em
Beiträge empirischer Musikpädagogik
bulletin of empirical music education research

Vol. 10
September 2019
ISSN: 2190-3174

Sonja Nonte
Georg-August-Universität Göttingen

Jens Knigge
Nord University, Campus Levanger, Norway

Tobias C. Stubbe
Georg-August-Universität Göttingen

Differentielle Itemfunktionen und Messinvarianz in standardisierten Musiktests – eine Frage der Testfairness

Differential Item Functioning and Measurement Invariance of Standardized Music Assessments – an Issue of Test-Fairness

Zusammenfassung

Testfairness gilt als zentrale Voraussetzung für die Interpretation von Gruppenunterschieden und damit auch für das Ableiten von Handlungsempfehlungen im Bildungssektor. Oftmals werden dabei die Gruppenvergleiche anhand sozialer Referenzkategorien vorgenommen. Der Vergleich von Merkmalen zwischen verschiedenen Gruppen ist in der Regel jedoch kontextabhängig, sodass die Ergebnisse, sofern relevante Kontextmerkmale nicht kontrolliert werden, verzerrt sein können. Der Beitrag verdeutlicht, wie solch ein Test-Bias identifiziert und letztendlich Testfairness hergestellt werden kann. Basierend auf den Daten von Musikkompetenztests aus zwei Studien mit n = 499 Sechstklässlerinnen und Sechstklässlern (Studie I) sowie n = 773 Siebtklässlerinnen und Siebtklässlern (Studie II), zeigten sich im Rahmen von Messinvarianzanalysen unter Nutzung von konfirmatorischen Faktorenanalysen (CFA) sowie Analysen zu Differentiellen Itemfunktionen (DIF) Hinweise auf einen systematischen Test-Bias für die Gruppierungsmerkmale Geschlecht, häuslicher Sprachgebrauch, Erfahrung mit Gesangs- und Instrumentalunterricht sowie kognitive Grundfähigkeiten. Implikationen, Einschränkungen und zukünftige Herausforderungen für die Gewährleistung von fairen Vergleichen in der musikpädagogischen Forschung werden diskutiert.

Schlagwörter: *Item-Response-Theorie, Kompetenzmessung, Messinvarianz, Musikpädagogik, Testkonstruktion*

Summary

Test-fairness is one of the most important requirements when results are interpreted with regard to group-differences. Often, group comparisons are made along social categorizations. However, the comparison of groups is context sensitive, i.e. results will be biased as long as only a limited number of context-variables are taken into account. The aim of this paper is twofold: First, we describe and

discuss the relevance and theoretical assumptions of test fairness notably in view of an evidence based educational policy. After different aspects of test fairness are highlighted, two major statistical methods (Measurement Invariance Analysis, Differential Item Functioning) are described. In a second step, we use data sets from two studies to investigate test bias for two different music assessments. The samples consist of n = 499 sixth graders (study I) and n = 773 seventh graders (study II) respectively. We define subgroups of the samples in accordance with theoretical assumptions and findings from previous studies about disadvantaged groups in music assessments. We find evidence for test bias with regard to gender, language usage, cognitive ability and experience with music tuition. The results indicate the relevance of test fairness in music assessments, especially, if recommendations for educational practice and policy are made.

Keywords: *competencies, item response theory, measurement invariance; music education, testing*

1. Einleitung

Wenngleich auf methodologischer Ebene qualitative Studien in der empirischen Musikpädagogik seit längerem überwiegen, so ist doch für die vergangenen beiden Jahrzehnte ein durchaus beachtlicher Anstieg quantitativer Arbeiten zu verzeichnen (Niessen & Knigge, 2018), insbesondere, wenn Studien berücksichtigt werden, die im Schnittfeld von Musikpsychologie und Musikpädagogik zu verorten sind. Kennzeichnend ist für diesen Bereich unter anderem, dass zunehmend standardisierte und psychometrische Testverfahren zum Einsatz kommen. Exemplarisch sind die verschiedenen, in den letzten Jahren entwickelten musikbezogenen Kompetenztests zu nennen (im Überblick s. Hasselhorn & Knigge, 2018), genauso wie Erhebungsinstrumente aus den Bereichen der Offenohrigkeits- (z. B. Louven, 2016), Musikalitäts- (z. B. Law & Zentner, 2012; Müllensiefen, Gingras, Musil & Stewart, 2014) und Selbstkonzeptforschung (z. B. Fiedler & Spychiger, 2017; Nonte, 2013). Häufig werden die entwickelten Messinstrumente dazu eingesetzt, um bestimmte Personengruppen (Teilstichproben) miteinander zu vergleichen (bspw. Kinder, die an einem speziellen musikpädagogischen Programm teilnehmen, mit Kindern, die lediglich ‚normalen' Musikunterricht erhalten). Die psychometrischen Kriterien werden jedoch nur in wenigen Fällen reflektiert und überprüft, obgleich sie die Grundlage für die Durchführung entsprechender Vergleiche sind (s. Abschnitt 2). Der vorliegende Beitrag widmet sich diesem Desiderat und diskutiert dabei zunächst zwei methodische Vorgehensweisen im Kontext der C*onfirmatory Factor Analysis* (CFA) sowie der *Item Response Theory* (IRT) (s. Abschnitte 2 und 3), um diese anschließend anhand empirischer Daten, die aktuellen musikpädagogischen Studien entnommen sind, zu veranschaulichen (s. Abschnitte 4–7).

2. Hintergründe

2.1 Vergleiche in der empirischen Bildungsforschung

Das Thema der fairen Vergleiche ist in der empirischen Bildungsforschung, nicht zuletzt aufgrund der steigenden Bedeutung von Schulleistungsvergleichsstudien, allgegenwärtig (Fiege, Reuther & Nachtigall, 2011). Im Sinne einer evidenzbasierten Bildungspolitik werden Ergebnisse aus Leistungsvergleichsstudien auf Klassenebene aggregiert und anschließend an Schulen zurückgemeldet, um so im besten Fall Schulentwicklungsmaßnahmen auf Unterrichtsebene anzustoßen. Dabei

sehen Fiege et al. (2011) die adäquate Analyse der Daten als Voraussetzung für Veränderungsprozesse auf Unterrichtsebene.

Im Rahmen von Schulleistungsvergleichsstudien stellt sich darüber hinaus häufig die Frage der Referenz. Bezugnehmend auf die von Heckhausen (1974) formulierten Normen (kriterial/sachlich, individuell oder sozial) werden beispielsweise Ergebnisse aus Vergleichsarbeiten vornehmend kriterial/sachlich, anhand von erreichten Kompetenzniveaus, zurückgemeldet. Der kriteriale Vergleich kann entsprechend nur dort erfolgen, wo empirisch validierte und standardisierte Instrumente zur Erfassung von domänenspezifischen Kompetenzen vorliegen. In vielen Fächern trifft dies bislang jedoch erst teilweise zu. So liegen beispielsweise für das Fach Musik lediglich für die Teilbereiche der musikbezogenen Reproduktion (Hasselhorn, 2015) und Rezeption (Jordan, Knigge, Lehmann, Niessen & Lehmann-Wermser, 2012) Kompetenzmodelle und darauf basierende Kompetenztests vor (s. a. Hasselhorn & Knigge, 2018). Abgesehen von Leistungsdaten stehen oftmals auch motivationale Merkmale von Schülerinnen und Schülern sowie ihrer Lerngruppen im Vordergrund von Evaluationen. Liegen keine kriterial/sachlichen Bezugsnormen vor, dann erfolgt die Ergebnisinterpretation oftmals anhand von sozialen Kriterien, beispielsweise in Form von Gruppenvergleichen. So vergleicht Nonte (2013) das musikalische Selbstkonzept von Grundschulkindern, die Instrumentalunterricht erhalten, mit denen, die keinen Instrumentalunterricht erhalten. Der Vergleich von Merkmalen zwischen verschiedenen Gruppen ist in der Regel jedoch stark kontextabhängig, sodass die Ergebnisse, sofern relevante Kontextmerkmale nicht kontrolliert werden, als unfair gelten. Der Einbezug von sämtlichen Kontextmerkmalen in Form von Kovariaten ist in der Forschungspraxis aus testökonomischen Gründen oftmals nicht realisierbar, sodass sich Studien meist auf einige der zentralen Einflussfaktoren wie das Geschlecht, den sozioökonomischen Status oder die Schulform konzentrieren, um (einigermaßen) faire Vergleiche zu gewährleisten (Fiege et al., 2011).

Neben der Berücksichtigung etwaiger Störvariablen beziehungsweise disparitätserzeugender Merkmale für die Gewährleistung von fairen Vergleichen kommt auch den (antizipierten) Konsequenzen der Ergebnisrückmeldung eine große Bedeutung zu, der jedoch von der Testfairness im statistischen Sinne zu trennen ist. Der Begriff der Testfairness soll im Folgenden kurz skizziert werden.

2.2 Testfairness

Der Begriff der Testfairness ist grundsätzlich sehr weit und umfasst unter anderem auch übergreifende Fragen sozialer Gerechtigkeit (Kane & Bridgeman, 2017). Kline (2013) und Messick (1998) führen zwei zentrale Aspekte von Testfairness an. Dies ist zum einen die soziale Komponente (*consequential aspect*) und zum anderen die statistische Komponente (*evidential aspect*). Die soziale Komponente umfasst in erster Linie intendierte und nicht intendierte Konsequenzen der Folgen der Ergebnisnutzung. Hier wird der Aspekt der Verteilungsgerechtigkeit angesprochen. So wird ein Test als unfair empfunden, wenn beispielsweise eine bestimmte Gruppe auf der Grundlage der Ergebnisrückmeldung benachteiligt wird, etwa in Form von gewährten beziehungsweise nicht gewährten Fördermaßnahmen oder aufgrund einer Kategorisierung und damit möglicherweise einhergehenden Stigmatisierung bestimmter Gruppen.

