Latente Variablenmodelle in der empirischen Bildungsforschung
–
die Schärfe und Struktur der Schatten an der Wand

Kumulative Dissertation zur Erlangung des akademischen Grades eines Doktors der Philosophie (Dr. phil.)

an der Fakultät Erziehungswissenschaft und Soziologie
der Technischen Universität Dortmund

vorgelegt von

Dipl.-Päd. Michael Schurig

bei Prof. Dr. Wilfried Bos
und Prof. Dr. Tobias C. Stubbe

Dortmund 2017

EIDESSTATTLICHE ERKLÄRUNG

Hiermit versichere ich schriftlich und eidesstattlich gemäß § 11 Abs. 2 PromO v. 08.02.2011/08.05.2013:

1. Die von mir vorgelegte Dissertation ist selbstständig verfasst und alle in Anspruch genommenen Quellen und Hilfen sind in der Dissertation vermerkt worden.
2. Die von mir eingereichte Dissertation ist weder in der gegenwärtigen noch in einer anderen Fassung an der Technischen Universität Dortmund oder an einer anderen Hochschule im Zusammenhang mit einer staatlichen oder akademischen Prüfung vorgelegt worden.

Ort, Datum	Michael Schurig

3. Weiterhin erkläre ich schriftlich und eidesstattlich, dass mir der „Ratgeber zur Verhinderung von Plagiaten“ und die „Regeln guter wissenschaftlicher Praxis der Technischen Universität Dortmund“ bekannt und von mir in der vorgelegten Dissertation befolgt worden sind.

Ort, Datum	Michael Schurig

Mein Dank Euch allen,
die Ihr mich auf diesem Weg begleitet habt.
Ich hoffe, Ihr wisst,
was Ihr mir bedeutet.

Inhaltsverzeichnis

ABBILDUNGSVERZEICHNIS

TABELLENVERZEICHNIS

1. EINFÜHRUNG

In Platons Phaidon wird herausgestellt, dass für die Erklärungen von Sachverhalten eine einzelne sinnliche Beobachtung nicht hinreichend sei, erst eine indirekte, begriffliche Reflexion liefere echtes Wissen (Szlezák, 2003). Im Höhlengleichnis der Politeia (Buch VII) wird dies sinnbildlich umrissen. Es seien Menschen ihr ganzes Leben an eine Wand gekettet, sodass diese nur eine Höhlenwand und nicht die Höhle hinter sich sehen können. Die Gefangenen sehen nur das Licht eines Feuers hinter sich an der Wand, nicht aber das Feuer selbst. Von Trägern, deren Schatten nicht an die Wand geworfen werden, werden Gegenstände zwischen dem Feuer und der Wand, aber hinter den Gefangenen hin- und hergetragen. Die Gegenstände können von den Gefangenen als Schatten wahrgenommen werden. Manche der Träger reden miteinander, andere schweigen, da die Höhle Echos wirft, wirkt es aber, als ob die Gegenstände sprächen. Die Wirklichkeit jener Gefangenen stellt sich also durch die Betrachtung der Schatten dar (Loewenthal, 2004). Platon bemängelt die Deutung der Wirklichkeit durch die Gefangenen in der Höhle als eine „Wissenschaft der Schatten“, bei der die Weisheit der Gefangenen sich vor allem durch die scharfe Beobachtung der Schatten durch Einzelne und „das beste Gedächtnis daran, was vor, nach und mit [den Schatten] zu kommen pflegte, und das geschickteste Vorhersagen des künftig Kommenden […]“ (Loewenthal, 2004, S. 251) ausdrückt. Ein Schritt aus der Höhle heraus ist nicht empirisch, sondern theoriebasiert, ein Verlassen des Platzes in der Höhle und eine Hinwendung zum Feuer oder sogar zu der Sonne außerhalb der Höhle (eine Allegorie für das Gute) könne nur in der intelligibelen Welt, also einer Welt, die nur über den Intellekt erfasst werden kann, vorgenommen werden (Szlezák, 2003, S. 35f). Während das Gleichnis vor allem die Notwendigkeit eines philosophischen Bildungswegs als Befreiungsprozess umreißt, enthält es auch eine Kritik der empirischen Wissenschaft. Die Beobachtung wird klar von der Erklärung abgetrennt und mahnt zur realistischen Beschränkung des empirisch erklärbaren Gegenstandsbereichs sowie einer Kritik an einer unzureichenden Reflexion von Einschränkungen der Empirie.

Im Bildungsbereich zeichnet sich eine Entwicklung in Richtung einer daten- und evidenzgestützten Politik ab, wie sie auch in anderen Bereichen, zum Beispiel der Gesundheit oder der Wirtschaft bereits vor längerer Zeit begonnen hat (Bromme, Prenzel & Jäger, 2014, S. 4). Forschung soll verstärkt zur Versachlichung von Debatten beitragen. Allerdings setzt dies methodisch angemessene Forschung, Forschungssynthesen und ein sozialwissenschaftlich fundiertes Verständnis der öffentlichen Debatten und Auffassungen zu einer Daten- und Evidenzbasierung voraus (Bromme & Prenzel, 2014, S. 1). Dies umfasst die Reflexion der wissenschaftlichen Methoden.

Wissenschaftliche Methoden sind Werkzeuge, um Erkenntnisziele zu erreichen. Eine zentrale Bedeutung haben Methoden zur Formulierung und Prüfung wissenschaftlicher Aussagen. Die Verwendung latenter Variablen ist dabei in den Bildungswissenschaften und auch in anderen Disziplinen, beispielsweise der Medizin oder der Ökonomie, üblich und weit verbreitet (Glymour, Scheines & Spirtes, 2014), da die Verwendung latenter Variablen ein inhaltlich generalisierbareres Schließen erlauben als manifeste Variablen (z.B.Little, Lindenberger & Nesselroade, 1999; Mislevy & Haertel, 2006). Latente Variablen erlauben Rückschlüsse auf

Kompetenzen (vgl. Klieme & Hartig, 2008), wahrscheinlichkeitsbasierte Klassenzugehörigkeiten, die Zusammenhangsstrukturen hypothetischer Konstrukte und insbesondere den Fehlergehalt, respektive die Präzision von Messungen. Als latente Variablen werden in den Human- und Sozialwissenschaften weitläufig Größen verstanden, welche nicht direkt beobachtet werden (können) und erst durch eine formal-mathematische Operationalisierung von manifesten, also sichtbaren, Indikatoren beobachtbar gemacht werden (z. B. Bollen, 2002; Sedlmeier & Renkewitz, 2013; Skrondal & Rabe-Hesketh, 2004). Technisch werden latente Variablen als Zufallsvariablen begriffen, deren realisierte Werte verborgen sind. Sie werden als mathematische Größen abgeleitet, indem ein statistisches Modell die latenten Variablen mit manifesten beobachteten Indikatoren verbindet. Es erfolgt eine formale Repräsentation objektivierter Realität durch eine mathematische Struktur. Dies wird über Verbindungen der Struktur auf die objektivierte Welt, die Entitäten im Modell, die Beschreibung des wissenschaftlichen Vorgehens, das gemeinsame Verständnis wissenschaftlichen Vorgehens und ein gemeinsames Sprachverständnis befördert (Dempster, 1998, S. 252). Je nach Wissenschaftsdisziplin werden die Modelle unter verschiedenen Namen geführt, beispielsweise *random effects*, *common factor*, *latent classes*, *underlying Variables* oder auch *frailties* (Skrondal & Rabe-Hesketh, 2007, S. 712).[1]

Damit substanzwissenschaftliche Schlüsse mittels latenter Variablen möglich werden, müssen diese technischen Entitäten mit theoretischen Konstrukten und Annahmen begründet werden. Erst diese Verbindung legitimiert die Verwendung in erläuternder Funktion und zur Prüfung theoretischer Annahmen. Ungeklärt ist aber der theoretische Status von den latenten Variablen selbst, also wie diese definiert sind. Sollte angenommen werden, dass latente Variablen Repräsentationen realer Entitäten sind, oder etwa nützliche Erfindungen (z. B. Borsboom, Mellenbergh & van Heerden, 2003)? Ist beispielsweise eine Definition als unbeobachtbares theoretisches Konstrukt vor dem Hintergrund sich weiterentwickelnder Messtechnik hinreichend, wenn also ein Konstrukt zukünftig messbar gemacht wird?

Außerdem müssen diese Methoden, damit sie als wissenschaftlich gelten können, überdisziplinäre Gütekriterien erfüllen. Hier gelten die Haupt- und Nebengütekriterien von Tests und Messungen (vgl. Lienert & Raatz, 1998). Während die Gütekriterien konzeptionell leicht verständlich erscheinen, ist die Prüfung dieser, insbesondere für Modelle mit latenten Variablen, nicht immer einfach und untrennbar mit der Spezifikation, Identifikation und Evaluation der Modelle verbunden. Dabei hat es sich stark eingebürgert, eine Mehrzahl von Testgütekriterien heranzuziehen, diese auf der Basis von Normen und Faustregeln zu betrachten und Tests und Messungen dann als annehmbar oder unannehmbar zu betrachten (Newton & Shaw, 2013). Dies umfasst beispielsweise „erfolgreiche“ Prüfungen auf Signifikanz der

[1] Auch wenn für verschiedene Modelltypen deutsche Benennungen existieren werden in dieser Arbeit zumeist englische Bezeichnungen verwendet, da die englischsprachige Fachliteratur reichhaltiger ist und die Übersetzungen nicht immer einheitlich vorgenommen wurden.

Modellanpassung (z. B. einen Wert $p \geq .05$ bei einem χ^2-Test auf Modellanpassung[2]) ebenso wie erwünschte Grade von interner Konsistenz (z. B. einen Cronbachs-Alpha von >.7) oder ein angemessen erscheinender relativer Modellanpassungsindex (z. B. Werte von $> .95$ bei einer maximalen Anpassungsgüte von 1). Eine multikriteriale Betrachtungsweise ist zu begrüßen, da alle Werte auf distinkte Weise bestimmt werden und verschiedene Gewichtungen bei der Modellbewertung vornehmen, sodass ein einzelner Wert nicht in der Lage sein kann, in jeder Test- oder Messsituation angemessen scharf zwischen Annahme und Zurückweisung zu trennen. Dies hat aber den Nachteil, dass das konkrete Vorgehen bei der Zusammenstellung der Tests, beim Einsatz diverser Prüfmechanismen und bei der Auswertung von Test- und Messinstrumenten häufig nicht hinreichend reflektiert wird und Fehlschlüsse zu erwarten sind (Döring & Bortz, 2016, S. 442). Dies umfasst auch und insbesondere die Gewichtung der Indizes zueinander in Abhängigkeit zu der Test- oder Messsituation und der Datengrundlage.

Es existiert bereits methodologische Literatur, welche viele der hier vorgestellten Annahmen, Modelltypen und Prüfmechanismen einzeln oder auch in einem generalisierten Rahmen zusammenträgt (z. B. Rost, 1996). Aber dies geschieht vornehmlich einer statistischer Perspektive, welche auf die formalen Definitionen der latenten Variablen fokussiert, und bleibt anteilig auf einzelne Modelltypen (z.B. Reinecke, 2014) oder Computeranwendungen (z.B. Muthén & Muthén, 1998-2015) begrenzt. Diese Arbeit trägt daher theoretische Grundannahmen zu latenten Variablen in verschiedenen gängigen Ableitungsformen zusammen und verknüpft diese mit den Grundannahmen der Mess- und Testtheorie sowie den Bedingungen zur Verwendung quantitativ empirischer Ergebnisse als Evidenzform für die Beobachtung stochastisch kausaler Zusammenhänge. Hierzu werden die Hauptgütekriterien von Tests und Messungen sowie deren Bedeutung für latente Variablenmodelle diskutiert und mit Strategien der Modellevaluation verknüpft. In dieser Arbeit wird dabei außerdem insbesondere die Versprachlichung formaler Definitionen und Schreibweisen angestrebt, um die impliziten Charakteristika der Verfahren, welche in der Anwendung eine gewichtige Relevanz haben, herauszustellen und so einen reflektierten und passgenauen Einsatz dieser statistischen Modelltypen zu vereinfachen. Damit richtet sich diese Arbeit insbesondere an Anwender latenter Variablenmodelle in einem geisteswissenschaftlichen Kontext.

Dafür wird zuerst die Bedeutung statistischer Modelle mit latenten Variablen in den Bildungswissenschaften ausgeführt (Kap. 2) und in der Folge werden Theorien zur Definition latenter Variablen vorgestellt (Kap. 3). Es wird argumentiert, dass die verbreiteten Alltagsdefinitionen nicht ausreichend sind, und es wird ein definitorischer Rahmen vorgestellt, der sowohl eine theoretische als auch formale Anknüpfung erlaubt. Zudem werden verschiedene zentrale Modelltypen, deren Eigenschaften sowie eine generalisierte Betrachtungsweise kurz umrissen (Kap.3.1). Die Basis für die Verknüpfung statistischer Modelle und deren theoretischen Annahmen bilden die Mess- und Testtheorie. Die impliziten Annahmen und deren Implikationen werden in Kapitel 3.2 herausgestellt. Sofern eine theoretische Verknüpfung und eine

[2] Die Logik des χ^2-Tests auf Modellanpassung ist entgegen üblichen χ^2-Tests umgekehrt. Es wird also in der Nullhypothese angenommen, dass die Daten gegenüber einer angenommenen Modellspezifikation konsistent sind. Damit eine gute Anpassung des Modells nachgewiesen werden kann, muss der Test also ein nicht-signifikantes Ergebnis aufweisen.

Modellbildung gelungen ist, können Evidenzien angesammelt werden, über die auf vorläufige stochastische Kausalität geschlossen werden kann (Kap. 3.3). Zur Bewertung vorliegender Evidenz sind wissenschaftliche Normen und Standards von Bedeutung. Dazu werden die Hauptgütekriterien wissenschaftlicher Tests und Messungen (Kap. 4) besprochen und deren Anwendbarkeit und Prüfbarkeit für latente Variablenmodelle ausgeführt. In der Folge werden Strategien der Modellidentifikation und -evaluation sowie implizite Anwendungsprobleme und Anwendungslösungen mit einzelnen Strategien zusammengefasst (Kap. 5). Ergänzend werden Einschätzungen zu Strategien der Modellevaluation gegeben und ein Raster von Modellgütebeurteilungen vorgestellt (Kap. 5.3). Als Rahmung für verschiedene Artikel, die auf die substanzwissenschaftliche Anwendung von latenten Variablenmodellen fokussieren (vgl. Anhang 2), werden in dieser Arbeit die gemeinsamen Grundlegungen zu diesen zusammengetragen und expliziert, während die Artikel exemplarisch verwendet werden, um die praktische Anwendung von Prinzipien zu verdeutlichen.

2. LATENTE VARIABLENMODELLE IN DEN BILDUNGSWISSENSCHAFTEN

Um die Bedeutung von Modellen mit latenten Variablen für die Bildungsforschung einordnen zu können, ist ein überblicksartiger Rückblick in die Geschichte der Bildungsforschung hilfreich, da sich das Spektrum dieser Disziplin in den vergangenen 50 Jahren zunehmend erweitert hat (Edelmann, Schmidt & Tippelt, 2012, S. 45). In den 1960er-Jahren wurde vor dem Hintergrund des „Sputnik Schocks“, welcher den technologischen Vergleich mit dem Ostblock in die Bildungsdebatte einbrachte, die Frage nach der tatsächlichen Effizienz von Bildungseinrichtungen unter dem Schlagwort der deutschen Bildungskatastrophe (Picht, 1964) laut. Generell wurde eine stärkere Rolle der empirischen Sozialforschung in der Erziehungswissenschaft gefordert, welche parallel zu der Hermeneutik Idee und Wirklichkeit abgleichen sollte (Roth, 1963, S. 117). Adorno betonte darüber hinaus die hervorzuhebende Bedeutung der Trennung von Überlegungen zu Normen und Zielen der Erziehung und Bildung sowie der empirischen Tatsachenforschung, wobei Norm- und Zielfragen selbstverständlich auch zu bearbeiten seien (Adorno et al., 1993). Brezinka (1975) verschärfte diese Kritik unter Rückbezug auf den kritischen Rationalismus Karl Poppers (1994) und verlangte die Trennung von pädagogischem Verstehen und erziehungswissenschaftlichem Erkennen. Den Empfehlungen des Deutschen Bildungsrats (1974, S. 16) nach hat die Bildungsforschung den Auftrag, Voraussetzungen und Möglichkeiten von Bildungs- und Erziehungsprozessen im institutionellen und gesellschaftlichen Kontext zu untersuchen, Lehr- Lernprozesse in schulischen und außerschulischen Bereichen zu analysieren und nichtinstitutionalisierte Sozialisationsprozesse zu thematisieren, was bis heute Bestand hat (Tippelt, 1998, S. 240).

Insgesamt verharrte die pädagogische empirische Forschung trotzdem weitgehend in der Schattenzone, und kritische Rückmeldungen an die Politik in der notwenigen Qualität und Intensität (Buchhaas-Birkholz, 2009, S. 27) blieben ebenso lange aus wie die Rezeption von Ergebnissen (Cortina, Baumert, Leschinsky, Mayer & Trommer, 2008, S. 45). Dies zeigte sich rückblickend, als die Teilnahmen an TIMSS/III (*Third International Mathematics and Science*

Study; Baumert, Bos & Lehmann, 2000) sowie in der Folge an den Studien PISA (*Programme for International Student Assessment*) und IGLU (*Internationale Grundschul-Lese-Untersuchung*) einen internationalen Rückstand des deutschen Bildungssystems aufzeigten. Als Ursache für die Missachtung der Forschung zur Schuleffektivität führen Bos und Postlethwaite (2000) das langjährige Fehlen von vergleichenden Schulleistungsstudien, also eine mangelhafte Einschätzung der Leistungsfähigkeit des Schulsystems, an. Erst nachdem der internationale Vergleich erfolgt war, erfuhr die Schulforschung, im Besonderen die Schuleffektivitätsforschung, eine erhöhte Beachtung. Demnach sind die Teilnahmen an TIMSS/III, PISA und IGLU als einschneidend für das Forschungsvolumen in der empirischen Bildungsforschung und die öffentliche Anerkennung anzusehen (vgl. Schwippert & Goy, 2008).

