpsychologie. Laaber: Laaber (Handbuch der Systematischen Musikwissenschaft, 3), S. 74–100.

Mislevy, R. J.; Steinberg, L. S.; Almond, R. G. (2002): On the Roles of Task Model Variables in Assessment Design. In: Irvine, S. H. (Hg.): Item generation for test development. Mahwah: Lawrence Erlbaum, S. 97–128.

Moosbrugger, H.; Kelava, A. (2007): Qualitätsanforderungen an einen psychologischen Test (Testgütekriterien). In: Moosbrugger, H.; Kelava, A. (Hg.): Testtheorie und Fragebogenkonstruktion. Heidelberg: Springer, S. 7–26.

Nauck-Börner, C. (1987): Wahrnehmung und Gedächtnis. In: La Motte-Haber, H. de (Hg.): Psychologische Grundlagen des Musiklernens. Kassel: Bärenreiter (Handbuch der Musikpädagogik, 4), S. 13–115.

Niessen, A. (2008): Leistungsmessung oder individuelle Förderung? Zur Funktion und Gestaltung von Aufgaben im Unterricht. In: Schäfer-Lembeck, H.-U. (Hg.): Leistung im Musikunterricht. Beiträge der Münchner Tagung 2008. München: Allitera (Musikpädagogische Schriften der Hochschule für Musik und Theater München, 2), S. 134–152.

Niessen, A. (2009): Kompetenzlernen und verstehende Auseinandersetzung mit Musik. Eine Antwort auf Christoph Richter. In: Diskussion Musikpädagogik, Nr. 44, S. 58–60.

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 27.08.2009.

Nold, G.; Rossa, H. (2007): Hörverstehen. In: Beck, B.; Klieme, E. (Hg.): Sprachliche Kompetenzen - Konzepte und Messung. DESI-Studie (Deutsch Englisch Schülerleistungen International). Weinheim: Beltz (DESI Ergebnisse, 1), S. 178–196.

Prenzel, M.; Häußler, P.; Rost, J.; Senkbeil, M. (2002): Der PISA-Naturwissenschaftstest: Lassen sich die Aufgabenschwierigkeiten vorhersagen? In: Unterrichtswissenschaft, Jg. 30, H. 2, S. 120–135.

Richter, C. (2008): Musikunterricht "von unten". Curriculare Arbeit und aufbauender Unterricht von den Schülern aus. In: Diskussion Musikpädagogik, Nr. 37, S. 11–20.

Richter, C. (2009): Musikunterricht am Scheidewege. Kompetenzlernen oder verstehende Auseinandersetzung mit Musik. In: Diskussion Musikpädagogik, Nr. 42, S. 7–8.

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

Runfola, M.; Swanwick, K. (2002): Developmental characteristics of music learners. In: Colwell, R.; Richardson, C. P. (Hg.): The new handbook of research on music teaching and learning. A project of the Music Educators National Conference. New York: Oxford Univ. Press, S. 373–397.

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

Strauss, A. L. (1994): Grundlagen qualitativer Sozialforschung. Datenanalyse und Theoriebildung in der empirischen soziologischen Forschung. München: Fink.