Die statistische Komponente hingegen wird auch als Test-Bias bezeichnet und umschreibt einen konstanten oder systematischen Bias aufgrund von bestimmten Personenmerkmalen oder einer bestimmten Gruppenzugehörigkeit (Kane & Bridgeman, 2017). Dieser Bias tritt dann auf, wenn

psychometrische Eigenschaften von Messwerten oder ihre Interpretation von nominalen Kategorien, wie beispielsweise Geschlecht, Ethnie oder anderen Charakteristiken abhängen, insbesondere, wenn diese Abhängigkeiten nicht theoretisch erwartbar sind (Reynolds & Suzuki, 2013). Er liegt etwa dann vor, wenn Personen aus zwei verschiedenen Gruppen denselben beobachteten Testscore, aber nicht denselben Status auf der geschätzten latenten Variablen haben (Amelang & Schmidt-Atzert, 2006). Die soziale Komponente und die statistisch-technische Komponente von Testfairness hängen nicht zwangsläufig zusammen. So bedeutet die Abwesenheit eines Test-Bias nicht, dass damit auch Fairness im Hinblick auf soziale Aspekte (Verteilungsgerechtigkeit) hergestellt ist, und auch ein im sozialen Sinne als fair wahrgenommener Test muss nicht unbedingt frei von einem Test-Bias sein. Die Vergleichbarkeit eines Messinstruments wird auch als Messäquivalenz (Messinvarianz) bezeichnet: „Vollständige Messäquivalenz bzw. Messinvarianz eines Messmodells liegt vor, wenn die Anwendung des Messmodells einer latenten Variablen in unterschiedlichen Stichproben bei gleichen Erhebungswerten der Indikatorvariablen auch die gleichen Messwerte für die latente Variable (Konstrukt- bzw. Faktorwerte und Konstruktmittelwert) erbringt“ (Weiber & Mühlhaus, 2010, S. 233). Es existieren unterschiedliche statistische Verfahren zur Überprüfung eines Test-Bias. Etabliert haben sich insbesondere zwei Verfahren, die im Folgenden kurz erläutert werden sollen.

2.2.1 Item Response Theory – Differential Item Functioning

Langjährige Tradition hat die Überprüfung der Test-Fairness im Rahmen der Item Response Theory (IRT). Im IRT-Ansatz wird anhand von differentiellen Itemanalysen, dem sogenannten *Differential Item Functioning* (DIF) überprüft, ob Itemschwierigkeit, Diskrimination und Rateeffekt zwischen den Gruppen variieren, obgleich sie dieselbe Fähigkeit aufweisen (Angoff, 1993; Stubbe, 2011). Die IRT geht von einem logistischen Zusammenhang des beobachteten Antwortverhaltens und dem zu messenden latenten Konstrukt aus (Schwab & Helm, 2015). Damit eignen sich DIF-Analysen als Screeningverfahren zur Analyse einzelner Items beziehungsweise eines Sets an Items (Camilli, 2013). Sie ermöglichen eine Entscheidung darüber, ob Testitems auf dieselbe Art und Weise in verschiedenen Gruppen von Testpersonen messen. Ein Beispiel für ein Item, welches mit hoher Wahrscheinlichkeit DIF aufweist, findet sich bei Schwabe (2014). Anhand einer hypothetischen Mathematikaufgabe wird deutlich, dass Aspekte wie die Lesekompetenz und fachspezifisches (nicht mathematikbezogenes) Wissen, die nicht als zentrale Prüfkriterien fungieren, zu einer Verzerrung (Bias) hinsichtlich der Schätzung der Personenfähigkeit führen kann. Bei der exemplarischen Aufgabe handelt es sich um eine Textaufgabe, bei der die zentrale Fähigkeit „Addition“ geprüft werden soll. Gegenstandsbereich der Textaufgabe ist ein Sinfonieorchester, wobei die Zusammenstellung des Orchesters im Text weiter anhand von Fachbegriffen wie „solistisch“ expliziert wird. Die abschließende Frage lautet „Wie viele Augenpaare richten Streicher der Berliner Philharmoniker auf den Taktstock des Dirigenten“ (Schwabe, 2014, S. 35). Die Ursache für DIF kann in diesem Fall domänenspezifisches (= musikbezogenes) Wissen sein, das die Wahrscheinlichkeit, das Item zu lösen, erhöht, aber nicht bei allen Testpersonen gleichermaßen vorhanden ist.

2.2.2 Confirmatory Factor Analysis – Messinvarianz

Darüber hinaus hat sich ein weiteres Verfahren etabliert, die Messinvarianzprüfung im Rahmen der C*onfirmatory Factor Analysis* (CFA), welches der Methode der Strukturgleichungsmodellierung zuzuordnen ist. Dieses Verfahren ist vergleichsweise jünger (Jöreskog, 1970) und ermöglicht eine auf Kovarianzen beruhende Betrachtung eines Messinstruments, bei dem der beobachtete Wert sich aus dem Intercept, der Faktorladung und dem Messfehler zusammensetzt (Meade & Lautenschlager, 2004). Hierbei wird ein einen linearer Zusammenhang zwischen dem Antwortverhalten und dem zu messenden latenten Konstrukt zugrunde gelegt. In der Praxis hat sich ein schrittweises Vorgehen etabliert, bei dem Schritt für Schritt verschiedene Grade an Messinvarianz anhand von Gruppenvergleichen überprüft werden (Schulte, Nonte & Schwippert, 2013; Steenkamp & Baumgartner, 1998). Damit wird sichergestellt, dass das zu messende Konstrukt über verschiedene Gruppen hinweg gleichermaßen reliabel und faktoriell valide anhand des genutzten Instruments erfasst wird. Zu diesem Zweck wird zunächst im Mehrgruppenmodell eine konfirmatorische Faktorenanalyse als Basismodell berechnet (*Multigroup-Confirmatory-Factor-Analysis*, MGCFA), wobei die einzelnen Parameter zwischen den Gruppen frei variieren. Eine Überprüfung der Gültigkeit der Passung des Modells wird anhand von inkrementellen und/oder absoluten Fit-Werten vorgenommen. Stimmt das hypothetische Modell mit den beobachteten Daten überein, dann kann angenommen werden, dass konfigurale Messinvarianz vorliegt. Diese kann hingegen verletzt sein, wenn beispielsweise das Konstrukt zu abstrakt ist, in verschiedenen Kulturen oder Gruppen unterschiedlich definiert wird, Probleme bei der Datenerhebung bestanden oder Übersetzungsfehler vorliegen. In einem nächsten Schritt werden die jeweiligen Faktorladungen über die Gruppen hinweg fixiert. Zeigt sich im Vergleich zum Basismodell keine signifikante Abweichung von den dort geschätzten Fit-Werten, kann angenommen werden, dass schwache faktorielle Messinvarianz vorliegt. Diese wird auch als metrische Messinvarianz bezeichnet und ermöglicht einen Vergleich von Beziehungsstrukturen etwa im Strukturmodell, nicht jedoch anhand der geschätzten latenten Mittelwerte. In einem weiteren Schritt werden neben den Faktorladungen auch die Intercepts (Schwellenwerte bei kategorialen Daten) über die Gruppen hinweg fixiert. Liegt erneut keine deutliche Abweichung von den Fit-Werten des Basismodells vor, wird strenge beziehungsweise skalare faktorielle Invarianz angenommen, welche die Voraussetzung für den Vergleich geschätzter latenter Mittelwerte ist. Anhand einer Fixierung der Faktorladungen, Intercepts und Residualvarianzen kann zusätzlich noch überprüft werden, ob für das untersuchte Instrument die Annahme strikter Messinvarianz bestätigt werden kann. Nur in diesem Fall kann davon ausgegangen werden, dass das Messinstrument reliabel und faktoriell valide und mit dem gleichen Ausmaß an Messfehlern das beobachtete Konstrukt misst. Meredith (1993) merkt an, dass dieses Ausmaß an Messinvarianz in der Realität nur schwer zu erfüllen ist.

2.2.3 Gegenüberstellung CFA und IRT

Analysen zu DIF im Kontext der IRT und Messinvarianzanalysen im Rahmen von CFAs haben sich in den vergangenen Jahren als Standardverfahren zur Überprüfung eines möglicherweise vorliegenden Test-Bias in der empirischen Bildungsforschung etabliert. Anhand einer Vielzahl an

Publikationen und Studien wurden bereits methodische Gegenüberstellungen der beiden Verfahren vorgenommen (siehe exemplarisch Kim & Yoon, 2011; Meade & Lautenschlager, 2004; Reise, Widaman & Pugh, 1993; Schwab & Helm, 2015). Zusammenfassend lässt sich konstatieren, dass beide Verfahren spezifische Stärken und Schwächen haben. So ist der CFA-Ansatz durch die Annahme eines linearen Zusammenhangs zwischen der latenten Variable und seinen Indikatoren analog zu den weit verbreiteten klassischen linearen Regressionsmodellen zu betrachten (Reise et al., 1993). In der IRT wird hingegen berücksichtigt, dass der Zusammenhang zwischen der latenten Variable und dem Antwortverhalten nicht notwendigerweise linear ist. Diese Einschränkung findet sich bei der IRT deshalb nicht, da hier Schwellenwerte (*thresholds*) geschätzt werden. Grundsätzlich finden sich mittlerweile jedoch auch im CFA-Ansatz Verfahren für die Schätzung kategorialer bzw. ordinaler Daten (Kim & Yoon, 2011; Millsap & Yun-Tein, 2004; Sass, Schmitt & Marsh, 2014). Messinvarianzanalysen im Rahmen von CFAs verfolgen einen globaleren Ansatz, der eine Ausweitung der jeweiligen Modelle im Rahmen des Strukturgleichungsansatzes ermöglicht und Beziehungen zwischen Faktoren zulässt, wohingegen DIF-Analysen sich insbesondere als Screening-Verfahren auf Einzelitemebene eignen. Zudem weisen Reise et al. (1993) darauf hin, dass CFA-Verfahren fortgeschrittener, einfacher und nutzerfreundlicher seien als IRT-Verfahren. Abgesehen von Herausforderungen wie beispielsweise Schätzproblemen und der Spezifikation von Modellen, die einen adäquaten Modellfit aufweisen, finden sich im Rahmen von CFAs eine größere Anzahl an Indizes zur Bestimmung der Modellgüte sowie implementierte Testverfahren zur Überprüfung von Messäquivalenz (X^2-Differenzentest). Diese finden sich nicht in vergleichbarer Weise im IRT-Ansatz. Auch bei einer Prüfung auf der Grundlage einer größeren Anzahl von Gruppen (Teilstichproben), scheint die CFA vorteilhaft zu sein. Die IRT liefert jedoch hinsichtlich Einzelitems oder einzelner Tests mehr psychometrische Informationen als CFAs (Meade & Lautenschlager, 2004).