Die Bildungsadministration hat unter Bezug auf die ersten Ergebnisse der internationalen Vergleichsstudien entschieden, dass bildungspolitische Entscheidungen zur Entwicklung von Strukturen und Bildungsgängen zukünftig zuverlässig an messbaren Schülerleistungen auszurichten seien (Zedler & Döbert, 2010, S. 33; Buchhaas-Birkholz, 2009, 28f). So wurde unmittelbar nach PISA 2000 vom Bundesministerium für Bildung und Forschung (BMBF) die theoretische Basis für nationale Bildungsstandards entwickelt (Klieme et al., 2009), eine nationale Bildungsberichterstattung wurde installiert (Autorengruppe Bildungsberichterstattung, 2008), eine fortlaufende Beobachtung der Chancengerechtigkeit im Schulsystem wurde eingerichtet (Bertelsmann Stiftung, Institut für Schulentwicklungsforschung Dortmund & Institut für Erziehungswissenschaft Jena, 2017) und die Teilnahme an weiteren internationalen Assessments wurde initiiert (zusammenfassend beiBuchhaas-Birkholz, 2009). Insgesamt kann attestiert werden, dass das Gewicht und der Stellenwert der empirischen Bildungsforschung, gemessen anhand der Zahl der Forschungsprojekte, seit dem Jahr 2000 enorm angestiegen ist. Initialisierte Programme umfassen das *Nationale Bildungspanel* (Blossfeld, Doll & Schneider, 2008), die *Förderinitiative Technologiebasiertes Testen* (z. B. Goldhammer, Frank, Rölke, Scharaf & Upsing, 2008), ein Förderschwerpunkt zur Kompetenzdiagnostik (vgl. Deutsches Zentrum für Luft- und Raumfahrt e. V. - Projektträger im DLR, 2013, 179ff) und ein strukturelles Rahmenprogramm zur Förderung der empirischen Bildungsforschung (BMBF, 2008). Insbesondere letzteres Papier zeigt wie kaum ein anderes eine neue, zweite empirische Wende in der Bildungspolitik an (Bromme et al., 2014, S. 4).

Welches Selbstverständnis der empirischen Bildungsforschung ist also aus diesen beiden Einschnitten erwachsen? Der Gegenstandsbereich der empirischen Bildungsforschung umfasst Prozesse und Ergebnisse von Bildung über die volle Lebensspanne und innerhalb sowie außerhalb von (Bildungs-)Institutionen (Prenzel, 2005, S. 12). Die empirische Bildungsforschung in ihrer aktuellen Ausprägung versteht sich dabei als interdisziplinäres Forschungsfeld (Prenzel, 2005); Edelmann et al., 2012; Schwippert, 2016) und weniger als Teildisziplin der Erziehungswissenschaft. Dieses Selbstverständnis wurde ebenso wie das Forschungsfeld bereits durch den Deutschen Bildungsrat (1974) geprägt, welcher in Empfehlungen zu Bedarfen und Entwicklungsplänen neben der Erziehungswissenschaft auch die Soziologie, die Psychologie und die Ökonomie als Referenzdisziplinen der Bildungswissenschaften nennt. Der Erziehungswissenschaft kommt dabei eine Sonderrolle zu, da sie für die pädagogische Orientierung konstitutiv ist und eine reflexive Zusammenführung

und Integration der Perspektiven zu leisten hat (Deutscher Bildungsrat, 1974). Dies gilt unverändert bis heute, auch in aktuelleren Veröffentlichungen wird die besondere Rolle der Erziehungswissenschaft im Kontext der Bildungsforschung wiederholt betont (vgl. u. a. Schmidt & Weishaupt, 2008; Tippelt, 1998; Wischer & Tillmann, 2009). Nichtdestotrotz wird die Interdisziplinarität kritisch diskutiert. So schreiben Zedler und Döbert, dass Untersuchungen im Kontext der Bildungsforschung „ […]von (eher wenigen) Erziehungswissenschaftlern, von Psychologen und Soziologen durchgeführt [werden], zunehmend in Kooperationen." (Zedler & Döbert, 2010, S. 34). Der Diskurs um die Durchführung umfasst dabei die disziplinäre (vgl. Tippelt & Schmidt, 2010) ebenso wie die politische Dimension (Dammer, 2015) der Deutungshoheit, also insbesondere die reflexive Zusammenführung von Ergebnissen und die Formulierung von Handlungsableitungen. Evidenzbasierte Bildungsforschung bedeutet zusammenfassend, interdisziplinäre theoriegeleitete empirische Forschung nach wissenschaftlichen Kriterien in empirisch-pragmatischer Hinsicht zu betreiben (vgl. Tippelt, 2009).

Die Methoden der empirischen Bildungsforschung orientieren sich dabei an der speziellen Forschungsfragestellung und dem disziplinären Ursprungskontext, umfassen aber generell all jene Methoden, die wissenschaftlich verteidigungsfähig sind (z.B. Terhart, 2012, S. 30). Die Förderung durch die Bildungspolitik, die Deutsche Forschungsgemeinschaft und private Stiftungen betraf aber nach dem Jahr 2000 in besonderem Maße empirisch-analytische, vornehmlich quantitative Forschungsprojekte, die in erster Linie steuerungsrelevanten, leistungs- oder kompetenzorientierten Problemen nachgehen (Zedler & Döbert, 2010, S. 33; Edelmann et al., 2012, 98f; Terhart, 2012, S. 23). Der prominente Platz, den quantitative Methoden in der aktuellen Bildungsforschung noch immer einnehmen, kann auch an der Öffentlichkeitswirksamkeit der Forschungsprojekte festgemacht werden (Bromme et al., 2014). Die Zentrierung und auch die öffentliche Wahrnehmung quantitativer Forschung warf und wirft in der Erziehungswissenschaft verschiedene Kritikpunkte auf. Herzog (2010) stellt beispielsweise den Mangel philosophisch-anthropologischer Rahmenbezüge in Forschungs-arbeiten heraus und kritisiert den konstitutionellen Bezug pädagogischer Forschung auf politische Probleme. Ein weiterer zentraler Kritikpunkt richtet sich an die Verkürzung der beobachteten Prozesse und Strukturen durch die verwendeten Methoden (vgl. Brügelmann, 2015). Andererseits hatte aber beispielsweise bereits Anderson (1961) festgehalten, dass man zum Zwecke der Messung nationaler Bildungsbemühungen auf quantitative Methoden, wie sie in der Psychologie üblich sind, zurückgreifen müsse. Zudem sei die Öffentlichkeitswirksamkeit generell für eine Adaption der Ergebnisse in der Praxis notwendig, und eine Kooperation von pädagogisch Tätigen und Forschern sei charakteristisches Merkmal empirischer Bildungsforschung (Bromme et al., 2014; Herzog, 2016). Aber die Übersetzung von empirischen Ergebnissen in handlungsleitendes Wissen ist ein komplexer Vorgang, welcher zu Kontroversen führen kann und muss. Diese Kontroversen sind in wissenschaftstheoretischer Sicht, beispielsweise im Sinne des Kritischen Rationalismus (Popper, 1994), Merkmale normaler Wissenschaft (Bromme et al., 2014, S. 8). Insbesondere kann aber Konfliktpotenzial in den Bildungswissenschaften ausgemacht werden, denn bereits John Dewey hat 1916 in einer Vorgängertheorie des Kritischen Rationalismus aufgezeigt, dass empirische Bildungsforschung weit davon entfernt ist, sichere Prognosen zu liefern, dass sie aber verschiedenartige

Einflussfaktoren in einer sich schnell wandelnden Umwelt pragmatisch beschreiben muss (Edelmann et al., 2012, S. 60). Um diese Beschreibung vorzunehmen, müssen wissenschaftliche Aussagen getroffen werden.

Wissenschaftliche Aussagen grenzen sich von Alltagsaussagen maßgeblich über die Anwendung von Standards respektive Qualitätskriterien für die Ableitung von Wahrheitsaussagen ab (Döring & Bortz, 2016, S. 84). Bei diesen handelt es sich um die Relevanz, die methodische und ethische Strenge sowie die angemessenen Präsentation (ebd., S. 90). Bei Edelmann et al. (2012, S. 83) sind drei zentrale Kennzeichen zielführender quantitativer Bildungsforschung analog als (1) die klare Formulierung konditionaler Forschungsfragen, (2) ein systematischer und strukturierter Forschungsprozess und (3) die Replikation der Befunde formuliert.[3] Natürlich wird dabei auf allgemeine Verfahren und Standards, welche mit der Sozialforschung und der Psychologie geteilt werden, zurückgegriffen. Lehrbücher wie von Atteslander (2010) und Diekmann (2012) aus der Soziologie oder Döring & Bortz (2016) oder Sedlmeier und Renkewitz (2013) aus der Psychologie weisen hohe Schnittmengen zu den speziellen Erfordernissen des Feldes auf und sind auch in der Bildungsforschung als Basisliteratur aufzuführen. Ebenso gewann die Psychometrie, also die Theorie und Methode des psychologischen Messens, in den vergangenen Jahren an Gewicht (Yousfi & Steyr, 2006). Es können aber Differenzen im Sprachgebrauch und bei der Schwerpunktsetzung ausgemacht werden, welche auf die jeweiligen Anwendungsbereiche und Forschungstraditionen zurückgeführt werden können. Während in der Soziologie beispielsweise nichtexperimentelle Umfragetechniken dominieren, sind in der Psychologie das Experiment oder das Quasi-Experiment besonders prominent (Diekmann, 2012, S. 22). Für die Bildungsforschung haben insbesondere Evaluationsmethoden und die Interventionsforschung eine besondere Bedeutung (Cronbach, 1982). Nicht zuletzt anhand der großen Zahl der eigenständigen Lehrbücher ist erkennbar, dass sich zunehmend auch ein spezifisch bildungswissenschaftliches Methodenrepertoire entwickelt (Gräsel, 2011, S. 25).

Die quantitativen Werkzeuge sind jedoch in all diesen Forschungstraditionen die gleichen. So stellt die absolute gemeinsame Basis eines quantitativen Repertoires in jedweder Disziplin die Verwendung von statistischen Variablen dar. Variablen sind im Gegensatz zu Konstanten veränderliche Werte, deren Ausprägungen als zufällig betrachtet werden (z.B. Saint-Mont, 2011, S. 78f). In der Regel werden Verteilungen von Variablen, also mehrere gleichartige Beobachtungen, zum Beispiel über die Zeit oder mehrere Personen hinweg, betrachtet. Im Rahmen von Analysen kommen Variablen unterschiedliche Rollen zu. Dabei sind

- unabhängige Variablen, also kausale oder wirkende Variablen (a),
- abhängige Variablen, also Variablen, auf die eine Wirkung erfolgt (b), und
- Konditionalvariablen und intervenierende Variablen, also weitere nicht-ignorierbare Einflüsse (c)

[3] Die Wichtigkeit der Replikation wissenschaftlicher Befunde, insbesondere zur Absicherung kleiner Effekte, wie sie in der Bildungswissenschaft nicht unüblich sind, oder der Klärung widersprüchlicher Befunde, wird in der Literatur mehrfach hervorgehoben (z. B. Edelmann, Schmidt und Tippelt, 2012, S. 83; Bromme, Prenzel und Jäger, 2014, S. 7).

hinreichend, um überprüfbare Zusammenhangshypothesen zu formulieren. Die Zusammenhänge, die mittels dieser Kernbausteine über formale Modelle abgeleitet werden, sind zumeist einfach und wirken teils nahezu trivial. Trotzdem kann es zu großen Unklarheiten und Missverständlichkeiten kommen.

Die beispielhafte Aussage „Kinder im deutschen Schulsystem mit Deutsch als erster Fremdsprache (c) erreichen eine höhere Leseleistung (b), wenn diese eine zusätzliche Förderung erhalten (a)." kann auf verschiedene Weisen kritisiert werden. Beispielsweise ist die die Form des Zusammenhangs, also die Frage danach, ob die Zusammenhangsfunktion beispielsweise linear oder quadratisch ist, unklar. Die Bedingung, dass Deutsch die erste Fremdsprache ist, inkludiert die Kinder von Diplomaten ebenso wie unbegleitete Flüchtlinge; daher könnte diese Formulierung zu kurz greifen. Ebenso stellen die Begrifflichkeiten der Leseleistung und der Förderung einen generellen Anspruch, der nicht eingehalten werden kann. Es wird behauptet, dass eine beliebige Förderungsform wirkt und dass die Leseleistung angemessen und umfassend durch die abhängige Variable abgebildet wird, was auch beim Einsatz differenzierter Tests nicht immer der Fall ist. Die sprachliche und methodische Schärfe ist also von besonderer Relevanz, da die verwendeten Variablen niemals eine perfekte Realisation der Wirklichkeit sein können und demnach die Wahrnehmung und Kommunikation der Unschärfen durch die Forschenden in direkter Verbindung zu dem Gewicht der abgeleiteten Wahrheitsaussagen steht. Die Schärfe der verwendeten theoretischen Konstrukte, der operationalen Realisationen dieser und der mathematischen Variablenmodelle bedingt dabei die Tragfähigkeit jeder darauf basierenden Erkenntnisse und der Formulierung und Summe der praktisch anwendbaren Wahrheitsaussagen. Dies hat insbesondere deshalb Gewicht, weil viele der interessierenden theoretischen Konstrukte nicht direkt beobachtet werden können, sodass es nicht nur zu Beobachtungs- oder Stichprobenfehlern, sondern auch zu Operationalisierungsfehlern kommen kann. Diese werden in statistischen Modellen welche latenten Variablen verwenden explizert und bewertet.

3. THEORIE LATENTER VARIABLEN

Menschen können die Welt wahrnehmen und auf diese reagieren, aber ihre Kognition bleibt begrenzt (Mulaik, 1995, S. 284). Es ist eine natürliche menschliche Praxis, Geschehnisse, die anders nicht durchdrungen werden können, auf der Basis von Konzepten zu erklären, zu verstehen und manchmal vorherzusagen, welche nicht direkt beobachtbar sind (Bollen, 2002, S. 605). Dafür ist die Verwendung latenter Variablen eine übliche Praxis in der empirisch-quantitativen Bildungsforschung. Der wissenschaftliche Nutzen von latenten Variablen liegt vor allem darin, Hypothesen zu testen und gegebenenfalls zu falsifizieren, während sprachliche Unschärfen und Unterschiede in individuellen Wahrnehmungen umgangen werden und Messfehler expliziert werden. Der Nutzer kann, basierend auf der Zusammenhangsstruktur von Beobachtungen oder ähnlichen Antwortmustern oder Verhältnissen von Variablen und Testteilnehmern, Schlüsse innerhalb der Stichprobe überprüfen und gegebenenfalls verallgemeinern. Dies steht im Gegensatz dazu, hochgradig konkrete Aussagen zu dem

Verhältnis spezifischer Variablen, also idiosynkratischer Variablen, machen zu können; es erlaubt also stärkere Verallgemeinerungen (z. B. Bollen, 2002; Skrondal & Rabe-Hesketh, 2004).

Wie die meisten statistischen Techniken ist die Modellierung latenter Variablen keine isolierte Rechenoperation, sondern Teil eines Prozesses, der je nach Anwendung ähnliche Ideen, Normen und Praktiken bezüglich des Umgangs mit wissenschaftlichen Daten aufweist (Borsboom, 2008, S. 26). Und obwohl es zahlreiche statistische Analysemodelle gibt, welche die Idee latenter Variablen aufgreifen, gibt es keine verbreitete generalisierbare Definition, welche alle möglichen Anwendungen umfasst (ebd.).

Einerseits gibt es in der Literatur eine große Zahl von Alltagsdefinitionen, beispielsweise die Bezeichnung als hypothetische Konstrukte (z.B. Rombach, 1970) oder „[…] something that scientist put together out of their imaginations“ (Nunnally, 1978, S. 96) oder Metaphern (Mulaik, 2004). Diese Definitionen können auch kombiniert werden: „[…] they are essentially hypothetical constructs invented by a researcher” (Everitt, 1984, S. 2). Diese Definitionen haben aber nur einen begrenzten Nutzen, denn Definitionen dieser Art bedeuten im Umkehrschluss, dass beispielsweise fachliche Motivation oder Kompetenzen nicht real wären und sind somit als problematisch einzuschätzen. Eine weitere übliche Definition zentriert die nur indirekt gegebene Messbarkeit (Döring & Bortz, 2016, S. 483; Everitt, 1984, S. 2; Jöreskog, Sörbom & Magidson, 1979, S. 105). Für diese Definition ist es problematisch, dass die Forschenden annehmen müssen, dass auch zukünftig niemals eine direkte Messbarkeit gegeben sein kann (Bollen, 2002, S. 606), da sich der definierte Gegenstand nicht bedingt durch Fortschritte in der Messtechnik verändern sollte. Zudem kann für übliche vorgeblich beobachtbare Variablen, zum Beispiel dem Geschlecht, angenommen werden, dass diese ebenso eigentlich theoretische Dimensionen repräsentieren (Borsboom, 2008, S. 28). Das Geschlecht, welches in der Praxis nahezu immer dichotom erfasst wird, meint beispielsweise häufig eigentlich ein Kontinuum zwischen Weiblichkeit und Männlichkeit. Demnach würde auch in diesen Fällen eine indirekte Messung stattfinden, da das Theorem durch die Art der Messung verkürzt wird. Außerdem wären verschiedene physikalische Phänomene, zum Beispiel Wärme als Energieform, die über °C oder °F bestimmt werden können, ebenso latente Konstrukte – denn diese können auch nur indirekt, beispielsweise über die Ausdehnung von Alkohol gemessen werden. Intuitiv würde dies aber abgelehnt werden. Zuletzt sind die hier aufgeführten Ansätze häufig nicht hilfreich für die formale Strukturierung der latenten Variablen, da keinerlei technische Annahmen gemacht werden.

Andererseits gibt es zahlreiche operationale oder formale Definitionen, welche aber häufig zu eng an einzelne statistische Modelle oder Annahmen gebunden sind. Operational können latente Variablen als reine Aggregate, also eine sparsamere Form der Datendarstellung, wie sie zum Beispiel durch Faktor-, Komponenten- oder Hauptachsenanalyse erstellt werden, betrachtet werden. Dies ist aber vornehmlich nur für deskriptive und explorative Funktionen der latenten Variablen angemessen, da hier a priori kaum überprüfbare Annahmen gemacht werden können (z.B. Kasper & Ünlü, 2013). Die Definition von latenten Variablenmodellen über die stochastische Unabhängigkeit der manifesten Variablen voneinander, also das Verschwinden jedweden Zusammenhangs, wenn eine latente Dimension definiert wurde, ist ebenso üblich, aber in

verschiedenen Modelltypen unterschiedlich restriktiv (vgl. Kap. 3.1; McDonald, 1996). So kann diese Annahme beispielsweise in Faktormodellen gelockert werden. Hier mangelt es an Vergleichen zwischen den Definitionen der einzelnen Modelle und deren Implikationen (Bollen, 2002, S. 606). Allgemeine Probleme werden bei fragmentierten operationalen Definitionen durch Probleme verschleiert, welche jeweils an eine begrenzte Zahl von Anwendungen gebunden sind (ebd.). Rein formale allgemeine Definitionen latenter Variablen, also beispielsweise die Definition von der latenten Variable y als imperfekte Funktion einer Variable x, $y = f(x) + e$, allein sind in der Beschreibung des Gegenstandes nicht hinreichend, da diese ausschließlich eine Zusammenhangsfunktion oder Erwartung von Beobachtungen auf eine latente Struktur beschreiben können (Borsboom, 2008, S. 27), ohne dabei eine theoretische Verknüpfung zu erlauben. Ohne eine substanzwissenschaftliche Fundierung misst ein Test nur einen substanzlosen Zahlenwert, ohne weitere Überbrückungen misst ein Test nur, was ein Test misst (vgl. Hartig, Frey & Jude, 2008).

Es sind also definitorische Unterscheidungen in die zu messende Eigenschaft und deren Operationalisierung oder dem a priori formulierten theoretischen Konstrukt und den a posteriori statistisch abgeleiteten Variablen, respektive deren Modellen nötig; es existiert ein Überbrückungsproblem (z. B. Yousfi & Steyr, 2006; Bollen, 2002; Steyer & Eid, 2001). Anteilig wird hier monokausal vom Operationalisierungsproblem gesprochen (z.B. Gadenne, 1984; 1994), womit das Problem ausschließlich methodenseitig, im Vorgang der Operationalisierung, verortet wird. Diese Sichtweise ist unter Rückbezug auf Friedrichs (1982, S. 53) problematisch, denn hier wird klar herausgestellt, dass eine Theorie überprüfbare Regeln formulieren sollte und ohne diese Regeln nur geringen Erklärungswert hat; es würde sich also de facto um eine implizite Theorie handeln. Dies macht das Überbrückungsproblem zu einem Problem der Theorie und der Operationalisierung gleichermaßen. In der Folge sollen die a priori und die a posteriori Explikationen getrennt definiert werden.

Die a priori Explikation oder auch Nominaldefinition entspricht dabei der Spezifikation des Konzeptes in die intensionale (Inhalt) und extensionale (Umfang) Bedeutung des Konstrukts und ist direkt mit dem Begriff der Validität verbunden (vgl. Kap. 4.3). Die intentionale Bedeutung wird über Eigenschaftszuordnungen vorgenommen, die extentionale Bedeutung über Verknüpfungen und Vergleiche zu anderen bereits definierten Begriffen (Döring & Bortz, 2016, S. 225; Friedrichs, 1982, S. 75). Das ICD-10 System zur Klassifikation von Krankheiten und verwandten Gesundheitsproblemen (Deutschen Institut für Medizinische Dokumentation und Information [DIMDI], 2014) ist ein exemplarisches Beispiel für die Sammlung von Explikationen. In den Bildungswissenschaften können die Debatten um die Kompetenzbegriffe als Lehrbeispiel für die Genese von a priori Explikationen herangeführt werden (Leutner, Klieme, Fleischer & Kuper, 2013; Weinert, 2002).

Die operationale Definition setzt an der a priori Explikation an und konstituiert eine möglichst passende Operationalisierung. Hier wird der theoretische Begriff durch die Angabe bezeichnender Sachverhalte oder messbarer Eigenschaften oder die Wahl von Indikatoren standardisiert, also in Korrespondenzregeln übersetzt (Friedrichs, 1982, S. 77). Steyer und Eid (2001, S. 3) sprechen von der Erstellung von *termini technici*, die von umgangssprachlich

vordefinierten Begriffen streng entkoppelt sind. Die mathematisch-formale Definition des interessierenden theoretischen Konstrukts wird dabei durch zentrale Korrespondenzregeln beschrieben. Eine beispielhafte a priori Definition, in diesem Fall einer lernbezogenen fächerübergreifenden Handlungs–kompetenz, findet sich bei Schurig, Wendt, Kasper und Bosf (2015, S. 37), während die operationale Definition, also die Beschreibung der technischen Herleitung dieser Kompetenzform, auf der Seite 39 (ebd.) gegeben wird. In den aufeinander aufbauenden Beiträgen von Schurig, Busch und Strauß (2012), Schurig und Busch (2014) und Busch, Schurig, Bunte und Beutler-Prahm (2015) wird die a priori Definition des einheitlichen Interessengegenstandes intensional aus der Theorie abgeleitet und dann verschiedenartig operationalisiert und die Operationalisierungen werden verglichen.

Operationalen Definitionen liegen dabei die formalen Ableitungen der jeweiligen latenten Größen zugrunde. In der Folge werden zwei allgemeine formale Definitionen latenter Variablen vorgestellt, wie sie bei Bollen (2002, S. 609ff) zusammengetragen worden sind, namentlich die Erwartungswertdefinition und die Definition über die lokale stochastische Unabhängigkeit.

Die Definition als Realisierung eines erwarteten Wertes (*true score*) ist am engsten mit der klassischen Testtheorie verknüpft. Der wahre Wert eines Individuums T_i auf einer Variable i ist gleich dem erwarteten Wert E einer beobachteten Variable Y_i.

$$T_i = E(Y_i)$$

Damit wird die Ausprägung der Variable behandelt, als ob diese wiederholt in einem hypothetischen Experiment beobachtet worden wäre, ohne dass eine Antwort eine andere beeinflusst hätte (Lord & Novick, 1968, S. 29–30). E ist also der hypothetische Mittelwert von Y, und der beobachtete Wert setzt sich zusammen als

$$Y_i = T_i + \varepsilon_i,$$

wobei ε_i einen Fehler oder ein Residuum[4], also eine nicht aufgeklärte Varianz beschreibt. Hier wird die Metrik der latenten Variable durch $E(Y_i)$ bestimmt. Zudem wird angenommen, dass e_i nicht mit T_i zusammenhängt, die Fehler mehrerer beobachteter Variablen und die beobachteten Variablen selbst nicht zusammenhängen und die latenten Variablen alle Kovarianz der Indikatoren determinieren. Die Annahme, dass die Indikatoren nicht zusammenhängen, wird als lokale stochastische Unabhängigkeit oder linear experimentelle Unabhängigkeit (Fischer, 1974, S. 33) bezeichnet. Diese Annahme kann gelockert und durch Vergleiche konfirmatorischer Modelle geprüft werden.[5] [6]

Die Definition als lokale stochastische Unabhängigkeit liegt auch Messmodellen der probabilistischen Testtheorie zugrunde. Der Unterschied zu deren Annahme in der klassischen

[4] Residuen werden üblicherweise als die Differenzen zwischen den durch das Modell gegebenen und den empirischen Kovarianzen begriffen (Skrondal & Rabe-Hesketh, 2004, S. 273).
[5] Es sei angemerkt, dass die Annahme dieser operationalen Definition damit de facto bedeutet, dass formative Modelle keine latenten Variablen enthalten (vgl. Kap. 2.1.1).
[6] Für eine Unterscheidung von starker und schwacher lokaler Unabhängigkeit vergleiche z. B. McDonald (1996).

Testtheorie ist, dass diese strikter ist und explizit geprüft wird (Bühner, 2011, S. 574). Ist also der Wert der latenten Variable η zweier Personen gleich, so ist die Wahrscheinlichkeit p einer Antwort A_1 und einer Antwort B_1 gleich dem Produkt der Einzelwahrscheinlichkeiten (Rost, 1996, S. 73)

$$p(A_1 , B_1) = p(A_1|\eta) * p(B_1|\eta).$$

Die Kernannahmen dieser Definition sind, dass die Messfehler unzusammenhängend sind, also keine Residualkovarianzen verbleiben, die beobachteten Variablen unzusammenhängend sind, also Eindimensionalität angenommen werden kann und wenigstens zwei beobachtete Variablen vorliegen.

Aber für beide Definitionen ist es problematisch, dass es wenn Zusammenhänge von Indikatoren nicht logisch zurückgewiesen (Bollen, 2002, S. 614) oder Faktorstrukturen nicht hinreichend determiniert werden können (Asparouhov & Muthén, 2009), durch die zentralen Modellannahmen faktisch zur kontraintuitiven Ablehnungen des Modells der latenten Variablen führen müsste. Die beispielhafte Freigabe einer Residualkovarianz, also dem Zusammenhang zweier Fehlerterme, in einem Strukturgleichungsmodell entscheidet hier über die „Existenz" einer möglichen Formalisierung des zu messenden theoretischen Konstruktes und würde somit gegebenenfalls auch die Theorie methodisch unfundiert falsifizieren. Zudem verbietet es die Verwendung dieser Ansätze allein, Residuen als latente Konstrukte zu begreifen, welche ihrerseits als solche verstanden und analysiert werden könnten.

In Abgrenzung zu diesen modellspezifischen Definitionen für latente Variablen wirft Bollen (2002, S. 611) eine formale Definition als Stichprobenrealisierung auf. Eine latente Variable sei eine Zufallsvariable, für welche es keine Stichprobenrealisierung gibt, also keine Variablenwerte in den Rohdaten vorliegen. Dies umfasst die Perspektive, dass alle Variablen latent sind, bis eine stichprobenspezifische Beobachtung vorliegt. Das bedeutet auch, dass Variablen in einer Studie latent und in einer anderen manifest sein können oder sogar in einer einzelnen Studie erst latent und dann manifest sein können, nachdem nämlich Scores abgeleitet wurden. Dieser minimalistische Ansatz erlaubt, die Inklusion einer großen Zahl von Anwendungen und formale Definitionen können angeknüpft werden (Bollen, 2002).

Borsboom (2008) erweitert diese Definition dahingehend, dass für manifeste Variablen die Inferenz der Datenstruktur auf die Variablenstruktur mit Sicherheit vorgenommen werden kann, während diese Inferenz für latente Variablen messfehlerbehaftet ist. Es geht auch hier die Grundannahme ein, dass die meisten theoretischen Konstrukte in der Psychologie wie auch der Bildungswissenschaft latenter Natur sind. Eine theoretische Variable kann demnach nur dann als manifest behandelt werden, wenn Datenmuster einen deterministisch-kausalen Zusammenhang zu der Variable aufweisen, die Variable die einzige Quelle für Variation in den Beobachtungen ist und die Kardinalität in Variablen- und Datenstruktur gleich ist. Nur dann liegt eine bedeutungsgleiche Struktur vor (vgl. Kap. 3.2). Demnach müsste die Annahme, dass Variablen auch theoretisch manifest abbildbar seien, häufig fallen gelassen werden und jedwede Verletzung dieser Bedingungen als eine pragmatische Reduktion angesehen werden, die im Prozess der Datengenerierung mitgedacht werden muss. Oder anders gewendet: Jede Variable in

den Sozialwissenschaften ist eine latente Variable, bis eine perfekte Realisation des theoretischen Konstrukts gefunden wurde; demnach muss jede Verkürzung latenter Variablen auf eine „beobachtete" Variable plausibel sein (Borsboom, 2008, S. 49). Dies fordert einen reflektierten Umgang mit allen Variablen von Interesse, denkt den pragmatischen Reduktionsgedanken quantitativer Forschungstradition mit und erlaubt eine Verknüpfung zu verschiedensten formalen Definitionen. Die Unterscheidung von latenten und manifesten Variablen wird auf Unterschiede in der das Erkenntnis betreffenden, also der epistemischen Verfügbarkeit der Variablen zurückzuführen (vgl. Borsboom, 2008).

Zusammenfassend ist aber die klare Trennung und Gewichtung von den a posteriori und den a priori Definitionen latenter Variablen im Forschungsprozess häufig nicht gegeben. Der explorative Charakter latenter Klassenanalysen (vgl. Kap. 3.1) oder explorativer Faktoranalysen eignet sich beispielsweise insbesondere, um latente Variablen datengetrieben abzuleiten und wiederum weniger stark, um Hypothesen zu testen, womit eine starke Betonung der a posteriori Definition vorliegt. Dies kann beispielsweise im Beitrag von Schurig et al. (2012) nachvollzogen werden. Hier existierte keine hinreichende Kenntnis die Form einer latenten Struktur, respektive die hypothetische Struktur bildete sich in den Daten nicht ab, so dass eine alternierende Struktur operational erschlossen wurde. Steyer und Eid (2001, S. 2) betonen diesbezüglich beispielsweise, dass während der Entdeckung eines Gebietes noch grobe Skizzen genügen, aber spätestens bei dessen Erschließung und Kultivierung Präzision bei der Überbrückung vonnöten sei. Die Ansprüche an die operationalen Definitionen variieren also je nach Erkenntnisinteresse. Die Definitionen bedingen sich gegenseitig, die a prori Definition muss prüfbar sein und die a posteriori Definition eine angemessene Präzision aufweisen. Dabei entspricht eine Annäherung der a posteriori Definition an die a priori Definition einer höheren Generalisierbarkeit. Die selbstreflektierte Verortung einer Studie auf diesem Kontinuum, ist bei Ergebnisdarstellungen klar herauszustellen.

Dabei wird die Verwendung latenter Variablen in verschiedenen Bereichen als problematisch betrachtet, exemplarisch sei B. F. Skinner als prominenter Kritiker genannt (vgl. Glymour et al., 2014), der davon ausgeht, dass die Vorhersagekraft von Experimenten nicht durch die Verwendung mentaler Konzepte als interveneriende Variablen geteigert werden könne. Selbst in der Statistik werden latente Variablenmodelle häufig als eher obskure statistische Modelltypen betrachtet (Skrondal & Rabe-Hesketh, 2004, S. 1), die Varianz niemals vollständig aufklären und es Forschern erlauben, „subjektive" Vertrauensintervalle für Falsifikationen zu definieren (vgl. Kap. 4.3). Weitere Vorwürfe umfassen die Verwendung approximativer Fit-Maße (vgl. Kap. 5.2.1.2) und die Annahme, zu einfach auf Kausalität zu schließen (vgl. Bollen & Pearl, 2014; vgl. Kap. 3.3). Die Kritikpunkte werden in den folgenden Kapiteln einzeln aufgegriffen und bearbeitet. Im Folgenden sollen Konzepte und die basalen Modelltypen vorgestellt werden, die zum weiteren Verständnis notwendig sind.

3.1. MODELLTYPEN

Statistische Modelle können entsprechend ihrer Funktion kategorisiert werden, und dementsprechend können ihre Eigenschaften variieren. Generell können drei Formen von Zusammenhängen einzeln oder gemeinsam modellbasiert abgebildet werden, der empirische, der stochastische und der prädiktive Zusammenhang (Dempster, 1998, S. 253).

- Bei empirischen Zusammenhangsanalysen werden vorliegende Realisierungen mit gegebenen Verteilungen verglichen, um auf deren mathematische Ähnlichkeit zu schließen.
- Stochastische Zusammenhänge basieren auf den Wahrscheinlichkeiten des Auftretens von Realisierungen von Zufallsvariablen.
- Prädiktive Zusammenhänge beschreiben die geschätzte Performanz eines Objektes bei gegebenen Attributen.