Allgemein wird angenommen, dass die Kombination beider Verfahren ein ganzheitliches Bild zum Vorliegen eines möglichen Test-Bias zeichnen kann. So schreiben Meade und Lautenschläger (2004) „researchers and practitioners may receive an incomplete ME/I [Measurement Equivalence/Invariance] picture of the psychometric properties of a measure by using only one methodology" (S. 383).

Auch im Bereich der musikpädagogischen Forschung bildet die Überprüfung von Test-Bias auf der Grundlage von DIF- und Messinvarianz-Analysen einen vielversprechenden Ansatz für faire Vergleiche. Dies gilt in besonderem Maße, da Kompetenzmodelle im Bereich der Musik nur vereinzelt vorliegen (Hasselhorn & Knigge, 2018) und damit kriteriale Vergleiche, beispielweise in Leistungs- und Evaluationsstudien, bislang kaum etabliert sind.

3. Forschungsstand

Im Folgenden werden Befunde aus aktuellen Studien vorgestellt, denen jeweils eines der beiden berichteten Analyseverfahren zugrunde liegt. Da es bis heute nur wenige musikpädagogische Arbeiten gibt, die sich mit dem Aspekt Test-Bias befassen, werden exemplarisch auch Studien aus dem Bereich der Pädagogik und pädagogischen Psychologie herangezogen.

In der Musikpädagogik befassten sich insbesondere Arbeiten im Kontext des KoMus-Projekts (Kompetenzmodell im Fach Musik) mit DIF-Analysen: Jordan (2014) verwendet einen vorliegenden Item-Bias als potentielles Ausschlusskriterium für die Selektion von Testitems. Sowohl für das

Geschlecht der Schülerinnen und Schüler als auch für die die außerschulische Instrumental- beziehungsweise Gesangserfahrung findet sie für eine kleinere Anzahl von Items einen Item-Bias; eine genauere Untersuchung der Item-Charakteristika, die zu diesem Bias führen könnten, erfolgt jedoch nicht. Knigge (2010) nutzt DIF-Analysen ebenfalls zur Überprüfung der Qualität der KoMus-Testitems und findet für sieben der insgesamt 179 Items einen Item-Bias bezüglich des Geschlechts der Schülerinnen und Schüler. Explorative inhaltliche Analysen der betroffenen Items deuten darauf hin, dass dies mit genderspezifischen Musik- und Instrumentenpräferenzen zusammenhängen könnte.

Für den musikpädagogischen Bereich finden sich zudem Hinweise darauf, dass die im Kontext der Studie zum Instrumentalunterricht in Grundschulen (SIGrun) eingesetzte Skala zur Erfassung des musikalischen Selbstkonzepts ebenfalls zwischen Mädchen und Jungen (Nonte, 2012), aber auch im zeitlichen Verlauf von der ersten bis zur vierten Klasse, Parametervariationen aufweist. Für das genutzte Messinstrument zur Erfassung des musikalischen Selbstkonzepts konnte damit das geforderte Mindestmaß an skalarer Messinvarianz nicht bestätigt werden, sodass Mittelwertvergleiche zwischen Jungen und Mädchen sowie über verschiedene Erhebungszeitpunkte nicht sinnvoll interpretiert werden können.

Einen breiteren Forschungsstand bieten benachbarte Fachdisziplinen wie die Psychologie oder die Erziehungswissenschaft. Interessant sind hier zunächst Studien, die auf der Basis von Mehrgruppen-Modellen Verstöße gegen die Annahme von Messinvarianz, beispielsweise aufgrund eines kulturellen Bias, berichten. Hierzu gehört unter anderem die Studie von Cheung und Rensvold (2002), die für das Persönlichkeitsinventar der Big-Five im Rahmen eines interkulturellen Vergleichs eine weitere Dimension fanden, die zwar in China, aber in westlichen Industrieländern so nicht beobachtet werden konnte. Damit zeigt sich ein Verstoß gegen die Annahme der konfiguralen Invarianz, sodass beispielsweise Mittelwertvergleiche auf der Grundlage des genutzten Messinstruments zwischen den beteiligten Staaten nicht sinnvoll vorgenommen werden können. Auch im Rahmen von internationalen Schulleistungsvergleichsstudien zeigt sich ein Verstoß hinsichtlich der Annahme, dass das zur Erfassung des Leseselbstkonzepts genutzte Instrument in einer Auswahl von Staaten mit einem ähnlichen wirtschaftlichen und kulturellen Hintergrund vollständig reliabel misst. Hier konnte das Vorliegen von skalarer Messinvarianz für das untersuchte Konstrukt nicht bestätigt werden (Schulte et al., 2013).

Auf der Grundlage eines Vergleichs einer Auswahl von deutschsprachigen Staaten fand Stubbe (2011) DIF für den Lesekompetenztest, der im Rahmen der Studie PIRLS (Progress in International Reading Literacy) eingesetzt wurde. Diesen Bias führte er auf die unterschiedliche Übersetzung des Erhebungsinstrumentes zurück. Hinsichtlich relativer Stärken und Schwächen von bestimmten Gruppenmitgliedern zeigten Walther, Schwippert, Lankes und Stubbe (2008), dass Mädchen und Jungen für bestimmte Aufgabentypen im Fach Mathematik eine höhere Lösungswahrscheinlichkeit aufwiesen, die unabhängig von ihrer Personenfähigkeit war. Sie argumentierten, dass Jungen vor allem bei komplexeren Aufgaben besser abschnitten als Mädchen und Mädchen gegenüber Jungen einen Vorteil beim Lösen von Routineaufgaben zu haben schienen. Auch Schwabe (2014) nahm DIF-Analysen im Kontext der Internationalen Grundschul-Lese-Untersuchung (IGLU 2001 und IGLU 2011) vor. Sie konnte zeigen, dass offene Antwortformate besonders häufig mit einem DIF einhergehen, da die zugrundeliegenden komplexen kognitiven Lösungsprozesse insbesondere für Schülerinnen und Schüler mit geringen Lesekompetenzen relativ gesehen zu einer geringeren Lösungswahrscheinlichkeit führen.

Ausgehend vom aktuellen Forschungsstand können als zentrale disparitätserzeugende Merkmale das Geschlecht sowie die außerschulische Instrumental- beziehungsweise Gesangserfahrung identifiziert werden. Es zeigen sich aber auch Hinweise auf den Einfluss kognitiver Fähigkeiten von Schülerinnen und Schülern auf die Bearbeitung von positiv und negativ gepolten Items, einem sogenannten Methodenartefakt. Hier zeigte Nonte (2012) für das Selbstkonzept der Schulfähigkeit, dass die latente Modellierung eines Methodenfaktors, der zusätzliche itemspezifische Varianz der invertierten Items beinhaltet, die beobachteten Daten besser beschreibt als ein Modell ohne solch einen Faktor. Ohne diesen Faktor zeichnet sich eine Benachteiligung von Schülerinnen und Schülern mit geringen kognitiven Fähigkeiten ab, was einen Test-Bias nahelegt.

Bislang liegen für Kompetenzmodelle im Bereich Musik, aber auch für andere musikbezogene Erhebungsinstrumente wie beispielsweise zur Erfassung von Interesse, Selbstkonzept und Freude am Instrumentalspiel nur sehr wenige Studien vor, die einen möglichen Test-Bias systematisch in den Blick nehmen und hinsichtlich praktischer Implikationen diskutieren. Dieses Desiderat wird im Folgenden aufgegriffen, indem anhand von zwei exemplarischen Studien aus dem musikpädagogischen Kontext die Anwendung von DIF- und Messinvarianzanalysen vorgestellt und Befunde abschließend im Hinblick auf die praktische Relevanz diskutiert werden.

4. Fragestellungen

Auf der Grundlage von Forschungsbefunden zu disparitätserzeugenden Merkmalen im Bereich der musikalischen Kompetenz- und Selbstkonzeptforschung lassen sich vier zentrale Bereiche identifizieren, für die ein Test-Bias wahrscheinlich erscheint. Dies ist das Geschlecht, die Erfahrung mit Gesangs- und Instrumentalunterricht, die kognitiven Grundfähigkeiten und der Sprachgebrauch in der Familie, wobei die beiden letztgenannten Konstrukte nicht konkret aus rezenten musikpädagogischen oder musikwissenschaftlichen Studien abzuleiten sind, sondern aus Befunden anderer Fachdisziplinen (s. o.). Zu vermuten ist, dass Jungen, Kinder mit geringen kognitiven Grundfähigkeiten, geringen Vorerfahrungen mit Gesangs- und Instrumentalunterricht sowie mit einem nicht-deutschen Sprachgebrauch zu Hause bei der Bearbeitung standardisierter Musiktests benachteiligt sind.