Diese Zusammenhänge können einzeln oder gemeinsam in einem einzelnen Modell auftreten. So können statistische Modelle approximieren, erklären oder vorhersagen. Dabei gibt es in der Modellgenese, also der Erstellung der Modelle, ebenfalls drei Traditionen: das Anpassen von Modellen an vorliegende Daten, die Verknüpfung von stochastischen Modellen mit Beobachtungen und die Konstruktion formaler Verbindungen (Dempster, 1998, S. 257). Während die erste Tradition primär Strukturen aufdeckt, macht die zweite Tradition Approximationen an „wahre Modelle“ und in dritter Tradition können, in Anknüpfung an eine der ersten Traditionen, Verknüpfungen abgeleitet werden. Modelle sind immer zweckgebunden und werden zumeist nur einem Modellzweck zugeordnet.[7] Die Ableitung eines Messmodells mit latenten Variablen entspricht dabei der Verknüpfung eines stochastischen Modells mit Beobachtungen. Steyer und Eid (2001, S. 8) formulieren die wichtigsten Ziele von latenten Messmodellen parallel als (1) Überbrückung von Theorie und Empirie, (2) Explikation der logischen Struktur eines theoretischen Begriffs, (3) den Einbezug des Messfehlers in das Modell und (4) den Einbezug situationaler Effekte. Die über die Messmodelle abgeleiteten latenten Variablen können dann in Zusammenhangs- und Prädiktionsmodellen, also Vorhersage-modellen, verwendet werden. Im Rahmen von Strukturgleichungsmodellen wird von einer Differenzierung in Messmodelle und Strukturmodelle gesprochen. Diese Benennung wird hier auf andere Modelltypen übertragen. In der Abbildung 1 sind zwei Messmodelle und ein Strukturmodell abgetragen, welche modellbasiert gemeinsam verarbeitet werden. Die beobachteteten Werte werden dabei als Quadrate oder Rechtecke gezeichnet und können Indikatoren für latente Variablen darstellen, welche als Ovale oder Kreise auftauchen. Einzigartige Varianzen, also Fehlerterme oder Residuen, werden als doppelköpfige Pfeile auf einzelnen Variablen dargestellt, Kovarianzen würden als doppelköpfige Pfeile zwischen verschiedenen Variablen gezeichnet werden. Nur abhängige, also endogene Variablen, verfügen über Residuen. Regressive Zusammenhänge werden als Pfeile mit einem einzelnen Kopf dargestellt, der Pfeil entspricht der Wirkrichtung. Die latente Variable y_2 wird also zurückgeführt auf die latente Variable y_1. Es existieren Unterschiede zwischen verschiedenen

[7] Eine Ausnahme bilden beispielsweise Explorative Strukturgleichungsmodelle (ESEM; Asparouhov und Muthén, 2009).

grafischen Schreibweisen und deren Notation für latente Variablenmodelle (Boomsma & Hoogland, 2001; Kline, 2011; Reinecke, 2014), aber diese verkürzte Form ist für die Exempel in dieser Arbeit hinreichend.[8] In dem vorliegenden Beispiel gibt es zwei latente Variablen, welche regressiv miteinander verbunden sind. Jede der latenten Variablen hat jeweils drei distinkte Indikatoren.

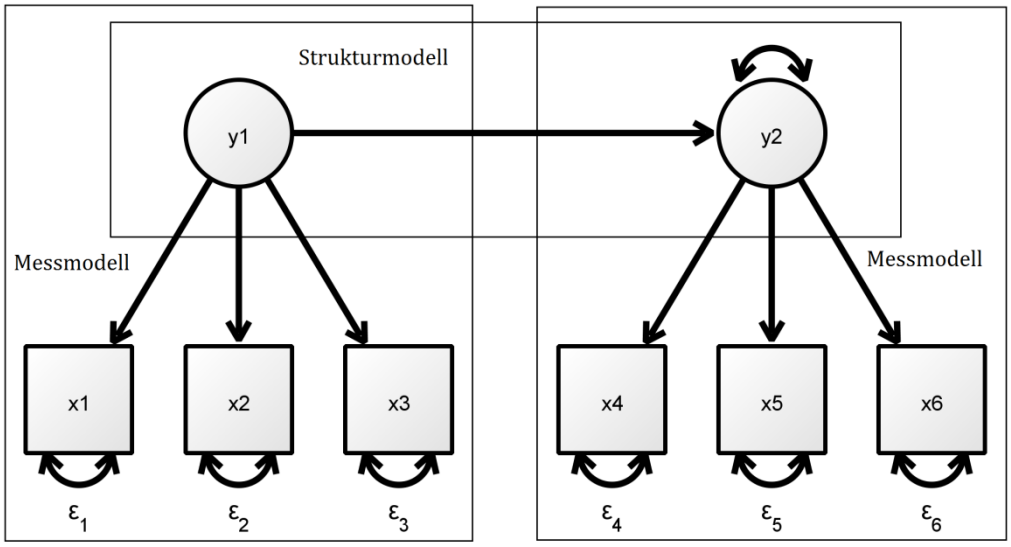


Abbildung 1: Mess- und Strukturmodell[9]

Die Richtung der regressiven Zusammenhänge, analog auch der Faktorladungen oder Ladungsgewichte, zwischen einer latenten Variable und ihren Indikatoren entspricht dabei der Annahme, dass die Kovarianz der Indikatoren durch die latente Variable determiniert wird; man spricht von einem reflexiven Modell (Bollen, 2002, S. 616). Alternativ werden in formativen Modellen die latenten Variablen durch die Beobachtungen determiniert (Bollen, 2002, S. 616). In verschiedenen Anwendungsformen kann diese Spezifikation angemessener sein, beispielsweise bei der Ableitung eines sozio-ökonomischen Status oder, genereller, bei der Bildung von indikatorgestützten Indizes. In der Abbildung 2 sind jeweils ein formatives und ein reflexives Modell abgetragen. Für formativ begründete Modelle gilt, dass die latenten Variablen vollständig operational definiert und messfehlerbehaftet sind. Damit kommt diesen keine theoretische Bedeutung jenseits der Bedeutung ihrer Indikatoren zu, entsprechend wichtig ist die Auswahl dieser. Die Indikatoren bilden die Bausteine des operationalen Konstrukts und können nicht ausgetauscht werden. Es liegt somit ein grundlegend anderes Verständnis von der Gültigkeit und

[8] Außerdem können latente Variablenmodelle in Matrixschreibweise und formal als paralleles Gleichungssystem dargestellt werden. Die grafische Schreibweise erfreut sich aber insbesondere wegen der leichten Vermittelbarkeit der Modellannahmen großer Beliebtheit.
[9] Alle Modelldarstellungen wurden mit Ωnyx von Oertzen, Brandmaier und Tsang (2014) erstellt.

Güte einer Messoperation vor, und die allgemeine Messtheorie (vgl. Kap. 3.2) kann hier keine Anwendung finden (Bühner, 2011, S. 34ff). Zur Modellevaluation können ausschließlich die Ladungsgewichte herangezogen werden und verschiedene Werkzeuge zur Modellbeurteilung, wie z.B. die Bestimmung der internen Konsistenz einer Skala (vgl. Kap. 4.2), können nicht sinnvoll angewendet werden (Bollen, 1989, S. 222). Die Eigenschaften von latenten Variablen dieser Art und den theoretischen Konstrukten, auf die geschlossen werden soll, lassen sich inhaltlich nicht mit reflexiven Ableitungen vergleichen. Ableitungsformen für formative Modelle sind beispielsweise die häufig genutzte Hauptkomponentenanalyse (Backhaus, Erichson, Plinke & Weiber, 2016, S. 412; Kasper & Ünlü, 2013),[10] eine Datenreduktionstechnik, welche gewichtete Summenscores erzeugen kann (Bartholomew, 2004).

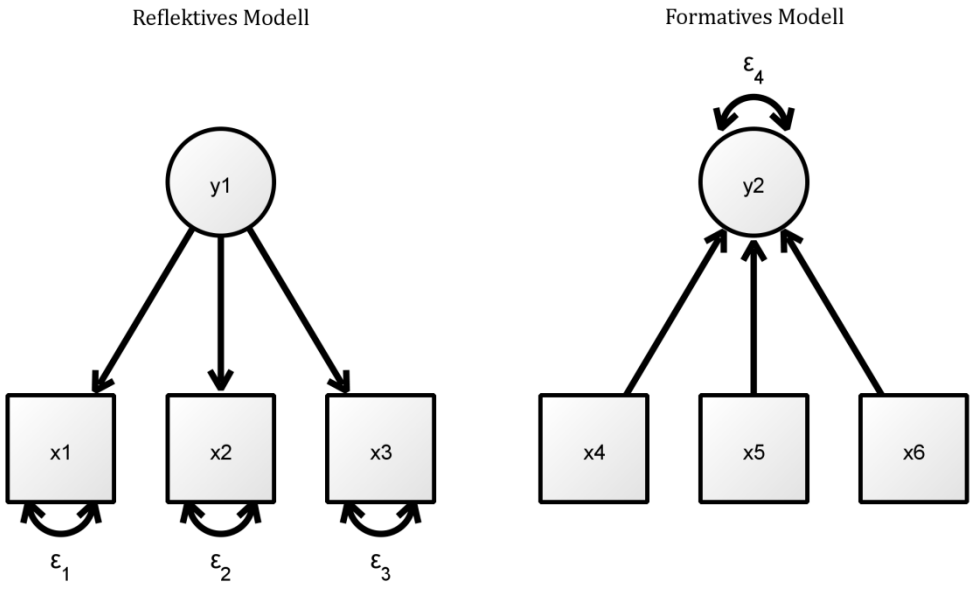


Abbildung 2: Reflexive und Formative Modelle

Die überwiegende Anzahl von Forschungsarbeiten im Kontext latenter Variablenmodelle nimmt an, dass Indikatoren reflektiv, also Manifestationen eines dahinterliegenden theoretischen Konstrukts, sind, auch wenn dies nicht immer methodisch so klar wie in der Abbildung 2 dargestellt determiniert werden kann (vgl. Kap. 3.1.2 und 3.1.3). Ohne diese Annahme kann ein Schluss auf dieses Konstrukt nur schwer erfolgen, da die Auswahl der Indikatoren die einzige Möglichkeit ist, eine Überbrückung zwischen der a priori und der a posteriori Definition zu leisten, es ist also eine große Vorkenntnis um konkrete Inhaltsbereiche, also die intensionale Explikation des theoretischen Konstruktes, notwendig. Fehlspezifikationen können im Strukturgleichungsansatz zu inkonsistenten Parameterschätzungen führen (Bartholomew, 2004). In dieser Arbeit werden maßgeblich reflektive Modelle behandelt.

[10] Vergleichbares gilt ebenso für die Hauptachsenanalyse. Explorative Faktormodelle hingegen können entsprechend den Modellannahmen formativ oder reflektiv sein.

Nahezu jedes Modell, das eine Art von theoretischer Struktur formal mit einer beobachteten Struktur verbindet, kann als latentes Variablenmodell begriffen werden (Borsboom, 2008, S. 26), und die Anzahl und Benennung verschiedener Modelltypen ist bestenfalls als verwirrend zu bezeichnen. Latente Variablenmodelle werden in vielen empirischen Disziplinen unter unterschiedlichen Namen verwendet. Skrondal und Rabe-Hesketh (2004, ix) formulieren dazu „We strongly believe that progress is hampered by the use of ‚local' jargon leading to compartmentalization.", und beziehen sich dabei sowohl auf mangelnden überdisziplinären Austausch als auch auf mangelnden Austausch zwischen den Nutzern verschiedener Modellansätze. Der in den Bildungswissenschaften prägende Ansatz der Klassifikation latenter Variablenmodelle bezieht sich auf die Metrik der manifesten und der latenten Variablen.

Als traditionelle Modelltypen werden dabei an verschiedenen Stellen latente Klassifikationsmodelle, also die latente Profil und die latente Klassenanalyse (z.B. Bacher & Vermunt, 2010), Item-Response-Modelle (Embretson & Reise, 2000) und Faktormodelle (Bollen, 1989) betrachtet. Es handelt sich bei diesen um Modelle, welche Werte von latenten Variablen oder Konstrukten ableiten (können) und bei denen die zugehörigen beobachteten Variablen als indirekte Beobachtungen oder fehleranfällige Messungen betrachtet werden (Skrondal & Rabe-Hesketh, 2007, S. 34). Ein Klassifikationsschema basierend auf der Metrik der theoretischen und beobachteten Variablen für traditionelle latente Variablenmodelle ist in Tabelle 1 gegeben. Eine weitere Unterscheidung der Modelltypen betrifft die Aggregationsobjekte. Während bei Faktormodellen und Item-Response-Modellen meist Variablen zusammengefasst werden, werden bei Klassifikationsanalysen üblicherweise die Testobjekte zusammengefasst.

Tabelle 1: Klassifikationsschema latenter Variablenmodelle (vgl. Skrondal & Rabe-Hesketh, 2007)

		Theoretische/ Latente Variable	
		kontinuierlich	kategorial
Beobachtete Variable	kontinuierlich	Common Factor Model/ Faktormodell Structural Equation Model (SEM)/ Strukturgleichungsmodelle Linear Mixed Model/ Gemischtes lineares Modell	Latent Profile Model (LPA)/ Latentes Profilmodell
	kategorial	Latent Trait Model/ Item Response Modelle (IRM)	Latent Class Model (LCA)/ Latentes Klassenmodell

Die Definition kategorialer (diskreter) und kontinuierlicher (stetiger) Variablen ist dabei nicht immer eindeutig und nicht gleichzusetzen mit der inhaltlichen Beschreibung als qualitative Variablen ohne Rangordnung und quantitative Variablen mit Rangordnung. Zufallsvariablen

können auch kategorial sein, wenn sie (endlich) viele oder abzählbar unendlich viele Ausprägungen haben. Kategoriale Variablen können, müssen aber keine logische Reihenfolge haben. Hilfreich ist eine Einteilung in die Möglichkeiten der anwendbaren Skalenniveaus.

Zur Theorie der Skaleneinteilung hat Stevens (1946) einen Grundstein gelegt. In der Statistik sind kategoriale Variablen nominalskaliert oder ordinalskaliert (vgl. Kap. 3.2.1). Sie können auch metrisch betrachtet werden, wenn nur wenige Schwellen vorliegen. Kontinuierliche Variablen hingegen werden intervallskaliert, verhältnisskaliert oder absolut skaliert (Fahrmeir, Künstler, Pigeot & Tutz, 2003). Zwischen den Beobachtungswerten, die in den Variablen erfasst werden und den Skalenniveaus existieren sowohl inhaltliche Grauzonen als auch mathematische Probleme. So lassen sich beispielsweise Schulnoten sowohl als ordinalskaliertes Merkmal als auch als intervallskaliertes Merkmal begreifen (Döring & Bortz, 2016, S. 5ff). [11] Zusammenfassend können kontinuierliche Variablen geordnet und gemessen werden, während kategoriale Variablen nur zählbar sind.

Während die Frage nach der „korrekten“ Metrisierung maßgeblich manifeste Variablen betrifft, kann auch die Metrik latenter Variablen empirisch nicht immer zweifelsfrei nachgewiesen werden. Während die Metrik in der Operationalisierung determiniert wird, ist sie nicht immer unabhängig von der Natur der vorliegenden Indikatoren (z.B. Borsboom et al., 2003). Grundsätzlich gilt, da die latente Variable nicht gemessen wurde, dass die Maßeinheit durch die Forschenden festgelegt werden muss. Dies determiniert die Wahl des Modells. Über das vorgestellte Kategorienraster hinweg sind zudem Mischformen und Kombinationen möglich. So können beispielsweise mehrere distinkte Systeme von kontinuierlichen latenten Variablen erstellt werden, die ihrerseits latente Klassen definieren (Rost, 1990). Zudem existieren Item-Faktor-Modelle, also Modelle mit ordinalen manifesten Items im Strukturgleichungsansatz (Wirth & Edwards, 2007). Einen Überblick über Mischformen gibt Muthén (2007). Grundlegend ist es möglich, die Verarbeitungen in einem einzelnen Schritt vorzunehmen, was für die Schätzung der Standardfehler vorteilhaft ist. Da die Modelle, insbesondere im Falle probabilistischer Modelle, aber schnell zu komplex werden, um geschätzt zu werden (vgl. Kap. 5.1), ist eine Verarbeitung in zwei Schritten nicht unüblich. Beispielsweise werden bei Jaekel, Schurig, Florian und Ritter (im Erscheinen [2017]; Beitrag 6) die Personenscores von IRM als Indikatoren in einem SEM und bei Schurig et al. (2015) als Indikatoren einer LPA verwendet.

Um die Differenzen und Ähnlichkeiten der im Raster vorgestellten Modelltypen zu geben, werden im folgenden Abschnitt stark verkürzte Einführungen der Modelltypen gegeben.

[11] Generell werden in der Forschungsarbeit häufig ordinalskalierte Variablen als intervallskaliert begriffen und verwendet. Insofern die Verteilungen der verwendeten Variablen hinreichend normal und die Stichprobenumfänge groß genug sind, ist dies unbedenklich; es kann jedoch zu Problemen führen, wenn Deckeneffekte oder schiefe Verteilungen vorliegen.

3.1.1. FAKTORMODELLE

Das faktoranalytische Modell (FA) ist eines der populärsten Verfahren, wenn an Techniken für den Umgang mit latenten Variablen gedacht wird (Bollen, 2002, S. 623). Das Konzept wurde 1906 von Spearman eingeführt und insbesondere durch Jöreskog (z.B. Jöreskog et al., 1979) erweitert. Wenn von Faktormodellen die Rede ist, meint dies heute üblicherweise konfirmatorische Modelle (*Confirmatory Factor Analysis*; CFA), welche wiederum im Rahmen von Strukturgleichungsmodellen gedacht werden und als Teile eines Strukturmodells verwendet werden können, um Zusammenhänge zwischen latenten Variablen aufzuklären. Alternativ existieren explorative Faktormodelle (*Exploratory Factor Analysis*; EFA), in welchen latente Dimensionen nicht prädeterminiert, sondern aus der Ladungsstruktur abgeleitet werden; hierbei handelt es sich konzeptionell um einen operativen Ansatz der Definition theoretischer Strukturen. Strukturgleichungsmodelle (SEM) können als Kombination von Regressionsmodellen und Pfadanalysen verstanden werden und sind in einer großen Zahl von Lehrbüchern vorgestellt worden (vgl. Bollen, 1989; Kline, 2011)[12]. FA werden üblicherweise nicht verwendet, um Personenwerte abzuleiten, sondern um nachzuweisen, dass eine klassische Ableitung eines Wertes, zum Beispiel als Summenscore, angemessen ist oder, um auf Zusammenhangsstrukturen mit latenten Variablen zu schließen.