Der Beitrag geht demnach drei zentralen Fragestellungen nach:

1. Inwieweit kann für den in Studie I eingesetzten Musiktest zur Erfassung der Hörwahrnehmung skalare faktorielle Messinvarianz bestätigt werden, wenn das Geschlecht, die Erfahrung mit Instrumentalunterricht, die wöchentliche Übungszeit, die kognitive Verarbeitungsgeschwindigkeit sowie die kognitiven Fähigkeiten (Figurenanalogien) als zentrale Ungleichheitsdimensionen kontrolliert werden?
2. Inwieweit kann für den in Studie II eingesetzten Musiktest ebenfalls skalare faktorielle Messinvarianz bestätigt werden, sofern die zentralen Ungleichheitsdimensionen Geschlecht, Sprachgebrauch in der Familie und Erfahrung mit Instrumental-/Gesangsunterricht berücksichtigt werden?
3. Sofern in Frage 2 die Annahme von skalarer faktorieller Messinvarianz widerlegt wurde, welche Items weisen einen signifikanten DIF auf und in welchem Zusammenhang steht dies möglicherweise mit bestimmten Aufgabencharakteristika?

5. Erhebungsinstrumente & Datengrundlage

Die Analysen wurden auf der Grundlage von Datensätzen durchgeführt, die aus zwei voneinander unabhängigen Studien stammen. Die Studien verfolgen eigenständige Forschungsfragen, die teilweise auch bereits publiziert sind (s. u.). Für den vorliegenden Zusammenhang werden lediglich Teile der Datensätze genutzt, um im Rahmen von weiterführenden Analysen Aspekte der Testqualität und schließlich -fairness zu untersuchen, die in den ursprünglichen Studien keine oder nur eine untergeordnete Rolle spielten.

Studie I – ProBiNi (Stubbe, Nonte, Haas & Krieg, 2019)

Die Studie zur ProfilBildung an Niedersächsischen Gymnasien und Integrierten Gesamtschulen (ProBiNi, Laufzeit 2016–2019, DFG-Projektnummer: 312968144) ist eine Längsschnittstudie mit einem quasi-experimentellen Design und nimmt unter anderem die Entwicklung von Schülerinnen und Schülern in Musik- und MINT-Klassen sowie von Schülerinnen und Schülern in Vergleichsklassen von der fünften bis zur siebten Jahrgangsstufe in den Blick (Stubbe, Nonte, Haas & Krieg, 2019). Zu Beginn der sechsten Jahrgangsstufe wurden insgesamt 527 Schülerinnen und Schüler befragt. Fälle mit einem Missing-Anteil von mehr als 80 Prozent auf den interessierenden Variablen wurden ausgeschlossen, sodass das Analysesample Angaben von n = 499 Schülerinnen und Schülern umfasst (Ausschluss von 5.3 % der Fälle). Dabei handelt es sich um fünf MINT-Klassen und sechs Bläser-/Orchesterklassen mit jeweils der gleichen Anzahl an Vergleichsklassen in denselben Schulen. Insgesamt nahmen zum zweiten Messzeitpunkt 21 Klassen aus zehn Schulen an ProBiNi teil. Das durchschnittliche Alter der Schülerinnen und Schüler zum zweiten Erhebungszeitpunkt betrug M = 11.46 Jahre (SD = 0.44). Fehlende Werte wurden nicht imputiert, da in *Mplus 7.1* (Muthén & Muthén, 1998–2012) ein geeignetes Schätzverfahren (FIML) implementiert ist, welches fehlende Werte modellbasiert schätzt (Full Information Maximum Likelihood; Lüdtke, Robitzsch, Trautwein & Köller, 2007).

Der in Studie I (ProBiNi) verwendete Musiktest setzt sich zum zweiten Erhebungszeitpunkt aus insgesamt 25 Items zusammen. Von diesen 25 Items stammen 11 Items aus dem KoMus-Test (Kompetenz-Dimension 1: Hörwahrnehmung und musikalisches Gedächtnis; Jordan et al., 2012) sowie 14 Items aus dem IOWA Test of Music Literacy, Level 1 (tonal/rhythm concepts, listening, reading; Gordon, 1991). Der Test wurde als Paper & Pencil-Test administriert und dauerte etwa 15 Minuten. Geschulte Testleiterinnen und Testleiter führten die Erhebung mit den Schülerinnen und Schülern durch und spielten die jeweiligen Musikstücke über ein Abspielgerät ab, welches zentral im Raum positioniert wurde.

Alle Items wurden für diesen Beitrag dichotomisiert. Eine erste psychometrische Überprüfung des Gesamttests als 1-PL Modell im IRT-Ansatz ging mit einer geringen Reliabilität einher (EAP/PV = .53), sodass für den vorliegenden Beitrag auf eine Überprüfung im IRT Ansatz verzichtet wurde. Es zeigte sich jedoch für eine Auswahl an 7 Items, dass im Rahmen von CFAs ein Messmodell geschätzt werden konnte, für das ein guter Modell-Fit vorliegt (X^2 = 36.51; *df* = 31; *p* = n.s.; *RMSEA* = 0.03; *CFI* = 0.97; *TLI* = 0.96).

Schülerinnen und Schüler der Studie I nahmen zusätzlich an einer standardisierten Befragung zu ihren Interessen, den fachspezifischen Selbstkonzepten, zur Unterrichtsqualität sowie zu ihrem sozioökonomischen Hintergrund teil. Diese Items wurden mit Hilfe eines standardisierten Fragebogens erhoben. Als Ungleichheitsdimensionen wurden folgende Merkmale berücksichtigt:

- das Geschlecht (Junge/Mädchen),
- die Erfahrung mit Instrumentalunterricht in Jahrgang sechs oder/und fünf (ja/nein),
- die verwendete Zeit für das Üben des Musikinstruments (keine/bis zu einer Stunde pro Woche/mehr als eine Stunde pro Woche),
- die kognitive Informationsverarbeitungsgeschwindigkeit (niedrig/mittel/hoch; gemessen anhand des Zahlen-Verbindungs-Tests ZVT, Oswald & Roth, 2016) sowie
- eine Subdimension der kognitiven Fähigkeiten (Figurenanalogien, KFT 4-12+R; N2 Figurenanalogien; Heller & Perleth, 2000) (niedrig/mittel/hoch).

Bei den Gruppierungsvariablen *kognitive Informationsverarbeitungsgeschwindigkeit* (ZVT) und *kognitive Fähigkeiten* (KFT, N2) wurden relative Differenzierungskategorien gewählt. Die Gruppen wurden so gebildet, dass sie die Stichprobe in etwa drei gleich große Untergruppen aufteilt.

Studie II – Prädiktoren der Kompetenz „Musik wahrnehmen und kontextualisieren" (Harnischmacher & Knigge, 2017)

Harnischmacher und Knigge (2017) befassten sich ebenfalls mit der Erfassung musikalischer Kompetenzen von Schülerinnen und Schülern. Die Studie ist eine Querschnittstudie und wurde im Jahr 2015 durchgeführt. Im Rahmen von zwei Schulstunden (etwa 90 Minuten) wurden Schülerinnen und Schülern der Jahrgangsstufe sieben computer- und webbasiert (Moodle) getestet. Die Schülerinnen und Schüler (n = 773) stammten aus 9 Gymnasien und einer Integrierten Sekundarschule. 57.8 Prozent der Schülerinnen und Schüler waren weiblich und das Durchschnittsalter betrug 12.68 Jahre (SD = 0.63).

Der in Studie II genutzte Musiktest umfasst 27 Items, die eine Kurzversion des KoMus-Tests darstellen (KoMus_short). Die Items verteilen sich auf vier Kompetenz-Dimensionen, wobei die Gesamtreliabilität des Testes bei EAP/PV = .77 liegt. Die Reliabilitäten für die jeweiligen Dimensionen betragen .62, .77, .71 und .60. Anhand von bootstrapping-Verfahren und X^2-Tests konnte Rasch-Modellgültigkeit bestätigt werden. Auch im Rahmen einer CFA wiesen die guten Modell-Fits auf die Annahme von Eindimensionalität hin (X^2 = 351.17; df = 275; $p \leq 0.01$; $RMSEA$ = 0.02; CFI = 0.96; TLI = 0.96). Alle Items wurden für den vorliegenden Beitrag dichotomisiert. Auch hier wurden fehlende Werte mit Hilfe des in Mplus basierten Schätzers (FIML) modellbasiert geschätzt.

Zusätzlich füllten die Schülerinnen und Schüler einen Online-Fragebogen aus, in dem sie unter anderem demographische Angaben sowie Angaben zum zeitlichen Umfang des Musikunterrichts, zum Kompetenzerleben im Musikunterricht, zum Musikinteresse in der Familie, zu kulturellen Aktivitäten, zur Motivation im Musikunterricht, zu ihren Zensuren in Musik, zur Instrumental-/Gesangspraxis sowie zu Migration und zum Sprachgebrauch machten (für detaillierte Informationen zu den genutzten Erhebungsinstrumenten s. Harnischmacher & Knigge, 2017, S. 12ff.). Für den vorliegenden Beitrag werden folgende Ungleichheitsdimensionen aus Studie II berücksichtigt:

- das Geschlecht (Junge/Mädchen),
- der Sprachgebrauch in der Familie (spricht immer deutsch/spricht manchmal oder nie deutsch) sowie
- die Erfahrung mit Instrumental- oder Gesangsunterricht (keine/mehr als ein Jahr).

6. Datenanalyse I: Überprüfung von Test-Bias auf der Basis von CFAs

6.1 Methodisches Vorgehen

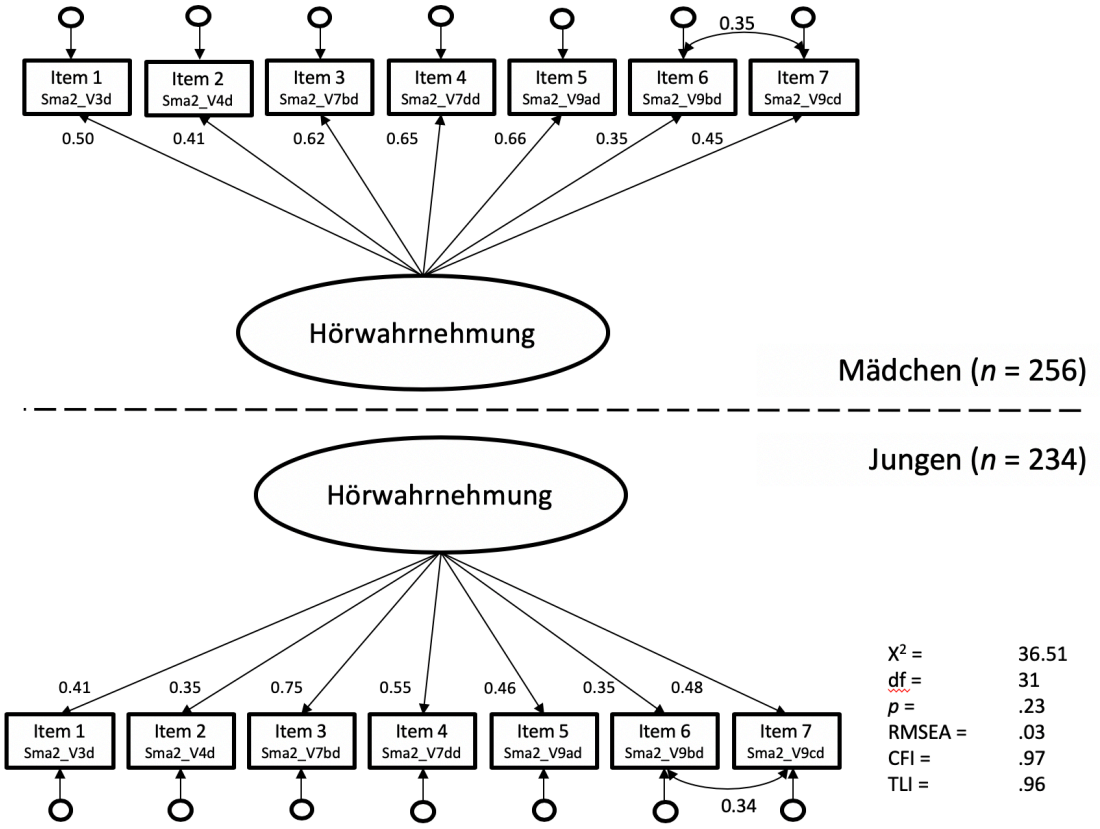
Die Überprüfung von Messinvarianz wird anhand von Mehrgruppen-Analysen im Kontext von CFAs vorgenommen. Wie bereits beschrieben erfolgt die Überprüfung für die möglicherweise für einen Bias verantwortlichen Gruppenmerkmale schrittweise (Steenkamp & Baumgartner, 1998). In einem ersten Schritt wird simultan für die Gruppen jeweils eine CFA geschätzt, in der alle Parameter frei variieren. Dieses Modell bildet den Ausgangspunkt für den Vergleich mit weiteren Modellen, bei dem schrittweise einzelne Parameter restringiert werden, und wird im Folgenden als Basismodell bezeichnet. Es prüft zudem das Vorliegen von Konstruktäquivalenz, die bestätigt werden kann, wenn die inkrementellen und absoluten Fit-Werte den beispielsweise von Hu und Bentler (1999) vorgeschlagenen Richtwerten entsprechen ($RMSEA \leq 0.06$, $CFI \leq 0.95$, $TLI \leq 0.95$, $X^2/df \leq 3$). Da es sich um kategoriale Daten handelt, wird ein robuster Schätzer (WLSMV) verwendet. In einem zweiten Schritt werden die Faktorladungen und Schwellenwerte über die Gruppen hinweg fixiert. Die jeweiligen Fit-Werte werden mit dem Basismodell verglichen. Anhand eines in *Mplus* implementierten X^2-Differenzentests wird automatisiert ermittelt, ob das restringierte Modell signifikant vom Basismodell abweicht. Ist dies nicht der Fall, kann daraus geschlossen werden, dass strenge (skalare) Invarianz vorliegt. Da es sich um kategoriale Daten handelt und damit keine Intercepts geschätzt werden können, wird in *Mplus 7.1* keine automatisierte Überprüfung von metrischer sowie strikter Invarianz vorgenommen.

6.2 Ergebnisse

Fragestellung 1: Testung auf Messinvarianz in Studie I (ProBiNi)

In einem ersten Schritt wird im Rahmen einer konfirmatorischen Faktorenanalyse ein Basismodell berechnet, in dem Faktorladungen, Schwellenwerte und Residualvarianzen zwischen den Gruppen frei variieren. Abbildung 1 zeigt dies für die Gruppierungsvariable *Geschlecht*.

Das Modell, in dem alle Parameter über die Gruppen frei variieren, eignet sich als Modell zur Überprüfung der konfiguralen Messinvarianz. Ist der Modell-Fit gut, kann angenommen werden, dass die Faktorstruktur über die Gruppen identisch ist und damit konfigurale Invarianz vorliegt. In einem nächsten Schritt wird die skalare faktorielle Messinvarianz überprüft. In diesem Modell werden Faktorladungen und Intercepts über die Gruppen hinweg gleichgesetzt. Ist die Abweichung des Modells zur Überprüfung skalarer faktorieller Invarianz vom vorherigen Messmodell marginal (der X^2-Differenzentest ist nicht signifikant), kann angenommen werden, dass das latente Konstrukt über die Gruppen hinweg auf derselben Metrik abgebildet werden kann. Dies ist, wie bereits beschrieben, die Voraussetzung für sinnvolle Mittelwertvergleiche und damit auch für die Umsetzung von fairen Vergleichen. In *Mplus* wurden für alle Merkmale, für die angenommen werden kann, dass sie ursächlich für einen Test-Bias sein könnten, Mehrgruppenvergleiche nach dem vorherigen Schema durchgeführt. Die Ergebnisse sind in Tabelle 1 dargestellt.



Anmerkungen: standardisierte Pfadkoeffizienten. Alle dargestellten Werte sind signifikant.

Abbildung 1: CFA: Mehrgruppenmodell mit Gruppierungsvariable Geschlecht in Studie I (ProBiNi)

Tabelle 1: Überprüfung von Messinvarianz im Rahmen von Mehrgruppenmodellen in Studie I (ProBiNi).

	X^2	df	X^2/df	p ≤	RMSEA	CFI	TLI	$X^{2\text{-Difference}}$	df	p ≤
Geschlecht										
Konfigural	33.49	26	1.29	n.s.	0.03	0.97	0.95			
Skalar	42.76	31	1.38	n.s.	0.04	0.96	0.94	9.43	5	n.s.
Instrumentalunterricht[1]										
Konfigural	26.71	26	0.68	n.s.	0.01	0.99	0.99	-	-	-
Skalar	-	-	-	-	-	-	-	-	-	-
Übungszeit										
Konfigural	44.68	39	1.15	n.s.	0.03	0.96	0.94			
Skalar	56.29	49	1.15	n.s.	0.03	0.95	0.94	11.73	10	n.s.
ZVT										
Konfigural	45.72	39	0.65	n.s.	0.03	0.98	0.96			
Skalar	69.68	49	1.42	0.05	0.05	0.93	0.90	22.76	10	0.05
KFT										
Konfigural	33.58	39	0.86	n.s.	0.00	1.00	1.05			
Skalar	38.99	49	0.80	n.s.	0.00	1.00	1.07	6.15	10	n.s.

Anmerkung: grau hinterlegte Felder kennzeichnen einen Verstoß gegen Messinvarianz; n.s. = nicht signifikant.
[1] keine Schätzung skalarer MI aufgrund einer negativen Residualvarianz.

Die Ergebnisse deuten darauf hin, dass der in Studie I (ProBiNi) eingesetzte Musiktest für zwei von fünf Gruppierungsmerkmalen keine Messäquivalenz aufweist. Dies ist zum einen die Gruppierungsvariable *kognitive Informationsverarbeitungsgeschwindigkeit* (niedrig/mittel/hoch) und zum anderen die Gruppierungsvariable *Erfahrung mit Instrumentalunterricht in Jahrgang sechs und/oder fünf* (ja/nein). Für das Modell skalarer faktorieller Invarianz in letztgenannter Gruppe treten Schätzprobleme auf, die darauf beruhen, dass für eines der Items in der Gruppe der *Kinder ohne Erfahrung mit Instrumentalunterricht* negative Residualvarianzen geschätzt werden. Bei dem Item handelt es sich um einen Abgleich einer vorliegenden Notation mit einem Instrumentalstück, welches in der Testsituation für die gesamte Klasse über Lautsprecher abgespielt wird. Die latent geschätzten Mittelwertdifferenzen der musikalischen Fähigkeiten sind in Tabelle 2 dokumentiert.

Tabelle 2: Geschätzte latente Mittelwertdifferenzen im Musiktest der Studie I (ProBiNi).

	M	*SE*	$p \leq$
Geschlecht			
Jungen (*n* = 234)	0.00		
Mädchen (*n* = 256)	0.10	0.08	n.s.
Übungszeit			
keine (*n* = 212)	0.00		
bis zu einer h/Woche (*n* = 133)	0.31	0.14	0.05
mehr als eine h/Woche (*n* = 87)	0.30	0.14	0.05
KFT			
niedrig (*n* = 162)	0.00		
mittel (*n* = 161)	0.28	0.13	0.05
hoch (*n* = 171)	0.38	0.17	0.05

Anmerkung: n.s. = nicht signifikant.