Die Grundannahme der FA ist, dass sich der Wert einer beobachteten Variablen zusammensetzt aus eine spezifischen Komponente, welche die latente Variable darstellt, und einem zufälligen Messfehler[13], der nicht mit anderen Fehlern und weiteren latenten Variablen zusammenhängt. Diese Annahmen liegen ebenso Regressionsmodellen zugrunde. Damit ist die FA auf der formalen Definition als *„true score"*-Modell aufgebaut, erweitert diese aber. Was die FA und Regressionsmodelle trennt, ist die Inklusion unbekannter Größen über Kovarianzstrukturen der beobachteten Größen. Die Annahme gilt, dass die unbeobachtete Variable ein Prädiktor der beobachteten Variablen ist. Die latente Variable wird als Kovarianz der beobachteten Indikatoren aufgefasst. Die basale Statistik von Faktormodellen ist damit die Kovarianz zweier beobachteter kontinuierlicher Variablen x und y, welche bestimmt wird als das Produkt der Pearson-Korrelation r zwischen beiden Variablen und deren jeweiligen Standardabweichungen SD und stellt also ein unstandartisiertes Zusammenhangsmaß da.

$$cov_{xy} = r_{xy} SD_x SD_y$$

Daraus lässt sich ableiten, dass die Analyse von Kovarianzstrukturen zwei Ziele haben kann, (1) die Zusammenhangsmuster der beobachteten Variablen zu verstehen und (2) so viel Varianz wie möglich durch das Modell zu erklären (Kline, 2011, S. 10). Der Teil des Modells, welcher Hypothesen über Varianzen und Kovarianzen enthält, ist die angenommene Kovarianzstruktur

[12] In dieser Arbeit wird der Einfachheit halber in der Folge von Faktormodellen (FA) gesprochen. Dies meint reflektive konfirmatorische Faktormodelle. Wenn andere Modelltypen gemeint sind, wird dies expliziert.

[13] Kenny (1979) schlägt vor, den Anteil unaufgeklärter Varianz als unbegreiflich zu verstehen, da er, in Ermangelung eines besseren Begriffs, den freien Willen darstellt – die Fähigkeit von Menschen in Situationen außerhalb spezifischer externer oder interner Einflüsse zu handeln. Diese Sichtweise erschöpft sich aber für Untersuchungen in denen die Fälle einzelne Personen repräsentieren.

(ebd.), Modellannahmen werden also als Restriktionen[14] in der Kovarianzmatrix umgesetzt. Im Falle einer eindimensionalen CFA wird also angenommen, dass die Ausprägungen der beobachteten Variablen durch die Kovarianz aller beobachteten Variablen determiniert werden. Dies begründet die alternative Benennung von Faktormodellen als Kovarianzstrukturmodelle (Reinecke, 2014).

Das lineare Modell der FA kann für Messmodelle, entsprechend einer linearen Regression, geschrieben werden als

$$\boldsymbol{x} = \boldsymbol{\Lambda_x \xi} + \boldsymbol{\delta},$$

wobei $\boldsymbol{x}$ einen $1 * q$ Vektor auf einer Zahl von q Variablen darstellt, $\boldsymbol{\Lambda_x}$ eine $q * k$ Matrix von Regressionsgewichten, also den Faktorladungen. $\boldsymbol{\xi}$ ist ein $1 * k$ Vektor von zugrunde liegenden Dimensionen k und $\boldsymbol{\delta}$ ein $1 * q$ Vektor, der die Residuen darstellt (Bollen, 1989, S. 10ff; Kaplan, 2009, S. 40ff).[15] Die Ausprägungen der manifesten Variablen sind determiniert als die Faktorladungen auf eine latente Variable und deren Residuen. Dabei wird angenommen, dass der Erwartungswert der latenten Variablen und der Fehler gleichermaßen 0 ist und diese nicht zusammenhängen. Aller Zusammenhang zwischen beobachteten Variablen, die eine latente Größe repräsentieren sollen oder, im mathematischen Sinne richtiger, durch diese in reflektiven Modellen determiniert werden, wird also in einer Kovarianzmatrix $\boldsymbol{\Sigma}$ zusammengefasst. Wobei $\boldsymbol{\Sigma}$ der $q * q$ Populationskovarianzmatrix entspricht. $\boldsymbol{\Phi}$ ist die $k * k$ Matrix der Varianzen der latenten Variablen und Kovarianzen, $\boldsymbol{\Theta_\delta}$ ist die diagonale Matrix der Fehler (Kaplan, 2009, S. 41). Ein Apostroph zeigt dabei eine Transponierung an.

$$\boldsymbol{\Sigma} = cov(xx') = \boldsymbol{\Lambda_x \Phi \Lambda_x'} + \boldsymbol{\Theta_\delta}$$

Um der latenten Variable im FA-Ansatz eine Metrik zu geben, können Restriktionen auf jeweils zumindest ein Ladungsgewicht (unit loading identification; ULI) oder die Varianzen der latenten Variablen (unit variance identification; UVI) eingeführt werden (Kline, 2011, S. 127). Üblicherweise werden die Ladungen oder Varianzen auf 1, notwendigerweise aber nur auf einen positiven Wert fixiert (ebd.).[16] Für latente Variablen, die ein Messmodell aufweisen, ist dabei zu bedenken, welche beobachtete Variable fixiert wird. In Abhängigkeit von deren Trennschärfe und Gewicht kann dies zu verschiedenen Modellgüteschätzungen führen (ebd.). Ebenso ist eine Fixierung von Varianzen nicht zielführend, wenn die Varianz der latenten Variablen beobachtet werden soll oder deren Gleichheit bei unabhängigen Stichproben nicht angenommen werden kann. Da im faktoranalytischen Ansatz Fehlerterme als latente Variablen behandelt werden, gilt

[14] Der Begriff Restriktion kann missverständlich sein. De facto ist jede Restriktion eines Modells eine Annahme gegenüber einem (saturierten) Obermodell, in welchem alle Parameter frei variieren und alle Parameter zusammenhängen, das Obermodell wird restringiert. Das Gegenteil ist das maximal restringierte Nullmodell, bei dem alle Kovarianzen mit 0 spezifiziert werden (vgl. Bentler, 1980).

[15] Eine Vertiefung der zugrunde liegenden Matritzenalgebra kann und soll an dieser Stelle nicht gegeben werden. Spezifische Vertiefungen zu diesem Thema finden sich bei beispielsweise bei Bollen (1989), Kaplan (2009) oder Reinecke (2014).

[16] Die Fixierung des Ladungsgewichts des ersten Indikators auf 1 ist in vielen Softwarepaketen (z.B. Mplus und lavaan) die Standardeinstellung.

die Notwendigkeit der Skalierung ebenso für diese, üblicherweise werden hier die Ladungsgewichte auf 1 fixiert, da die nicht aufgeklärten Varianzanteile von Interesse sind.

3.1.2. ITEM-RESPONSE-MODELLE

IRM modellieren das probabilistische Zustandekommen von Antworten auf Items und zentrieren dabei auf einzelne Items und weniger auf den Test im Ganzen. IRM wurden in den 1960er-Jahren vor allem durch die Arbeit von Rasch (1980) populär und durch Lord und Novick (1968) stärker verbreitet. Während das Rasch-Modell ursprünglich nur für dichotome Antworten gedacht war, gibt es inzwischen zahlreiche Erweiterungen auch für polytome Antworten und multiple Einflussfaktoren. IRM werden üblicherweise verwendet, um Personenwerte abzuleiten, die Bezeichnung als „Messmodelle" ist also inhaltlich zutreffender als im FA Ansatz. Eine umfassende Einführung findet sich beispielsweise bei Embretson und Reise (2000).

Im Rahmen von IRM wird das Antwortverhalten von Personen als nicht-lineare Funktion der intrapersonellen Personenfähigkeit und der Eigenschaft der zu beantwortenden Items begriffen. Die zu schätzenden Parameter sind also die Eigenschaften der Items und die Eigenschaften der einzelnen Testteilnehmer. Beide werden auf einer einheitlichen Skala von logarithmisierten Wahrscheinlichkeiten (Logits) abgetragen. Einparametrische Modelle (1PL) wie das Rasch-Modell für dichotome (Rasch, 1980) oder das Partial-Credit-Modell für polytome Antwortmöglichkeiten (Wright & Masters, 1982), schätzen die Personenfähigkeit und die stichprobenweit einheitliche „Schwierigkeit" oder Lage des Items. Zweiparametrische Modelle (2PL) wie das 2PL-Modell nach Birnbaum (1968) oder das Generalized-Partial-Credit-Modell (Muraki, 1992; GPCM) schätzen einen weiteren Itemparameter, den Diskriminanz- oder Steigungsparameter, welcher eine vertiefende Betrachtung des Zusammenhangs zwischen dem Item und dem Antwortverhalten erlaubt. Ein 3PL-Modell (Birnbaum, 1968) ergänzt wiederum einen sogenannten Rateparamer.[17] Exemplarisch beschränkt sich die folgende Darstellung auf das 1PL-Modell.

Im 1PL-Modell ist die Itemschwierigkeit die einzige modellierte Eigenschaft des Items, und sie leitet sich aus den Antworthäufigkeiten ab. Diese ist für alle Testteilnehmer gleich. Der Personenparameter wird auf Basis der beobachteten Daten geschätzt und dessen Genauigkeit wird über den Standardmessfehler dieser Schätzung bestimmt (vgl. Kap. 4.2). In der Item-Response-Theorie wird also die Wahrscheinlichkeit, ein Item x zu beantworten (hier beispielsweise mit 1), von der Itemschwierigkeit σ, also der Schwierigkeit, auf diesem Item den Wert 1 zu erreichen, und dem Personenparameter θ determiniert (z. B. Embretson & Reise, 2000; Rost, 1996)

$$p(x = 1|\theta, \sigma) = \frac{\exp(\theta - \sigma_1)}{1 + \exp(\theta - \sigma_1)}.$$

Formal liegt hier eine logistische Funktion vor, die die Antworten in eine kontinuierliche Verteilung überführt. Die Annahmen gelten, dass die Itemcharakteristika, also die Form der logistischen Funktion, für alle Items homogen und die Items unabhängig voneinander sind

[17] Das unübliche 4PL Modell ergänzt zusätzlich einen Unachtsamskeitsparameter.

(Steyer & Eid, 2001, S. 233). Der angenommene Zusammenhang ist demnach nicht-linear. Die Schätzung der Parameter erfolgt über die Likelihoodfunktion (vgl. Kap. 5.2.1.1). Die Schätzung der Parameter σ und θ kann parallel erfolgen, üblicher ist allerdings eine getrennte Schätzung in mehreren Rechendurchläufen, um die Schwierigkeitsparameter zu fixieren. Es erfolgt eine bedingte Maximum-Likelihood-Schätzung. Diese und weitere Verfahren der Likelihood-Schätzung werden detailliert bei Embretson und Reise (2000) vorgestellt.

An der Grenze zwischen IRM- und CFA-Modellen haben sich zusätzlich Item-Faktor-Modelle etabliert (Muthén, 2007; Wirth & Edwards, 2007). Bei diesen handelt es sich um faktoranalytische Modelle mit ordinalen Indikatoren (vgl. Kap. 5.2.1.1). Diese werden derer erhöhren Robustheit gegenüber schiefen Verteilungen im Rahmen von SEM beispielhaft bei Schurig et al. (2012) und Busch et al. (2015) eingesetzt.

In IRM ist es nicht möglich durch die Modellspezifikation einfach zwischen reflektiven und formativen Modellen zu unterscheiden, da die Modelle assoziativ, also korrelativ sind. Die Differenzierung muss hingegen als theoretische Setzung gemacht werden.

3.1.3. LATENTE KLASSENMODELLE

Die Analyse kategorialer latenter Variablen wurde in den 1950er-Jahren durch Lazarsfeld für kontinuierliche manifeste Variablen eingeführt (1959) und später für kategoriale Variablen erweitert (Goodman, 1985; Lazarsfeld & Henry, 1968). Weitere Bezeichnungen sind beispielsweis finite Mischverteilungsmodelle, Clusterverfahren mit Messfehlern oder Modelle mit diskreten latenten Variablen. Beide Verfahren weisen substanziell hohe Ähnlichkeiten auf und werden in dieser Arbeit daher weitgehend zusammenfassend dargestellt. Latente Klassifikationsmodelle ordnen Testobjekte probabilistisch qualitativen, also kategorialen Klassen ohne Rangreihung zu und können damit als Form der Fuzzy-Clusteranalysetechniken (vgl. Kaufman & Rousseeuw, 1990) verstanden werden (Bacher & Vermunt, 2010). Dies kann konfirmatorisch geschehen, also indem eine bekannte Zuordnungsregel angewandt wird, oder explorativ, indem eine empirische Zuordnungsregel abgeleitet wird. Über die Organisation von Daten in dieser Form können Datenstrukturen exploriert, partitioniert und aggregiert werden. Die Zuordnung findet dabei auf der Basis hoher Ähnlichkeit der Antwortmuster innerhalb der Klassen oder geringer Ähnlichkeit zwischen den Klassen statt (Bacher, Pöge & Wenzig, 2010; Kaufman & Rousseeuw, 1990). Üblicherweise werden LCA/LPA verwendet, um unbekannte Zusammenhangsstrukturen zwischen den Testteilnehmern zu explorieren.

Die konkreten Modellannahmen besagen, dass den Daten eine nicht beobachtete Anzahl von Klassen zugrunde liegt, konstante Antwortmuster innerhalb einer Klasse vorliegen, die Klassen lokale stochastische Unabhängigkeit aufweisen und die Klassen exhautiv und disjunkt sind, also jedes Objekt klassiert werden kann und jedes Objekt nur einer Klasse angehört. Aller Zusammenhang zwischen den Indikatoren wird auf die Klassenzugehörigkeit zurückgeführt (Bacher & Vermunt, 2010). Unbekannt ist a priori wie viele Personen den einzelnen Klassen angehören, welche Personen welcher Klasse angehören und gegebenenfalls wie viele latente Klassen existieren. Das Optimierungskriterium, das maximiert wird, um ein Modell auszuwählen, ist die Likelihood (vgl. Kap. 5.2.1.1). Geschätzt werden die Modellparameter über

die beobachteten Antwortmuster. Dabei wird angenommen, dass bei einer perfekten Modellanpassung die „Lösung" x, also die Wahrscheinlichkeit ein Item i zu lösen oder eine Kategorien zu wählen (z. B. $k =$ „*nie*", „*manchmal*" *oder* „*immer*") innerhalb der Klassen g für alle Personen gleich sind, womit ein Personenparameter v entfällt.

$$p(x_{vi} = k|g) = p_{ikg}$$

Die Wahrscheinlichkeit, mit der eine Person Items in Kombination kreuzt, ist das Produkt der bedingten Wahrscheinlichkeiten aller Items n bei gegebener Klassenzugehörigkeit. Damit kann die Antwortmusterwahrscheinlichkeit $p(a_v)$, also die relative Häufigkeit eines Musters, ausgedrückt werden als:

$$p(a_v|g) = \prod_{i=1}^{n} p_{ikg}$$

Die unkonditionale Antwortmusterwahrscheinlichkeit ist das Produkt der bedingten Klassenzugehörigkeitswahrscheinlichkeiten mit der Summe aller relativen Klassengrößen π_g:

$$p(a_v) = \sum_{g=1}^{G} \pi_g \prod_{i=1}^{n} p_{ikg}$$

Daraus lässt sich mittels des Theorems von Bayes ableiten, wie hoch die Wahrscheinlichkeit ist, einer Klasse zuzugehören:

$$p(g|a_v) = \frac{\pi_g * p(a_v|g)}{p(a_v)}$$

Womit alle Modellparameter bis auf die Zahl der latenten Klassen geschätzt werden können; diese wiederum kann über Modellvergleiche abgeleitet werden.

Auch in LCA/LPA ist es nicht, möglich durch die Modellspezifikation zwischen reflektiven und formativen Modellen zu unterscheiden. Die Differenzierung wird als theoretische Setzung vorgenommen. Eine direkte Gegenüberstellung von der Verarbeitung eines einheitlichen längsschnittlichen Datensatzes auf der Basis von FA und latenten Klassenanalysen, bei welchen die unterschiedlichen Ansätze der a priori und a posteriori Definitionen kontrastiert werden können finden sich im Beitrag von Schurig und Busch (2014).

3.1.4. GENERALISIERTES LATENTES VARIABLENMODELL

Während die vier Modelltypen sich ideengeschichtlich getrennt voneinander entwickelt haben, wurde in den 1990er-Jahren festgestellt, dass sich diese in einem einheitlichen generalisierten Rahmen fassen lassen. Dabei entwickelten sich verschiedene Ansätze, die die einzelnen Modelle jeweils als Spezialfälle eines übergreifenden Modells begreifen, wie das latente Variablenmodell von Mplus (Muthén, 1984; Muthén & Muthén, 1998-2015), *Generalized Linear Latent and Mixed Models* (GLLAMM; Rabe-Hesketh, Skrondal & Pickles, 2004) oder das *Generalized Linear Item Response Model* (GLIRM; Moustaki & Knott, 2000). Eine Zusammenfassung der Ansätze findet

sich zum Beispiel bei Skrondal und Rabe-Hesketh (2007). Für einen Überblick zu den technischen Details siehe Skrondal & Rabe-Hesketh, 2004 Kapitel 3).

Die Essenz der generalisierten Modellformulierung ist die mathematische Spezifikation hierarchisch konditionaler Zusammenhänge. Über eine Link-Funktion f und den Erwartungswert E einer Matrix beobachteter Variablen $\mathbf{X}$ wird über eine lineare Funktion g auf eine latente Struktur $\boldsymbol{\theta}$ geschlossen (Borsboom, 2008, S. 26),

$$f\big(E(\mathbf{X})\big) = g(\boldsymbol{\theta}).$$

Generalisierte lineare Modelle umfassen mehrere regressive Ansätze für nicht notwendigerweise normalverteilte abhängige Variablen (Fahrmeir, Kneib & Lang, 2007, S. 189). Ein derartiges Modell charakterisiert sich zum einen durch die Wahl der Verteilungsannahme der latenten Variable, also dem Typ der Exponentialfamilie. In den Bildungswissenschaften sind dies in den allermeisten Fällen normale, binominale oder polynominale Verteilungen, möglich sind aber auch andere, zum Beispiel Poissonverteilungen. Zum anderen sind die Wahl der Responsefunktion, also des linearen Prädiktors, und die Definition und Auswahl der Kovariablen, also der Spezifikation eines Strukturmodells, relevant. Diese Generalisierung besitzt keine Gültigkeit für formative Modelle.

Die gemeinsame Basis aller vorgestellten Modelle sind die Annahmen der Mess- und Testtheorie. Da diese implizit bei der Verwendung von latenten Variablenmodellen sowie generell aller empirischen Repräsentationen durch Zahlenwerte mitgedacht werden müssen, werden beide zusammenfassend eingeführt.

3.2. MESSTHEORETISCHE GRUNDLAGEN

Messmodelle explizieren die Beziehung zwischen Theorie und Empirie und geben dem untersuchten theoretischen Begriff eine formale Struktur. Ein Messmodell ist dabei notwendige Voraussetzungen der Falsifizierbarkeit eines theoretischen Modells (Steyer & Eid, 2001, S. 8). Es ermöglicht eine Übertragung von der Theorie in die empirische Praxis. Wie ausgeführt wurde (vgl. Kap. 3), ist die Bildungsforschung ohne diese Übertragungen, unabhängig davon ob sie in qualitativer, quantitativer oder textwissenschaftlicher Tradition erfolgen, keine empirische Disziplin. Den Rahmen für dieses Vorgehen in quantitativer Tradition geben die Messtheorie (z.B. Orth, 1974) und Testtheorien (z. B. Fischer, 1974; Lienert & Raatz, 1998; Rost, 1996) vor. Sowohl die Mess- als auch die Testtheorie beschäftigen sich mit statistischen Messmodellen, die Unterschiede liegen dabei im empirischen Ausgangspunkt (Steyer & Eid, 2001; S. V). Dabei ist das „Messen […] sowohl historisch als auch methodisch gesehen eine der Grundlagen der Wissenschaft. Ohne die Durchführung exakter Messungen lässt sich die Entwicklung der empirischen Wissenschaften, insbesondere der Naturwissenschaften, nicht vorstellen.“ (Orth, 1974, S. 9).

3.2.1. MESSTHEORIE

Begründet wurde die moderne repräsentative Messtheorie durch den Psychologen Stevens (1946, 1957) und seine hierarchische Taxonomie von Skalenniveaus sowie der Zuordnung

jeweilig angemessener Rechenoperationen.[18] In der repräsentativen Theorie des Messens wird davon ausgegangen, dass in experimentellen Untersuchungen Beobachtungen oder Dingen numerische Werte zugeordnet werden können, um strukturerhaltend Fakten über diese zu repräsentieren. Es muss also eine eindeutige Repräsentation für eine beobachtete Relation gefunden werden. Diese erlangt Bedeutsamkeit, wenn sich der Wahrheitswert der auf den beobachteten Werten beruhenden Aussage nicht ändert, wenn unterschiedliche Skalen angelegt werden (vgl. Saint-Mont, 2011). Beispielsweise darf die relationale Aussage „A ist schwerer als B." nicht davon abhängen, ob in Pfund oder in Kilogramm gemessen wird. Die verschiedenen Skalenniveaus müssen also ineinander transformierbar sein, um bedeutsam zu sein. Stevens (1946) etablierte hierzu eine Hierarchie von Skalenniveaus, die definiert, welche mathematischen Operationen für entsprechend skalierte Variablen, welche Transformationen und welche Interpretationen zulässig sind. Die Bedeutsamkeit definiert damit die zulässigen Lage, Streuungs- und Zusammenhangsmaße. Die wichtigsten Skalentypen, also die Nominal-, die Ordinal-, die Intervall- und die Ratioskala, sowie deren Lagemaße, und die angemessenen Verteilungsfunktionen sowie zulässigen Verhältnisstatistiken sind in der Tabelle 2 abgetragen (Döring & Bortz, 2016; Fahrmeir et al., 2003; Stevens, 1946).[19]

Tabelle 2: Skalenniveaus (vgl. Stevens, 1946)

Skalenniveau	Definition	Mögliche Lagemaße	Mögliche Verteilungsfunktionen	Verwendbare Statistiken	Beispiele
Nominalskala	Reine Betitelung, keine Ordnung möglich	Modus	Häufigkeiten	Gleichheit/ Ungleichheit	Geschlecht, Familienstand
Ordinalskala	Ordnung, aber keine Interpretation der Abstände möglich	Median	Perzentile/ Spannweite	Größer/ Kleiner	Schulnoten, Straßennummern
Intervallskala	Wie Ordinalskala, aber Interpretation der Abstände möglich	Arithmetisches Mittel	Varianz/ Standardabweichung	Unterschied/ Distanz	Kalenderdatum
Ratio- oder Verhältnisskala	Wie Intervallskala, aber Sinnvoller Nullpunkt vorhanden	Geometrisches Mittel und harmonisches Mittel	Variablilität	Verhältnis	Blutdruck, Länge, Gewicht, Alter

[18] Es existieren weitere Messtheorien, die klassische und die operationale (vgl. Orth, 1974), welche aber für die Bildungswissenschaften nur eine geringe Relevanz haben, da entweder eine perfekte Realisierung bekannt ist, oder der Messprozess die Eigenschaft bestimmt. Die Messtheorie wird umfassend in drei Bänden von Krantz, Luce, Suppes und Tversky behandelt (Krantz, Luce, Suppes und Tversky, 2007; Luce, Krantz, Suppes und Tversky, 2007; Suppes, Krantz, Luce und Tversky, 2007).

[19] Ursprünglich umfasste die Taxonomie nur die vorgestellten Skalen. Später wurden aber bei Stevens (1957) noch die Log-Intervall und die Absolutskala ergänzt. Zudem existieren noch weitere Abstufungen, welche aber im sozialwissenschaftlichen Kontext nur eine geringe Relevanz haben.

Das Skalenniveau sagt dabei nichts darüber aus, ob eine Skala abzählbar oder nicht abzählbar, also kategorial oder kontinuierlich, respektive diskret oder stetig ist. Nur für Nominalskalen kann klar festgehalten werden, dass diese immer kategorial sind.

Es verbleibt das Skalierungsproblem, also die Frage danach wie konkret gemessen werden kann, wie also mehrere Beobachtungen in einen Zahlenwert überführt werden können (vgl. Saint-Mont, 2011; Kap. 2). Dies wird in der Regel über konstruktive Beweise vorgenommen. Dabei werden mathematische Objekte durch ihre mathematische Konstruktion begründet, wobei die objektivistische Annahme gilt, dass das Objekt existiert, aber seine Existenz erst durch die Konstruktion begründet wird. Für theoretische Konstrukte in den Bildungswissenschaften, wie auch generell in den Sozialwissenschaften, ergeben sich hier verschiedene Probleme. Zum einen sind die Begriffe häufig nicht so formuliert, dass eine Formalisierung möglich ist (vgl. Kap. 3), und zum anderen existiert kein Einheitensystem, welches eine direkte Messung ermöglichen würde. Demnach kann kritisiert werden, ob eine angemessene Repräsentation überhaupt gegeben werden kann. Hier setzt die Testtheorie an.

3.2.2. TESTTHEORIE

Der Begriff „Test“ ist dabei nicht eng gefasst zu verstehen. Nach Lienert und Raatz (1998, S. 7) umfassen Tests Verfahren zur Untersuchung von Persönlichkeitsmerkmalen, den Vorgang der Durchführung der Untersuchung, die Gesamtheit der Durchführung sämtlicher Requisiten, jede Untersuchung, soweit sie Stichprobencharakter hat und mathematisch-statistische Prüfverfahren. Zumeist sind allerdings strukturierte Tests gemeint, die aus einer Batterie vollstandardisierter Items bestehen. Dies umfasst beispielsweise standardisierte Leistungstests ebenso wie Sammlungen von Indizes zur Bestimmung der sozialen Lage oder Instrumente der Persönlichkeitsdiagnostik (z.B. Rost, 1996). In der Testtheorie ist die Beobachtungsbasis das Antwortverhalten von Personen auf Fragen oder Items oder Testverfahren. Die Realität wird abgeleitet, indem überzufällige Zusammenhänge und (Un-)Ähnlichkeiten kausal auf unbekannte Größen bezogen werden oder indem hypothetisierte Effekte betrachtet werden. Während es Unterschiede in der Beobachtungsbasis zwischen der Mess- und der Testtheorie gibt, ist das Ziel jedoch dasselbe: Es sollen Messwerte zugeordnet werden, um damit quantitative Gesetzmäßigkeiten zu formulieren (Steyer & Eid, 2001). Die bedeutsamsten Untertheorien sind die klassische Testtheorie und die probabilistische Testtheorie (vgl. Kap. 3). Die Testtheorie kann als Unterkategorie der Messtheorie begriffen werden, aber die Abweichung in der Beobachtungsbasis hat Folgen für die Theoreme der Eindeutigkeit, der Repräsentation und der Bedeutsamkeit. Strenggenommen lassen sich die Angemessenheit der Repräsentation und die Kardinalität der Messung, also die geforderte Eindeutigkeit, in den Sozialwissenschaften häufig nicht vollständig bestimmen (Diekmann, 2012, S. 297). Es ist und bleibt unbekannt, ob beispielsweise die Bestimmung einer Kompetenz auf einer Intervallskala perfekt eindeutig und repräsentativ ist (Döring & Bortz, 2016, S. 244; Moosbrugger & Kelava, 2008, S. 18f). Das Skalenniveau wird also per fiat als Konvention festgelegt (Bühner, 2011; Diekmann, 2012; Döring & Bortz, 2016). Aus der Passung des Skalenniveaus auf die Daten baut der Bedeutsamkeitsbegriff der Testtheorie auf, beschrieben wird damit hier also die Menge der

Aussagen, unter denen ein Wahrheitswert invariat ist. Da es keine Idealrepräsentationen von beispielsweise Einstellungen oder Kompetenzen gibt, wie sie in der Messtheorie verlangt werden, haben weitere testtheoretische Konzepte ein besonderes Gewicht. Um die Güte der Repräsentation theoretischer Größen einzuschätzen, müssen Tests formal identifizierbar (vgl. Kap. 5.1) und testbar (vgl. Kap. 5.2) sein (Steyer, 1989). Dabei wird heute nur noch selten davon ausgegangen, dass ein Modell vollständig zu den Daten passt oder nicht passt, hingegen wird die Güte der Anpassung bestimmt (Kline, 2011, S. 190). Während die Probleme der Identifizierbarkeit und Testbarkeit zentral in der probabilistischen Testtheorie behandelt wurden, zeigt Steyer (1989) auf, dass die Fragen auch für die Modelle der klassischen Testtheorie relevant sind.

Reaktionen werden also eindeutige Zahlenwerte zugeordnet, die möglichst repräsentativ sind. Diese können in ein mathematisches Mess- oder Testmodell überführt werden, wodurch durch Prüfungen auf Modellgültigkeit und eine Prüfung der Modellparameter Widersprüche im formalen Modell oder gegebenenfalls in der Theorie ausgemacht werden und Prognosen abgeleitet werden können.

Die Akzeptanz der vorgestellten imperfekten Repräsentation zu beobachtender Merkmale und der häufig unvollständigen Anpassung formaler Modelle an die Daten hat einen direkten Effekt auf die praktische Betrachtungsweise, welcher auch quantitativ-analytisch abgeleitete Zusammenhänge unterliegen müssen. Diese Akzeptanz reflektiert sich im Umgang mit empirischen Sätzen und deren Abgrenzung zu normativen Sätzen (Popper, 1994, S. 7ff) und dem daraus erwachsenden Verständnis von interessierenden substanzwissenschaftlichen Wirkungszusammenhängen.

3.3. KAUSALITÄT UND EVIDENZ

Die Kausalität, also die Beziehung zwischen Ursache und Wirkung, ist die Grundlage des Denkens über die moderne Welt. Viele der Fragestellungen mit denen sich Forschende auch in den Bildungswissenschaften beschäftigen, sind in ihrem Kern kausal, so bearbeitet beispielsweise die Schulforschung Fragestellungen zu den Wirkungen von Unterricht. Also kann die Frage nach Kausalität in Studien zu Effekten von Interventionen, in Erhebungen über Wahrscheinlichkeiten von Kontrafaktizität oder Überzufälligkeiten und in Mediationsanalysen aufkommen. Die Frage nach Wirkzusammenhängen ist also bedeutsam, aber auch schwer zu beantworten. Denn es kann keine gültige Induktion von speziellen Beobachtungssätzen in der Wissenschaft, also von empirischen Erkenntnissen, auf allgemeine Aussagen stattfinden, ohne dass hinreichende Rand- oder Anfangsbedingungen formuliert werden, welche wiederum die Einschränkung der Beobachtungen in die Aussage einbetten (Popper, 1994).

Auf Kausalität wird daher über aus der Theorie abgeleitete Hypothesen geschlossen, die begründete Vermutungen über den generellen Zusammenhang und die Richtung von Zusammenhängen zwischen Variablen darstellen. Die Hypothesen können dann Falsifikations–versuchen unterworfen werden. Lazarsfeld formuliert in der Übersetzung von Hummel und Ziegler (1976) zur Kausalität, dass x y verursacht, wenn

1. ein statistischer Zusammenhang besteht, also eine Falsifikation misslang,
2. die Ursache (x) der Wirkung (y) zeitlich vorangeht und
3. der Einfluss nicht verschwindet, wenn Drittvariablen (z) kontrolliert werden.

Eine ähnliche Formulierung findet sich in Anlehnung an John Stuart Mill, beispielsweise bei Shadish, Cook und Campell (2002, S. 6): „(1) the cause preceeds the effect (2) the cause was related to the effect, and (3) we can find no plausible alternative explanation for the effect other than the cause." Es ist offensichtlich, dass insbesondere der dritte Punkt bei beiden Definitionen problematisch ist und bei Shadish und Kollegen gelockert wurde, denn die Zahl möglicher Drittvariablen und alternativer Erklärungsansätze ist unbegrenzt. Nichtsdestotrotz wird über die quantitative Beobachtung und Interpretation von Zusammenhängen Kausalität *konditional* nachvollzogen. Die Konditionen, unter denen ein Zusammenhang formuliert und beobachtet worden ist, sind also ein Gütekriterium für die inhaltliche Belastbarkeit der Beobachtung und deren Ableitungen, folglich für die Generalisierbarkeit von Zusammenhangsaussagen. Alle inhaltlich möglicherweise bedeutenden Konditionen, also Drittvariablen oder alternative Erklärungsansätze, die nicht beobachtet, also kontrolliert werden, verbleiben als Verzerrungen, Störvariablen oder unbeobachtete Konfundierungen. Der beobachtete Zusammenhang ist hingegen konditional unabhängig von allen angemessen verarbeiteten Bedingungen, dann gilt ein Effekt von x auf y bei Kontrolle von z. Zudem umschreibt die Phrase „no plausible alternative explanation" (s. o.) das erkenntnistheoretische Prinzip des zureichenden Grundes (z.B. Shadish et al., 2002, S. 4). Jedes Erkennen muss demnach in angemessener Weise auf einen beobachtbaren Grund zurückgeführt werden, und es müssen Begründungszusammenhänge abgeleitet werden. Statistische Modelle sind ein Mittel, einen möglichst angemessenen Begründungszusammenhang herzustellen und zu replizieren, respektive verschiedene angemessene Begründungszusammenhänge miteinander zu vergleichen.

Ein bewährter Ansatz, Drittvariablen auszuschließen und einen zureichenden Grund zu isolieren, ist es, das Untersuchungsdesign in den Mittelpunkt zu stellen. Der Gold-Standard für kausales Schließen ist dabei das randomisierte Experiment (Bühner, 2011; Diekmann, 2012; Döring & Bortz, 2016). Die Logik folgt hier der Annahme, dass konfundierende Variablen damit auf eine Experimental- und eine Kontrollgruppe gleichverteilt werden können. Somit wäre ein kausaler Effekt gleich der Differenz der Effekte in der Experimental- und der Kontrollgruppe, da ausschließlich die interessierende unabhängige Variable zwischen den Gruppen variiert.

Die Anwendung des randomisierten Experimentes ist aber in der Bildungsforschung aus praktischen Gründen häufig unmöglich (Bromme et al., 2014, S. 14). Dies kann in den Kosten, dem Zeitaufwand, ethischen Gründen bei der Wahl der Zuweisungsmechanik in die Gruppen oder zum Beispiel weit praktischer in der Unkontrollierbarkeit der Forschungsumgebung (z. B. dem Klassenraum) begründet sein. Der Auftrag der Bildungsforschung ist außerdem die Bereitstellung handlungsleitender Informationen. Dafür müssen die gewonnenen Erkenntnisse eine hohe externe Validität (vgl. Kap. 4.3) aufweisen, also möglichst generalisierbar sein, was für experimentelle Settings in geringerem Maße zutrifft als für nicht randomisierte Studien (Döring & Bortz, 2016, S. 196). Daher haben vergleichende Analysen, Trendanalysen und Panelstudien eine besondere Bedeutung (Blossfeld, Maurice & Schneider, 2011). Zudem ist ein

experimentelles Setting auch immer ein atypisches Setting, welches für die Evaluations- und Interventionsforschung ungeeignet sein kann, die beide ein besonderes Gewicht für die Bildungsforschung aufweisen (Cronbach, 1982).