Es liegen keine signifikanten Unterschiede hinsichtlich der geschätzten latenten Musikkompetenz zwischen Jungen und Mädchen vor. Hinsichtlich der Übungszeit erreichen Schülerinnen und Schüler, die durchschnittlich bis zu einer Stunde die Woche auf ihrem Instrument üben, signifikant höhere Werte als Schülerinnen und Schüler, die nie üben (eingeschlossen diejenigen, die überhaupt kein Instrument spielen). Auch Schülerinnen und Schüler, die durchschnittlich mehr als eine Stunde pro Woche üben, haben signifikant höhere Werte als Schülerinnen und Schüler, die nie üben. Für die Differenzierungsvariable kognitive Grundfähigkeiten (KFT, N2) zeigt sich ebenfalls, dass Schülerinnen und Schüler mit vergleichsweise höheren kognitiven Grundfähigkeiten höhere Werte im Musiktest aufweisen als Schülerinnen und Schüler mit geringen kognitiven Grundfähigkeiten.

Fragestellung 1: Testung auf Messinvarianz in Studie II (Harnischmacher & Knigge, 2017)

Wie bereits für Studie I (ProBiNi) soll nun auch für den Datensatz aus Studie II (Harnischmacher & Knigge, 2017) überprüft werden, ob der eingesetzte Musiktest (KoMus_short) messinvariant bezüglich der zuvor identifizierten Ungleichheitsdimensionen ist. Zunächst wird auch hier ein Basismodell berechnet, in dem alle Parameter über die Gruppen frei variieren. Die Fitwerte können

auch hier als gut interpretiert werden ($X^2 = 351.17$; $df = 275$; $p = 0.001$; $RMSEA = 0.02$; $CFI = 0.96$; $TLI = 0.96$) und legen eine Annahme von konfiguraler Messinvarianz nahe. Anschließend wird erneut in Mplus ein Modell zur Überprüfung skalarer Messinvarianz mit fixierten Faktorladungen und Schwellenwerten im Rahmen von Mehrgruppenmodellen geschätzt. Weicht das Modell mit fixierten Parametern signifikant vom Modell konfiguraler Messinvarianz ab, dann liegt ein Verstoß gegen die Annahme von Messinvarianz für das latente Konstrukt für die betrachteten Gruppen vor. Die Ergebnisse dieser Analysen sind in Tabelle 3 dokumentiert.

Tabelle 3: Überprüfung von Messinvarianz im Rahmen von Mehrgruppenmodellen in Studie II (Harnischmacher & Knigge).

	X^2	*df*	X^2/df	$p \leq$	*RMSEA*	*CFI*	*TLI*	$X^{2\text{-Difference}}$	*df*	$p \leq$
Geschlecht										
Konfigural	605.15	550	1.10	n.s.	0.02	0.97	0.97			
Skalar	664.65	573	1.16	0.01	0.02	0.95	0.95	51.67	23	0.001
Instrumental-/Gesangsunterricht										
Konfigural	610.63	550	1.11	0.05	0.02	0.96	0.96			
Skalar	682.39	573	1.19	0.001	0.02	0.93	0.93	57.38	23	0.001
Sprachgebrauch										
Konfigural	602.44	550	0.68	n.s.	0.02	0.97	0.97			
Skalar	651.06	573	1.43	0.05	0.02	0.95	0.95	44.32	23	0.01

Anmerkung: grau hinterlegte Felder kennzeichnen einen Verstoß gegen Messinvarianz; n.s. = nicht signifikant.

Für den Musiktest (KoMus_short) zeigen sich für die Gruppierungsvariablen *Geschlecht, Erfahrung mit Instrumental-/Gesangsunterricht* sowie für den *Sprachgebrauch zu Hause* Verstöße gegen die Annahme der skalaren faktoriellen Messinvarianz. Dementsprechend sind latente Mittelwertvergleiche zwischen den Gruppen, beispielsweise zwischen Mädchen und Jungen, nicht zulässig.

Da für alle drei Gruppierungsvariablen Verstöße gegen die Annahme von skalarer Messinvarianz beobachtet wurden, folgen tiefergehende Analysen.

7. Datenanalyse II: Überprüfung von Item-Bias anhand von DIF-Analysen

7.1 Methodisches Vorgehen

Im Rahmen von DIF-Analysen soll der Frage nachgegangen werden, welche Items des in der Studie II (Harnischmacher & Knigge, 2017) eingesetzten Musiktests zu Verzerrungen führen, also einen Bias aufweisen. Zu diesem Zweck wurden DIF-Analysen mit der Software ConQuest (Version 2.0; Wu, Adams, Wilson & Haldane, 2007) durchgeführt. Aus methodischer Sicht basiert eine DIF-Analyse zunächst auf der getrennten Raschskalierung eines Tests für die interessierenden Subgruppen (z. B. Jungen und Mädchen). Dadurch erhält man für jedes Item eines Tests jeweils zwei (oder in Abhängigkeit von der Gruppenvariable auch mehrere) getrennt geschätzte Itemparameter. Der anschließend durchzuführende Vergleich der resultierenden Itemparameter kann grafisch anhand eines Streudiagramms veranschaulicht werden. Hierbei werden die Itemparameter der beiden Gruppen gegeneinander aufgetragen, wodurch jedes Item durch einen Punkt dargestellt wird, der idealerweise auf der Geraden (oder zumindest sehr nahe an dieser) liegen sollte. Wenngleich grafische DIF-Analysen sehr anschaulich sind, so ist es doch schwierig

auf dieser Grundlage zu entscheiden, ob ein „substantieller" DIF vorliegt. Dies ist vor allem deshalb schwer, da in der Testpraxis fast jedes Item einen gewissen Grad an DIF aufweisen wird (Wu & Adams, 2007, S. 71). Für die Entscheidung, ob ein Item einen substantiellen DIF aufweist, kann die Größe der Differenz der Itemschwierigkeiten festgelegt werden, ab der für ein Item von DIF ausgegangen wird. Außerdem ermöglicht ConQuest eine Signifikanztestung des vorliegenden DIF (Wu, Adams & Wilson, 1998). Tristán (2006) plädiert für die Einteilung des DIF-Ausmaßes in drei Kategorien: (1) bis 0.43 logits = vernachlässigbar, (2) zwischen 0.43 und 0.64 logits = gering bis moderat, (3) ab 0.64 logits = moderat bis groß (vgl. auch Jordan, 2014, S. 93f.). Wang (2000) schlägt eine etwas strengere Einteilung vor und spricht von einem substantiellen DIF bezüglich zweier Teilpopulationen bereits, wenn die Differenz der Itemschwierigkeiten signifikant ist und mindestens 0.50 logits beträgt. Diesem Vorschlag schließen wir uns an und bezeichnen im Folgenden Items als biased ab einer logit-Differenz von 0.50.

Abweichend von bisherigen Modellierungen der aus KoMus stammenden Items als Partial Credit Modell (vgl. Jordan et al., 2012; Knigge, 2010) wurden für diesen Beitrag alle Items dichotomisiert (richtig = 1, falsch = 0), sodass 25 Items in die Analysen übernommen werden konnten. Die Reliabilität des dichotomen Raschmodells beträgt .73. Anschließend wurden DIF-Analysen auf der Basis dichotomer Raschmodelle (1-PL) für die Gruppierungsvariablen *Geschlecht* (Junge/Mädchen), *Erfahrung mit Instrument-/Gesangsunterricht* (nein/ja) und dem *Sprachgebrauch zu Hause* (spricht immer deutsch/manchmal oder nie deutsch) geschätzt.

7.2 Ergebnisse

Aus Tabelle 4 wird ersichtlich, dass von den insgesamt 25 eingesetzten Items des KoMus_short-Tests jeweils maximal zwei Items pro relevanter Ungleichheitsdimension (s. Tab. 3) DIF aufweisen. Da ein Item sowohl in der Geschlechts- als auch der Musikpraxisdimension DIF aufweist, sind insgesamt also 4 von 27 Items betroffen.

Tabelle 4: Verteilung von Items mit DIF auf die verschiedenen Ungleichheitsdimensionen.

Konstrukt	DIF > 0.50 (Anzahl Items)
Geschlecht (weiblich / männlich)	2
Sprachgebrauch (nur Deutsch vs. Deutsch + weitere Sprache)	1
Musikpraxis (keine außerschulische Praxis / mind. 1 Jahr)	2

Eine inhaltliche Analyse der betroffenen Items ergibt ein uneinheitliches Bild: Einerseits sind bestimmte Interpretationen naheliegend, so beispielsweise bei der Betrachtung von Items mit DIF in der Dimension Sprachgebrauch. Hier werden Kinder, die zu Hause nicht immer Deutsch sprechen, von einem Item (D1(3)-10b-1/2) ‚begünstigt', das sich durch sehr wenig Text im Aufgabenstamm und den Antwortoptionen auszeichnet. Ähnlich naheliegend scheint die Interpretation für ein Item (D2-4b), bei dem ein Hörbeispiel musikalischen Fachausdrücken zugeordnet werden muss („forte" und „piano") – dass hier Kinder, die außerschulisch Instrumental- oder Gesangsunterricht erhalten, im Vorteil sind, verwundert kaum. Andererseits ist jedoch vor zu schnellen Hypothesenbildungen und Schlüssen Vorsicht geboten, denn werden auch die restlichen Items ohne DIF in die Überlegungen mit einbezogen, so wird deutlich, dass die identifizierten Itemcharakteristika zumindest

nicht alleinig für das gefundene DIF verantwortlich sein können. Denn es gibt im restlichen Itempool immer auch ähnliche Items, die etwa ebenfalls sehr wenig Text aufweisen und trotzdem zu keinem relativen Vorteil für Kinder, die zu Hause nicht immer deutsch sprechen, führen. Der Sachverhalt scheint also komplexer zu sein, als es die Hypothesen „wenig Aufgabentext begünstigt Kinder mit Migrationshintergrund“ oder „viel Fachterminologie begünstigt Instrumentalisten“ suggerieren. Vielmehr ist von einer Interaktion der verschiedenen Aufgabencharakteristika auszugehen, wie sie auch von Knigge (2010, Kap. 7) im Rahmen der regressionsanalytischen Identifikation von schwierigkeitsgenerierenden Aufgabenmerkmalen – jedoch ohne Berücksichtigung von Ungleichheitsdimensionen – gefunden wurde.