Ein Ansatz jenseitig von zufälligen Gruppenzuweisungen ist die künstliche Erstellung von Gruppen zur Konstruktion vergleichbarer Einheiten. Dabei können verschiedene Lösungsstrategien angewandt werden, um eine Annäherung an experimentelle Settings zu erzeugen. Die Zuweisung kann beispielsweise über die Homogenisierung von Gruppen, also auf der Basis kriterialer Zuweisungen (z. B. dem Geschlecht, Pre-Test-Scores oder Leistungsquantilen) passieren, außerdem kann eine Feststellung von vorhandener oder mangelnder Messäquivalenz (vgl. Kap. 5.2.3) zwischen den Vergleichsgruppen vorgenommen werden. Die Zuteilung kann a priori ebenso vorgenommen werden wie post-hoc und sie kann uni- oder multivariat sein, also auf einzelnen oder mehreren Kriterien beruhen. Hier hat das mathematische Rubin Causal Model (RCM; Rubin, 1974; Holland, 1986) einen großen Einfluss, nicht nur in der Bildungswissenschaft, sondern allgemeiner in den Sozialwissenschaften, der Medizin und der Statistik (vgl. Gangl, 2010). Es erweitert die Logik des randomisierten Experimentes auf Studien, bei denen keine Zuteilung in eine Experimental- und eine Kontrollgruppe vorgenommen werden kann. Das Fundamentalproblem der Kausalanalyse (Holland, 1986) wird umgangen, indem ein Forschungsdesign implementiert wird, welches eine wahrscheinlichkeitsbasierte Abschätzung eines kausalen Effekts mithilfe des Vergleichs empirisch beobachtbarer Ereignisse rechtfertigt (Gangl, 2010, S. 932). Demnach können kausale Effekte geschätzt werden, wenn Experimental- und Kontrollgruppe post-hoc durch ein mathematisches Zuweisungsmodell gebildet werden. Die zentralen Anforderungen an Anwendungen des RCM (Rubin, 1974, S. 699f) lassen sich dabei auf alle Bereiche der empirischen Bildungsforschung übertragen, die sich mit Wirkungszusammenhängen beschäftigen.

- Die Variabilität der abhängigen Variable muss a priori weitest möglich spezifiziert sein, um Wirkungszusammenhänge formulieren zu können.
- Konfundierende, aber irrelevante Variablen müssen ignoriert werden, da ansonsten gewünschte Ergebnisse forciert werden können, indem wir die *richtigen* irrelevanten Prädiktoren mit einbeziehen.

Analyseformen, die in den vergangenen Jahren steigend an Bedeutung gewonnen haben, um Kausalität nachzuvollziehen, sind auf dem RCM beruhende wahrscheinlichkeitsbasierte Matchingverfahren (*Propensity Score Matching*; z. B. Stuart, 2010) und die vor allem in der Ökonometrie verwendeten Regressions-Diskontinuitätsanalysen (Thistlethwaite & Campbell, 1960). Auf diese wird in dieser Arbeit aber nicht weiter eingegangen.

Der Begriff der Wahrscheinlichkeit, der auch dem RCM zugrunde liegt, hat generell in der Frage nach der Entdeckung kausaler Zusammenhänge mittels statistischer Hypothesen große Verbreitung gefunden (z.B. Pearl, 2013); die Debatten um Risiken jedweder Natur machen aber klar, welches Gewicht Wahrscheinlichkeiten heute für Schlussfolgerungen haben (Borovcnik, 2014, S. 11; vgl. Gigerenzer, 2004, 2013). Denn Wahrscheinlichkeitshypothesen können in

objektivistischer Perspektive falsifiziert werden, wenn dabei auch kein strenger Kausalbegriff zugrunde liegen kann (Schroeder-Heister, 2013, S. 212). In der Wissenschaftstheorie der allgemeinen Statistik wird diesbezüglich zwischen Kausalität und Ätialität unterschieden (vgl. Diaz-Bone & Weischer, 2014, S. 14). Während der harte Kausalitätsbegriff das Prinzip beschreibt, dass gleiche (Einzel-)Ursachen gleiche Wirkungen verursachen, besagt das Prinzip der Ätialit, dass gleiche allgemeine Ursachen eine gleiche Menge von empirischen Wahrscheinlichkeitsverteilungen bedingen. Diese Unterscheidung entspricht der zwischen deterministischer und stochastischer Kausalität (vgl. Steyer, 1992). Das Prinzip der stochastischen Kausalität beruht dabei auf (Über-)zufälligkeiten und Wahrscheinlichkeiten, während das Prinzip deterministischer Kausalität ausnahmslose Gültigkeit verlangt. Damit wird akzeptiert, dass eine empirisch-statistische Information per Definition niemals eine vollständige kausale Information sein kann und statistische Modelle keinen zweifelsfreien Beweis für kausale Zusammenhänge erbringen können (Pearl, 2000).

Dabei ist das Konzept der Wahrscheinlichkeit keinesfalls einheitlich. „Probabilitas“ meinte im Wortsinn die Glaubhaftigkeit einer durch Autorität gesicherten Meinung, was heute ebenso wenig wie Glück oder Schicksal zur theoretischen Annäherung an den Begriff verwendet wird (Gigerenzer, 2004, S. 19). Inzwischen haben sich zwei zentrale Begriffsverständnisse von Wahrscheinlichkeit entwickelt. (1) Der subjektive Wahrscheinlichkeitsbegriff versteht die Wahrscheinlichkeit als subjektives Sicherheitsmaß, welches aus Überzeugungen und Informationen gebildet wird. Die wichtigste theoretische Strömung ist hier Bayes‘ Wahrscheinlichkeitstheorie (ebd.). Wenn aus dem Vorwissen[20] heraus Annahmen zu der Wahrscheinlichkeitsverteilung gemacht werden können, werden diese in die Berechnung von Wahrscheinlichkeiten einbezogen. Bei der Erweiterung des Begriffs durch de Finetti (1981) wird festgestellt, dass Wahrscheinlichkeit hier als Ausdruck unzureichender Information zu verstehen ist. (2) Der objektive Wahrscheinlichkeitsbegriff besagt wiederum, dass keine Berechnung exakter Wahrscheinlichkeiten möglich ist, sondern Wahrscheinlichkeiten Ausdruck der Eigenschaft der Messobjekte und/oder von Messfehlern sind. Die sogenannten Frequentisten nehmen dabei an, dass sich aus hinreichend ähnlichen Referenzhäufigkeiten Wahrscheinlichkeiten ableiten lassen (Gigerenzer, 2004, S. 11). Eine striktere Sichtweise in dieser Begriffstradition ist die Betrachtung von Wahrscheinlichkeiten als Propensitäten, bei denen die Kenntnis der kausalen Struktur vorausgesetzt wird. Ein populärer Vertreter dieser Sichtweise ist Popper (1994), wobei dieser Ansatz den Anwendungsbereich stark einschränkt (Gigerenzer, 2004, S. 11).

Während also der subjektive Wahrscheinlichkeitsbegriff stark auf Annahmen beruht, baut der objektivistische auf Beobachtungen auf, was wiederum die Anwendungsbereiche determiniert. Die Irrtumswahrscheinlichkeit in objektiver Tradition ist also definiert als die Wahr–scheinlichkeit eines Stichprobenergebnisses unter der Geltung der Nullhypothese, während diese in subjektiver Tradition als das Zustandekommen der Nullhypothese bei gegebenem Stichprobenergebnis betrachtet wird (Döring & Bortz, 2016, S. 616).

[20] Dieses Vorwissen kann über theoretische Annahmen, Testkalibrierung, Expertenurteile oder sogar normativ gegeben werden.

Das Prinzip der Gültigkeitsprüfungen von Wahrheitsaussagen auf der Basis von probabilistischen und stochastischen (Modell-)Formulierungen stößt dabei sowohl in der statistischen als auch in der wissenschaftstheoretischen Tradition auf Kritik. Im Kern bezieht sich diese auf den Umgang mit den zu definierenden Schwellenparametern, Vertrauensbereichen und, im Falle der subjektiven Theorie, auf den inkludierten Verteilungsannahmen, denn bei einer zu liberalen Definition der Vertrauensbereiche werden Hypothesen nicht-falsifizierbar (Clauss, 1981, S. 157). Daher werden methodologische Kriterien und wissenschaftliche Standards als Basis für die Beurteilung von Evidenzen herangezogen.

Statt deterministisch zu schließen, wird also auf Evidenzen für den Nachweis von Zusammenhängen vertraut. Diese Evidenzen müssen dabei in quantitativer Tradition notwendigerweise statistisch bedingt sein, sie sind aber erst dann hinreichend bedingt, wenn diese theoretisch und sachlogisch abgeleitet wurden (Weiber & Mühlhaus, 2010, S. 16). Evidenz ist dabei semantisch nicht alltagssprachlich zu verstehen, meint also keine Offensichtlichkeit, sondern eine Nachweisbarkeit in angelsächsicher Sprachtradition (Jornitz, 2009, S. 68), der Begriff entspringt also eher einem naturwissenschaftlich-empirischen Konzept als einem geisteswissenschaftlich-hermeneutischen (Bromme et al., 2014, S. 6). Der Evidenzbegriff hat auch jenseits der Sozialwissenschaften in den vergangenen Jahrzehnten verstärkte Bedeutung erlangt, zum Beispiel im Bereich der Medizin (z.B. Little & Rubin, 2000) oder auch in der Rechtsprechung (z. B. §137f Sozialgesetzbuch V; 27.11.2016). Unsere Welt ist evidenzbasierter geworden (Borovcnik, 2014, S. 11) und sogar Physiker ersetzen kausale Paradigmen durch den Zufall, „[…] ‚stochastische Kausalität' ist heutzutage für viele kein schmutziges Wort mehr." (Steyer, 1992; XII).

Was jedoch vor dem Hintergrund komplexer Bildungsprozesse einerseits als zu starke Verknappung interpretiert werden kann (Herzog, 2010), kann andererseits im Sinne eines Kausalbegriffs wie bei Rubin (1974) als notwendige und pragmatische Komplexitätsreduktion verstanden werden. „Generell zeichnen sich empirische Zugänge durch eine hohe Realitätsorientierung und einen ausgeprägten Pragmatismus auf" (Bromme et al., 2014, S. 15). Aber Evidenz bleibt auch immer unabgeschlossen. Mangelhafte Designs, Stichproben oder ungeeignete statistische Verfahren sind Beispiele dafür, wie es möglich sein kann, schwache Evidenz oder scheinbare Evidenz zu erzeugen (Bromme et al., 2014, S. 7). Doch neue und bessere Evidenzen können jederzeit vorgelegt werden. Diese Vorläufigkeit von Evidenz ist demnach nicht als Schwachpunkt, sondern als Normalfall wissenschaftlicher Forschung zu verstehen und steht der Nutzung der Erkenntnisse in Bildungspolitik und Bildungsadministration nicht entgegen (Bromme & Prenzel, 2014). Das, was im Rahmen der Bildungs- und Sozialforschung also erreicht werden kann, ist, die belastbaren empirischen Grundlagen für evidenzbasierte Entscheidungen zu liefern (Bromme et al., 2014, S. 6ff). Idealtypisch sollte dabei sogar auf Befunde aus mehreren Studien mit gleichen und alternierenden methodischen Ansätzen verwiesen werden können, welche im Einklang mit einer

Hypothese sind, um jenseits methodologisch evidenter Beobachtungen auf Kausalität zu verweisen (Edelmann et al., 2012, S. 83).[21]

Vor dem Hintergrund der stochastischen Kausalität und des Verständnisses von Forschungsergebnissen als Evidenzen erscheint die Titulierung von Mess- und Wirkungsmodellen als „Kausalmodelle“, wie zum Beispiel bezogen auf SEM im Lehrbuch von Weiber und Mühlhaus (2010) oder im einflussreichen Beitrag von Bentler (1980), als schlecht haltbar. Diese Benennung liegt zwar in der Tradition der Interpretation von Pfadmodellen als Kausalmodelle (Hummell & Ziegler, 1976; z.B. Steyer, 1992), ist aber irreführend. Diesen Betrachtungsweisen ist gemein, dass immer ein stochastischer Kausalbegriff vorliegt, und es wird nahe gelegt, dies expliziter zu betonen, um Missverständnisse bezüglich eines unterschiedlichen Vokabulars, zu vermeiden. Per Definition handelt es sich bei allen hier behandelten latenten Variablenmodellen um stochastische Modelle[22] (Steyer & Eid, 2001; Teil II) und der zentrale Umgang mit deren Vertrauensbereichen, also den Schwellen, auf deren Basis Evidenz angenommen oder abgelehnt wird, ist gleichzusetzen mit dem Umgang mit Gütekriterien von Messungen respektive Tests. Sie ist gleichzusetzen, mit einem ernst zu nehmenden Versuch der Falsifikation sensu Popper (1994).

4. GÜTEKRITERIEN VON MESSUNGEN UND TESTUNGEN

Messungen und Tests im bildungswissenschaftlichen Kontext unterliegen Standards und Kriterien der Wissenschaftlichkeit und einer großen Zahl von Haupt- und Nebengütekriterien (Diekmann, 2012; Döring & Bortz, 2016, 2016; Friedrichs, 1982; Lienert & Raatz, 1998; Steyer & Eid, 2001), welche zusammengenommen die Qualität der empirischen Forschung belegen. Dabei hat eine starke Orientierung an den Gütekriterien originär psychologischer Tests zur Untersuchung von Persönlichkeitsmerkmalen stattgefunden, da diese besonders hohe Standards aufweisen (Bühner, 2011; Lienert & Raatz, 1998; Rost, 1996). Diese Kriterien entstammen häufig der Messtheorie (Diekmann, 2012, S. 247) und wurden unter der besonderen Berücksichtigung der Zusammenhänge zwischen Reliabilität und Validität in die Testtheorie übertragen. Die hierarchisch aufgebauten Hauptgütekriterien quantitativer Forschung (vgl. Lienert & Raatz, 1998) sind dabei

- die Objektivität, also die intersubjektive Nachvollziehbarkeit und die Unabhängigkeit von der forschenden Person,
- die Reliabilität, also die Zuverlässigkeit, Replizierbarkeit und formale Genauigkeit, und
- die Validität, also die Gültigkeit.

[21] Es steht außer Frage, dass Methodenpluralität nützlich und notnwendig ist, um Evidenz zu sichern. Die Diskussion dieser geht aber über den Rahmen der vorliegenden Arbeit hinaus, diese Einschränkung wird in der Diskussion aufgegriffen.

[22] An verschiedenen Stellen werden stochastische Modelle und probabilistische Modelle gleichgesetzt, was zu sprachlichen Unschärfen führen kann, wenn beispielsweise von Modellen der Item-Response-Theorie und Faktormodellen gleichermaßen gesprochen wird. Stochastik meint die ‚Kunst des Vermutens‘, liefert also einen Rahmen für allgemeine Schätzverfahren, während probabilistische Aussagen Wahrscheinlichkeitsaussagen sind.

Auf die Nebengütekriterien, die an anderen Stellen bereits umfassend dargestellt sind (z.B. Döring & Bortz, 2016, S. 449), soll hier nicht näher eingegangen werden. Hingegen soll ein erweiterter Rahmen für die Hauptgütekriterien gegeben werden, denn auch wenn die Begriffe scheinbar fest arriviert sind, gibt es aktive Diskurse zu deren Bedeutung, Einschätzung, Gewichtung und Belegung, welche auch und insbesondere für Untersuchungen mit latenten Variablenmodellen eine hohe Bedeutung haben.

4.1. OBJEKTIVITÄT

Die Objektivität ist das Kernkonzept von Wissenschaft (Mulaik, 2004, S. 425). Ergebnisse sollten davon unabhängig sein, wer ein Objekt vorgibt, wer das Objekt interpretiert und wer das Objekt betrachtet (Rost, 1996, S. 38). Während der erste Punkt wissenschaftsethisch betrachtet werden muss und der zweite Punkt in quantitativer Tradition durch die formale Sprache hergestellt werden kann, verdient der dritte Punkt vor dem Hintergrund latenter Variablenmodelle eine vertiefte Aufmerksamkeit.

Da davon ausgegangen werden muss, dass jede Sichtweise subjektiv ist, wird der Wert von Erkenntnissen an Methoden und Standards des Forschens gemessen, welche die Basis inter-subjektiver Übereinstimmung bestimmen. Kant gibt dazu an, dass, wenn ein Urteil über ein Objekt nur in einer Person gefunden werden kann, es sich nur um eine Überzeugung handle und nur subjektiv Gültigkeit besitzen würde. Wahrheit basiere aber auf übereinstimmenden Urteilen zum Objekt, welche regelhaft zusammengeführt würden (Mulaik, 1995). Dies ist selbstverständlich keine absolute Determination der Wahrheit, sondern ein Weg, Fälle subjektiver Gültigkeit auszumachen. Für das Erreichen einer Übereinstimmung muss angenommen werden, dass ein gemeinsames Urteil semantisch und kognitiv möglich ist. Der Logik nach entspricht eine größere Zahl von Wahrnehmungen einer höheren Präzision, was einem Erfahrungsprozess entspricht. Die Summierung der Erfahrungen zu einem Objekt zu einer Wahrheitsaussage, kann durch mehrere getrennte Beobachtungen (Perspektiven) eines oder eine Wiederholung einer Beobachtungsform mehrerer Subjekte oder durch eine Kombination erfolgen. Der gleichen Logik folgt der Hypothesentest. Eine Hypothese, also die Annahme einer Regelhaftigkeit, wird basierend auf Daten aufrechterhalten oder verworfen. Daraus ergibt sich vorläufige Objektivität (Mulaik, 2004, S. 438). Die Objektivität muss vorläufig bleiben, da neue Daten generiert werden oder alternative Tests vorgenommen werden könnten. Wir verwenden Logik und Mathematik, um über Gemeinsamkeiten und Differenzen von Erfahrungen und Gemeinsamkeiten von Beobachtungen zu berichten.