8. Diskussion und Ausblick

Im Rahmen des Beitrags wurden zentrale Methoden zur Überprüfung von Test- und Item-Bias vorgestellt. Bei der exemplarischen Anwendung auf zwei Datensätze aus musikpädagogischen Studien zeigte sich, dass eine entsprechende Überprüfung angezeigt ist, bevor Mittelwertvergleiche hinsichtlich verschiedener Subgruppen angestellt werden können. Denn obwohl die eingesetzten Testverfahren als etabliert anzusehen sind, traditionelle Gütekriterien (insbesondere die Reliabilität) erfüllt sind und auch probabilistische Verfahren zur Überprüfung der psychometrischen Qualität eingesetzt wurden, ergaben die Analysen des vorliegenden Beitrags sehr deutlich, dass für einzelne Personengruppen von einem Test- beziehungsweise Itembias auszugehen ist, die Tests also nicht für alle getesteten Personen die gleichen Eigenschaften aufweisen und entsprechend keine vergleichbaren Ergebnisse produzieren. Im vorliegenden Beitrag trifft dies auf die Ungleichheitsdimensionen Geschlecht, Sprachgebrauch zu Hause, Erfahrung mit Gesangs- und Instrumentalunterricht und den kognitiven Grundfähigkeiten zu, die basierend auf Befunden aus den Bereichen der Musikpädagogik, der Erziehungswissenschaft und der Psychologie identifiziert wurden. Im Anschluss an diese Erkenntnis sind nun verschiedene Strategien/Perspektiven denkbar:

- Bezüglich der Weiterverarbeitung der Daten (mit dem Ziel, Gruppenvergleiche mit bereits erhobenen Daten durchzuführen) bieten sich zunächst zwei Verfahrensweisen an. Die eine beinhaltet das Ausschließen derjenigen Items, die einen Bias aufweisen. Die Reduktion des Itempools mit dem Ziel der Erhöhung der Reliabilität des jeweiligen Messinstruments führt in der Regel zu homogeneren Items, die unter Umständen jedoch die Spezifizität und Besonderheiten von bestimmten Interventionen oder Treatments, aber auch besondere kulturelle Unterschiede zwischen den Vergleichsgruppen, unberücksichtigt lassen. So gehen eine höhere interne Konsistenz und Reliabilität häufig mit einer geringeren inhaltlichen Validität einher (Rippl & Seipel, 2008). Damit bietet sich durchaus ein weiteres Verfahren an, das die Modellierung der Items vorsieht, obgleich sie einen Bias aufweisen. Kann dieser Bias identifiziert werden, kann beispielsweise im Rahmen von CFAs die Modellierung von partieller Messinvarianz vorgenommen werden, sodass unter besonderen Bedingungen (Schulte et al., 2013; Meredith, 1993) die Durchführung fairer Vergleiche sichergestellt werden kann.
- DIF-Analysen sind darüber hinaus insbesondere in Bezug auf Testneuentwicklungen beziehungsweise Überarbeitungen eines bestehenden Testinstruments vielversprechend. Wie in

Abschnitt 7 gezeigt, ermöglichen sie die Hypothesenbildung bezüglich der Aufgabencharakteristika, die zu einem Item-Bias führen, und daran anschließend die Überarbeitung oder gezielte Entwicklung von Items mit spezifischen Inhalten/Eigenschaften. Deutlich wurde jedoch auch, dass die Analysemöglichkeiten auf dieser Ebene sehr stark vom Itempool und von der Personenstichprobe abhängen. So müssten optimalerweise verschiedene Itemeigenschaften systematisch variiert werden, um ihren Einfluss isolieren zu können und gleichzeitig relevante Personeneigenschaften in ausreichend großen Subgruppen der Stichprobe vertreten sein, sodass ein potentiell vorhandener Bias statistisch untersucht werden kann. Beides ist im Rahmen von Testentwicklungen/-überarbeitungen oftmals nur schwer sicherzustellen.

Die endgültige Entscheidung, wie mit einem Test beziehungsweise Item-Bias umzugehen ist, sollte im Einzelfall und mit Bezug zur Fragestellung und zu den erwarteten Konsequenzen getroffen werden. Darüber hinaus sind auch testökonomische Gegebenheiten entscheidend, wie etwa die Möglichkeit der sorgfältigen Überarbeitung und erneuten Administration der Testitems oder die Frage nach der Kontext- und Instruktionssensitivität des eingesetzten Tests (vgl. Naumann, Musow, Aichele, Hochweber & Hartig, 2019). Für die vorliegenden Befunde wäre eine systematische Überarbeitung von Items, die DIF aufweisen, für zukünftige Studien angezeigt. Dennoch können die verwendeten Erhebungsinstrumente genutzt werden, um Zusammenhänge mit anderen zentralen Merkmalen, beispielsweise im Kontext von Strukturgleichungsmodellen, zu identifizieren. Falls partielle Messinvarianz bestätigt werden kann, können zudem, unter gewissen Vorannahmen, Mittelwertvergleiche umgesetzt werden, was für die vorliegenden Daten noch geprüft werden müsste.

Mit diesem Beitrag konnte gezeigt werden, dass Testfairness ein zentrales Gütemerkmal von musikbezogenen Interventions- sowie Evaluationsstudien ist, wobei systematische Verzerrungen auch dort auftreten können, wo sie nicht erwartbar sind und dementsprechend auch nicht immer identifiziert werden. Im vorliegenden Beitrag wurde insbesondere die statistische Komponente in den Blick genommen, der sozialen Komponente (*consequential aspect*; Kline, 2013; Messick, 1998) kommt jedoch zusätzlich eine wichtige Bedeutung zu. So liefern Evaluationsstudien zu musikbezogenen Interventionen und Treatments, wie sie beispielsweise im Kontext des Programms „Jedem Kind ein Instrument“ (JeKi) vorgenommen wurden, Ergebnisse, auf deren Basis eine Weiterführung, Erweiterung oder gar Einstellung von Programmen entschieden wird. Unser Beitrag verweist auf die Notwendigkeit von differenziellen Analysen, um valide und reliable Aussagen über die Wirksamkeit des Treatments oder der Intervention in verschiedenen Personengruppen zu ermöglichen (Leutner, 2010). Aus wissenschaftlicher Perspektive sei hier einmal mehr auf eine fundierte methodische Ausbildung von Nachwuchswissenschaftlerinnen und Nachwuchswissenschaftlern hingewiesen. Hier sollte zumindest eine tiefergehende Reflexion von und methodenkritische Perspektive auf Evaluationsergebnisse(n) geschult werden. Nicht zuletzt sei auch noch erwähnt, dass eine voreilige Rückmeldung von Evaluationsergebnissen an etwaige Auftraggeber ohne Berücksichtigung differenzieller Effekte und ohne die notwendige Absicherung von Messäquivalenz problematisch sein kann. Hier stehen eine zeitnahe Rückmeldung und Verwertung der Ergebnisse einer adäquaten und sorgfältigen Analyse der Daten diametral gegenüber, sodass im Einzelfall über das Abwägen von Konsequenzen entschieden werden muss.

Literatur

Amelang, M. & Schmidt-Atzert, L. (Hrsg.) (2006). *Psychologische Diagnostik und Intervention* (4. Aufl.). Berlin: Springer.

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland, & H. Wainer (Hrsg.), *Differential item functioning* (S. 3–23). Hillsdale: Lawrence Erlbaum Associates.

Camilli, G. (2013). Ongoing issues in test fairness. *Educational Research and Evaluation, 19,* 104–120. doi: 10.1080/13803611.2013.767602

Cheung, G. W. & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9,* 233–255. doi: 10.1207/S15328007SEM0902_5

Fiedler, D. & Spychiger, M. (2017). Measuring "musical self-concept" throughout the years of adolescence with MUSCI_youth: Validation and adjustment of the Musical Self-Concept Inquiry (MUSCI) by investigating samples of students at secondary education schools. *Psychomusicology: Music, Mind, and Brain, 27*(3), 167–179. doi: 10.1037/pmu0000180

Fiege, C., Reuther, F. & Nachtigall, C. (2011). Faire Vergleiche? – Berücksichtigung von Kontextbedingungen des Lernens beim Vergleich von Testergebnissen aus deutschen Vergleichsarbeiten. *Zeitschrift für Bildungsforschung, 1,* 133–149. doi: 10.1007/s35834-011-0009-x

Gordon, E. (1991). *Iowa Test of Music Literacy*. Chicago: GIA Publications.

Harnischmacher, C. & Knigge, J. (2017). Motivation, Musizierpraxis und Musikinteresse in der Familie als Prädiktoren der Kompetenz „Musik wahrnehmen und kontextualisieren" und des Kompetenzerlebens im Musikunterricht. *Beiträge empirischer Musikpädagogik*, 8, 1–21. https://www.b-em.info/index.php/ojs/article/view/136

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

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

Heckhausen, H. (1974). *Leistung und Chancengleichheit* (Motivationsforschung, Bd. 2). Göttingen: Hogrefe.

Heller, K.A. & Perleth, C. (2000). *KFT 4-12+R - Kognitiver Fähigkeits-Test für 4. bis 12. Klassen, Revision*. Göttingen: Beltz.

Hu, L.-T. & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. doi: 10.1080/10705519909540118

Jöreskog, K. G. (1970). A general method for analysis of covariance structures. *Biometrika*, *57*(2), 239–251. doi: 10.1093/biomet/57.2.239

Jordan, A.-K. (2014). *Empirische Validierung eines Kompetenzmodells für das Fach Musik. Teilkompetenz „Musik wahrnehmen und kontextualisieren"*. Münster: Waxmann.

Jordan, A.-K., Knigge, J., Lehmann, A. C., Niessen, A. & Lehmann-Wermser, A. (2012). Entwicklung und Validierung eines Kompetenzmodells im Fach Musik – Wahrnehmen und Kontextualisieren von Musik. *Zeitschrift für Pädagogik, 58*, 500–521. urn: nbn:de:0111-pedocs-103923

Kane, M. & Bridgeman, B. (2017). Research on validity theory and practice at ETS. In R. E. Bennett, & M. von Davier (Hrsg.), *Advancing human assessment. The methodological, psycho-*

logical and policy contributions of ETS (Methodology of educational measurement and assessment, vol. 12, S. 489–552). Cham: Springer International Publishing. doi: 10.1007/978-3-319-58689-2_16

Kline, R. B. (2013). Assessing statistical aspects of test fairness with structural equation modelling. *Educational Research and Evaluation, 19,* 204–222. doi: 10.1080/13803611.2013.767624

Knigge, J. (2010). *Modellbasierte Entwicklung und Analyse von Testaufgaben zur Erfassung der Kompetenz „Musik wahrnehmen und kontextualisieren“*. Dissertation, Universität Bremen. http://nbn-resolving.de/urn:nbn:de:gbv:46-diss000120066

Kim, E. S. & Yoon, M. (2011). Testing measurement invariance: A comparison of multiple-group categorical CFA and IRT. *Structural Equation Modeling: A Multidisciplinary Journal, 18,* 212–228. doi: 10.1080/10705511.2011.557337

Law, L. N. & Zentner, M. (2012). Assessing musical abilities objectively: Construction and validation of the profile of music perception skills. *PloS ONE, 7*(12). doi: 10.1371/journal.pone.0052508

Leutner, D. (2010). Perspektiven pädagogischer Interventionsforschung. In Hascher, T. & B. Schmitz (Hrsg.), *Pädagogische Interventionsforschung. Theoretische Grundlagen und empirisches Handlungswissen* (S. 63–72). Weinheim: Juventa.

Louven, C. (2016). Hargreaves' "open-earedness": A critical discussion and new approach on the concept of musical tolerance and curiosity. *Musicae Scientiae*, *20*(3), 235–247. doi: 10.1177/1029864916633264

Lüdtke, O., Robitzsch, A., Trautwein, U. & Köller, O. (2007). Umgang mit fehlenden Werten in der psychologischen Forschung: Probleme und Lösungen. *Psychologische Rundschau, 58,* 103–117. doi: 10.1026/0033-3042.58.2.103

Meade, A. W. & Lautenschlager, Gary, J. (2004). *Same Question, Different Answers: CFA and Two IRT Approaches to Measurement Invariance*. Symposium at the 19th Annual Conference of the Society for Industrial and Organizational Psychology, Chicago, IL. https://pdfs.semanticscholar.org/0685/2710c39ae93bfa090fc553d06e01b9751c46.pdf

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrica, 58, 525–543*. doi: 10.1007/BF02294825

Messick, S. (1998). Test validity: A matter of consequence. *Social Indicators Research, 45,* 35–44. doi: 10.1023/A:1006964925094

Millsap, R. E. & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research*, *39*(3), 479–515. doi: 10.1207/S15327906MBR3903_4

Muthén, L. K. & Muthén, B. O. (1998–2012). *Mplus User`s Guide* (7. Aufl.). Los Angeles, CA: Muthén & Muthén.

Müllensiefen, D., Gingras, B., Musil, J. & Stewart, L. (2014). The Musicality of non-musicians: An index for assessing musical sophistication in the general publication. *PLoS ONE, 9*(2). doi: 10.1371/journal.pone.0089642

Naumann, A., Musow, S., Aichele, C., Hochweber, J. & Hartig, J. (2019). Instruktionssensitivität von Tests und Items. *Zeitschrift für Erziehungswissenschaft*, 22(1), 181–202. doi: 10.1007/s11618-018-0832-0

Niessen, A. & Knigge, J. (2018). Empirische Forschung in der Musikpädagogik. In M. Dartsch, J. Knigge, A. Niessen, F. Platz, & C. Stöger (Hrsg.), *Handbuch Musikpädagogik* (S. 451–456). Münster: Waxmann.

Nonte, S. (2012). Die Überprüfung von geschlechtsbezogener Messinvarianz des Fähigkeitsselbstkonzepts von Grundschülern in der Schuleingangsphase. *Empirische Pädagogik, 26,* 478–503. doi: 10.1007/s35834-013-0062-8

Nonte, S. (2013). Herausforderungen und Probleme bei der Entwicklung eines Instruments zur Selbsteinschätzung musikalischer Fähigkeiten im Grundschulalter. *Beiträge empirischer Musikpädagogik, 4,* 1–30. https://www.b-em.info/index.php/ojs/article/view/94

Oswald, W. D. & Roth, E. (2016). *Zahlen-Verbindungs-Test (ZVT)* (3. Aufl.). Göttingen: Hogrefe Verlag für Psychologie.

Reynolds, C. R. & Suzuki, L. A. (2013). Bias in psychological assessment. In I. B. Weiner (Hrsg.), *Handbook of Psychology* (S. 85). Hoboken: J. Wiley. doi: 10.1002/9781118133880.hop210004

Rippl, S. & Seipel, C. (2008). *Methoden kulturvergleichender Sozialforschung. Eine Einführung* (Lehrbuch). Wiesbaden: VS Verlag für Sozialwissenschaften.

Sass, D. A., Schmitt, T. A. & Marsh, H. W. (2014). Evaluating model fit with ordered categorical data within a measurement invariance framework: A comparison of estimators. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(2), 167–180. doi: 10.1080/10705511.2014.882658

Schulte, K., Nonte, S. & Schwippert, K. (2013). Die Überprüfung von Messinvarianz in international vergleichenden Schulleistungsstudien am Beispiel der Studie PIRLS. *Zeitschrift für Bildungsforschung, 3,* 99–118. doi: 0.1007/s35834-013-0062-8

Schwab, S. & Helm, C. (2015). Überprüfung von Messinvarianz mittels CFA und DIF-Analysen. *Empirische Sonderpädagogik, 7,* 175–193. doi: 10.1007/s35834-013-0062-8

Schwabe, F. (2014). *Leseleistungsdifferenzen bei spezifischen Schülersubgruppen: DIF-Analysen von Large-Scale Assessments*. Dissertation. Dortmund: Technische Universität Dortmund.

Steenkamp, J.-B. E. M. & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research, 25,* 78–90. doi: 10.1086/209528

Stubbe, T. C. (2011). How do different versions of a test instrument function in a single language? A DIF analysis of the PIRLS 2006 German assessments. *Educational Research and Evaluation, 17,* 465–481. doi: 10.1080/13803611.2011.630560

Stubbe, T. C., Nonte, S., Haas, M. & Krieg, M. (Hrsg.). (i. Vorb.). *Musik- und MINT-Profile an Niedersächsischen Gymnasien und integrierten Gesamtschulen. Ergebnisse der Studie ProBiNi.*

Walther, G., Schwippert, K., Lankes, E. -M. & Stubbe, T. C. (2008). Können Mädchen doch rechnen? Vertiefende Analysen zu Geschlechtsdifferenzen im Bereich Mathematik auf Basis der Internationalen Grundschul-Lese-Untersuchung IGLU. *Zeitschrift für Erziehungswissenschaft, 11,* 30–46. doi: 10.1007/s11618-008-0002-x

Weiber, R., & Mühlhaus, D. (2010). *Strukturgleichungsmodellierung. Eine anwendungsorientierte Einführung in die Kausalanalyse mit Hilfe von AMOS, SmartPLS und SPSS*. Berlin: Springer.

Wu, M. & Adams, R. (2007). *Applying the rasch model to psycho-social measurement: A practical approach,* Educational Measurement Solutions, Melbourne. Abgerufen von https://media.metrik.de/uploads/incoming/pub/Literatur/von%20Winfried/RaschMeasurement_Complete.pdf

Wu, M., Adams, R. & Wilson, M. (1998). *ConQuest: Generalised item response modelling software*. Melbourne: Australian Council for Educational Research.

Wu, M., Adams, R., Wilson, M. & Haldane, S. (2007). *ACER ConQuest: Version 2.0. Generalised Item Response Modelling Software*. Camberwell, Victoria: Acer.

Sonja Nonte
Georg-August-Universität Göttingen
Institut für Erziehungswissenschaft
Waldweg 26
37073 Göttingen (Deutschland)
E-Mail: snonte@uni-goettingen.de

Jens Knigge
Nord University
Faculty of Education and Arts, Campus Levanger
Røstad, Høgskoleveien 27
7600 Levanger (Norwegen)
E-Mail: jens.knigge@nord.no

Tobias C. Stubbe
Georg-August-Universität Göttingen
Institut für Erziehungswissenschaft
Waldweg 26
37073 Göttingen (Deutschland)
E-Mail: tstubbe@uni-goettingen.de

Elektronische Version / Electronic Version:
https://www.b-em.info/index.php/ojs/article/view/173
URN: urn:nbn:de:101:1-2018082855