In latenten Variablenmodellen werden die Indikatoren als einzelne Erfahrungswerte betrachtet, welche theoretisch kausal mit dem Objekt verknüpft sind. Die Kovarianz der Indikatoren ist die Übereinstimmung in der Beobachtung, welche je nach Modelltyp durch das Konstrukt erklärt wird.[23] Verbleibende Varianzen in Indikatoren werden als Messfehler (*true score* Modelle) oder Unschärfen (probabilistische Modelle) betrachtet, sind also je nachdem subjektive Verzerrungen oder ungemessene Eigenschaften der Subjekte.

[23] Die Indikatoren a priori allein ob hoher interner Konsistenz auszuwählen und von diesem Punkt aus auf Strukturen zu schließen ist irreführend und wird verzerrte Ergebnisse erzeugen (Little, Lindenberger & Nesselroade, 1999, S. 209).

Für statistische Modelle ist die minimale Anzahl von Beobachtungen und Erfahrungen durch Zahlen der Indikatoren und der Testteilnehmer gegeben. Die minimale Anzahl der Indikatoren pro latenter Variable wird durch die mathematische Identifizierbarkeit des statistischen Modells gesetzt und ist damit eine notwendige Bedingung für die Herstellung der Objektivität eines Modells. Nicht oder nur gerade identifizierte Modelle (vgl. Kap. 5.1) sind damit auf weitere Quellen angewiesen, um ihre Objektivität sicherzustellen.

Die nötige Anzahl der Testteilnehmer wird bei Hypothesentests über die verlangte Teststärke gegeben, also die Gefahr, einen Fehler zweiter Art zu begehen oder anders die Wahrscheinlichkeit, die Nullhypothese abzulehnen, wenn diese falsch ist (Cohen, 1977). Die Teststärke setzt sich aus dem Niveau des Alphafehlers (typischerweise $\alpha \leq .05$), der Größe der beobachteten Zusammenhänge und dem Stichprobenumfang zusammen. Ebenso sind die Parameterschätzungen und deren Standardfehler vom Stichprobenumfang abhängig. Zuletzt steht auch die Richtigkeit, also die der Eindeutigkeit (vgl. Kap. 5.1), der maximierten Wahrscheinlichkeit des Modells in Zusammenhang mit der Größe der Stichprobe (vgl. Gagne & Hancock, 2006). Somit werden die Anzahl der Erfahrungswerte (Items) und die Anzahl der Beobachtungen pro Erfahrungswert (Testteilnehmer) relevant. Generell kann festgehalten werden, dass Modelle mit mehr Indikatoren und großen Faktorladungen auf der Basis großer Stichproben immer eine höhere Wahrscheinlichkeit haben, rechnerisch zu konvergieren und eine bessere Ojektvität aufweisen. Die mindeste Zahl der Items kann über die Prüfung der Identifizierbarkeit sichergestellt werden. Diese wird vertiefend in Kapitel 5.1 vorgestellt. Für die Zahl der notwendigen Testteilnehmer existiert aber kein derart klarer Mechanismus.

Modellübergreifend gilt für latente Variablenmodelle ebenso wie generell für statistische Analysen, dass Analysen mit kleinen Stichproben weniger präzise Schätzungen, also größere Standardfehler, verursachen, die Teststärke sinkt und die Schätzungen weniger robust gegenüber Ausreißern sind (z.B. Fahrmeir et al., 2003). Die Zahl der notwendigen Testteilnehmer variiert insbesondere mit dem vorliegenden Modell und dessen Komplexität. Für SEM wird häufig eine optimale Zahl von 200 Testteilnehmern respektive fünf bis zehn Fällen pro zu schätzendem Parameter gegeben (Kline, 2011, S. 11–12). Generell wird davon ausgegangen, dass IRM eine höhere Anzahl von Testteilnehmern benötigen, um stabile Schätzungen der Item- und Personenparameter zu gewährleisten. Gleichzeitig ist es aber möglich, präzise Schätzungen im mittleren Fähigkeitenbereich auch bei kleiner Zahl von Testteilnehmern zu erreichen (vgl. Kap. 4.2). Dabei ist hier von erhöhter Relevanz, wie viele Item-Parameter zu schätzen sind. Bei Embretson und Reise (2000) finden sich grobe Empfehlungen, die von mindestens 50 Testteilnehmern für die einfachste Form, die 1PL-Modelle, über mindestens 500 für 2PL-Modelle, bis hin zu mindestens 1000 bei 3PL-Modellen. Bei politomen Indikatoren ändert sich dies dahingehend, dass Schwellen von der Anzahl der mindesten Beobachtungen pro Indikatorausprägung in Abhängigkeit zum Modelltyp ergänzt werden. Linacre (2002) nennt hier eine Zahl von mindestens zehn Beobachtungen. Wurpts und Geiser (2014) stellten in Analysen zu der notwenigen Qualität und der Anzahl der Items in LPA und LCA sowie den optimalen Stichprobenumfängen fest, dass alle drei Größen interagieren und damit die Schwelle der rechnerischen Lösbarkeit (und der Qualität) der Modelle bedingen. Während bei Finch und Bronk (2011) noch ein Stichprobenumfang von etwa 500 festgehalten wird, empfehlen Wurpts

und Geiser (2014) gleitende Übergänge bei einem mindesten Stichprobenumfang von 100, gegeben, dass gute Items (vgl. Kap. 5.2.3) in ausreichender Zahl und gegebenenfalls stabilisierende Kovariablen vorliegen. Auf die benötigte Teststärke von latenten Variablenmodellen und demnach die Ableitung des notwenigen Stichprobenumfangs kann über Monte-Carlo-Simulationen geschlossen werden (Muthén & Muthén, 2002).

Die Objektivität hat bei der Anwendung von vollstandardisierten psychologischen Tests oder Leistungstestbatterien nur eine geringe Bedeutung[24], da sie durch das Befolgen derer Normierung und die darauf aufbauenden Testmanuale gegeben ist (solange diese befolgt werden; Döring & Bortz, 2016, S. 442). Wie aufgezeigt wurde, hat sie aber eine gewichtige Bedeutung für die Formulierung, die Identifikation und die Evaluierung von latenten Variablenmodellen. Die Abhängigkeit von anderen Dingen als Beobachtungen und Erfahrungen, also zum Beispiel der Testsituation, wird maßgeblich im Bereich der Validität unter dem Begriff der Generalisierbarkeit aufgegriffen. Eine Objektivität im Sinne der Unabhängigkeit von anderen Bedingungen ist die spezifische Objektivität (vgl. Embretson & Reise, 2000; Rost, 1996). Diese besagt, dass das Ergebnis unabhängig von der Auswahl der Indikatoren ist. In den meisten gängigen quantitativen Testmodellen gilt, dass die Messwerte spezifisch objektiv sind (Rost, 1996, S. 38). Dies gilt aber nur, wenn das Modell hinreichend zuverlässig, also reliabel ist.

4.2. RELIABILITÄT

Die Reliabilität beschreibt die Zuverlässigkeit eines Tests, aber im engeren mathematischen Sinne beschreibt sie die Messgenauigkeit (Rost, 1996, S. 34), also wie gering oder stark ein Test durch einen Messfehler verzerrt ist oder wie konsistent die Ergebnisse sind. Mueller und Hancock (2008, S. 505) geben dazu an, dass latente Konstrukte theoretisch immer perfekt reliabel sind, aber imperfekte Messungen aufweisen. Dies kann Eigenschaft des Tests, des untersuchten Konstrukts, aber auch der Testteilnehmer sein, wenn zum Beispiel Aufgaben nicht verstanden oder Antwortmuster absichtlich verzerrt werden. Die Reliabilität wird für FA, IRM und LCA/LPA auf unterschiedliche Weisen bestimmt, daher wird in den Standards for Psychological and Educational testing (American Educational Research Association, American Psychological Association & National Council on Measurement in Education, 2014) neben dem klassischen Reliabilitätsbegriff der Begriff der Präzision aufgeführt, insofern die Konsistenz des Tests gemeint ist. Die Reliabilität oder Präzision ändert sich mit der Variabilität von Testscores über mehrere Replikationen der Tests und die Analysen dazu basieren auf der Variabilität der Items, Inhalte oder Testteilnehmer. Aber neben diesen quantitativen Dimensionen muss ebenso das Prinzip der Modellsparsamkeit oder der Parsimonität als qualitatives Gütemerkmal, welches mit der Zuverlässigkeit von Tests verhaftet ist, eingeführt werden. Das Prinzip besagt, dass zur Begründung oder Erklärung von Zusammenhängen so wenig Variablen und Annahmen wie nötig herangezogen werden sollten, um diese mathematisch und valide ausreichend aufzuklären. Die Logik hinter diesem Prinzip ist, dass eine Falsifikation weniger hypothetischer Annahmen leichter ist (Döring & Bortz, 2016, S. 57).[25]

[24] In der Entwicklung ist die Relevanz natürlich sehr hoch.
[25] Die Diskussion zu diesem Prinzip findet sich in dem Ansatz von Ockhams Rasiermesser begründet.

Faktormodelle

Für essenziell messfehlerbasierte Modelle wie die FA, wird zur Bestimmung der Reliabilität dieser Messfehler zentriert. Eine klassische Definition der Messgenauigkeit setzt hier die wahre Varianz Var_T in ein Verhältnis zu der beobachteten Varianz Var_E

$$Reliabilität = \frac{Var_T}{Var_E}.$$

Da die Varianz des Erwartungswerts per Definition immer größer ist als die Varianz des wahren Werts (vgl. Kap. 3), liegt die Reliabilität in dieser Operationalisierung immer zwischen 0 und 1.[26]

Übliche Verfahren zur Schätzung der Reliabilität sind die Test-Retest-Reliabilität, die Paralleltest-Reliabilität, die Interrater-Reliabilität und die geläufigste Methode, die Bestimmung der internen Konsistenz, welche zumeist über den Cronbach Alpha vorgenommen wird (z.B. Döring & Bortz, 2016, S. 444). Weiterhin existiert die Möglichkeit der Ableitung der Reliabilität über die Bestimmung einer Split-Half-Reliabilität, welche aufgrund verschiedener Probleme bezüglich der Verkürzung der Testlänge (s. u.; Yousfi & Steyr, 2006) nur noch historische Bedeutung hat. Die Test-Retest-Reliabilität wird dabei bestimmt als die Korrelation zweier zu unterschiedlichen Zeitpunkten erhobener Messwertreihen, die Paralleltest-Reliabilität als die Korrelation der Messwerte mit einem Testzwilling oder gegebenenfalls Itemzwillingen und die Interrater-Reliabilität durch die Korrelation mehrerer Urteilender. Der Cronbach Alpha ist definiert als die mittlere Split-Half-Reliabilität aller split-halfs, es wird also jedes Item wie ein Paralleltest behandelt und die Werte werden zusammengefasst (Döring & Bortz, 2016, S. 469). An verschiedenen Stellen wurden Schwellenwerte für die Interpretation von Reliabilitätsanalysen (z.B. .7; Kline, 2011, S. 70) vorgeschlagen, aber die Höhe der Koeffizienten ist von verschiedenen Faktoren abhängig, die in keinem Zusammenhang mit der Zuverlässigkeit eines Tests stehen (Cortina, 1993; Dunn, Baguley & Brunsden, 2014; Lord & Novick, 1968; Raykov, 1997; Rost, 1996, S. 356; Schmitt, 1996). Beispielsweise weist die Testlänge, also die Anzahl der homogenen Items, einen direkten Zusammenhang mit der Reliabilität auf. Für Test-Retestanalysen kann zudem festgehalten werden, dass die Länge der zeitlichen Intervalle, in denen eine Testwiederholung stattfindet, einen Einfluss auf den Wert hat. Zudem senkt die inhaltliche Heterogenität der Items den Wert, auch wenn diese Variabilität theoretisch angemessen ist. Somit wird an dieser Stelle davon Abstand genommen, Empfehlungen für konkrete Schwellenwerte zu wiederholen.

Generell sind diese zusammengefassten Werte vor dem Hintergrund hochflexibler Faktormodelle häufig nicht hinreichend. Denn die Qualität der Messungen, in Wiederholung oder zwischen Items, darf in diesen bedingt variieren, und (Residual-)Varianzheterogenität kann vorkommen. Und von der Homogenität mehrerer Messungen hängt ab, welche konkrete Formel zur Schätzung des Modells verwendet werden kann. Daher existieren im Rahmen der Klassischen Testtheorie verschiedene Messmodelle. In groben Kategorien handelt es sich um das

[26] Es gibt verschiedene, situativ alternative, Ansätze zur Bestimmung der Reliabilität, beispielsweise das relative Verhältnis von der Varianz zwischen Gruppen und innerhalb von Gruppen, bei Mehrgruppenvergleichen, z.B. Shrout und Fleiss (1979), bei denen die Reliabilität über *F*- Werte oder χ^2-Werte abgeleitet wird. Dabei können auch Werte außerhalb dieser Wertespannweite vorkommen.

parallele, das τ–äquivalente und das kongenerische Messmodell (Bühner, 2011; Steyer & Eid, 2001). Wobei die Modelle steigend weniger restriktiv sind und demnach mehr Parameter nötig sind, um das Modell zu identifizieren (vgl. Kap. 5.1). Zusammengefasst wird

- im parallelen Modell angenommen, dass die Messungen (also die Faktorladungen der einzelnen Indikatoren) identisch sind und die Varianzen der Messfehler der Indikatoren gleich sind.
- Im τ –äquivalenten Modell dürfen diese Messfehler variieren, im essenziell τ –äquivalenten Modell dürfen diese Messfehler variieren, und die Ladungsgewichte dürfen um additive Konstanten über alle jeweiligen Indikatoren hiweg verschoben sein, und
- im kongenerischen Modell dürfen die Ladungen und die Messfehler variieren, so lange die Ladungen linear ineinander überführbar sind (Bühner, 2011, S. 147–151).

Beispielsweise muss bei der Erstellung vom Summenscores von Indikatoren, einem üblichen Vorgehen im Rahmen der klassischen Testtheorie, angenommen werden, dass mindestens τ-Äquivalenz vorliegt, da jeder Indikator exakt mit dem gleichen Gewicht in den Wert eingeht. Bis zu diesem Modelltyp - oder korrekter, falls die Annahmen des τ–äquivalenten Modells gelten - ist die interne Konsistenz ein angemessenes Reliabilitätsmaß. Für kongenerische Modelle hingegen ist die Bestimmung der internen Konsistenz als Reliabilitätsmaß aufgrund der Verletzung der Annahme der τ-Äquivalenz nicht mehr angemessen (Döring & Bortz, 2016, S. 468). Das konfirmatorische Faktormodell (CFA) ohne Gleichheitsrestriktionen zwischen den Parametern entspricht dabei dem kongenerischen Modell und kann auf seine Gültigkeit getestet werden, die Verwendung eines Scores wäre also dann angemessen, wenn das Modell positiv evaluiert wurde (vgl. Kap. 5.2). Damit werden die Begrenzungen der Klassischen Testtheorie für den Umgang mit theoretischen Konstrukten im Ansatz der CFA erweitert. Da im Rahmen von Modellen der KTT keine Trennung vom beobachteten Wert und dem *true score* angenommen werden kann, wird ein Summenscore oder ein einfaches Aggregat der beobachteten Variablen als *true score* betrachtet werden. Die Eigenschaften der Items haben somit nur für die vorliegende Stichprobe Gültigkeit. In der CFA wird die individuelle Ausprägung von latenten Variablen hingegen von den Eigenschaften der Items getrennt, womit eine Vergleichbarkeit über verschiedene Stichproben, Gruppen und die Zeit möglich wird, da angenommen und geprüft werden kann, dass nur der individuelle Anteil variiert. Dieser Vorteil gilt für IRM und LCA/LPA ebenso.[27]

Alternative Vorschläge zur globalen Prüfung der Reliabilität in Messmodellen mit geringeren Restriktionen oder gelockerten Annahmen oder bei mehrdimensionalen Tests gab es zum Beispiel bei Raykov (1997) über die Composite Reliability oder durch die Formulierung eines Omega-Wertes (McDonald, 1999), sowie dessen Weiterentwicklungen (vgl. Dunn et al., 2014). Der Omega-Wert Ω wird analog zum Alpha-Koeffizienten bestimmt als das Verhältnis der quadrierten Summe der Faktorladungsgewichte β zu der quadrierten Summe der Faktorladungsgewichte sowie deren Residuen ε. Unter der Annahme der τ–Äquivalenz ist der Omega-Wert gleich dem Alpha-Wert.

[27] Diese Stichprobenunabhängigkeit wird häufig als Merkmal von IRM hervorgehoben. Aber auch in Modellen der Faktoranalyse kann diese angenommen und geprüft werden (vgl. Kap. 5.2.3).

$$\Omega = \frac{(\beta_1 + \beta_2 + \cdots + \beta_n)^2}{(\beta_1 + \beta_2 + \cdots + \beta_n)^2 + (\varepsilon_1 + \varepsilon_2 + \cdots + \varepsilon_n)}$$

Der Omega-Wert wird aber praktisch selten berichtet oder verlangt, da es häufig informativer ist, direkt die Ladungsstruktur zu inspizieren.

Die Feststellung der internen Konsistenz von latenten Faktormodellen wird also maßgeblich über die Prüfung einzelner Modellparameter und deren Homogenität vorgenommen. Diese kann getestet werden, indem Modelle mit und ohne Gleichheitsrestriktionen auf den Faktorladungen und Residuen gegeneinander geprüft werden. Somit wird die Reliabilität hier über die Inspektion der Ladungsstruktur der Indikatoren für jedes Teilmodell und im Zusammenspiel von, soweit vorhanden, Teilmodellen vorgenommen.[28]

Item-Response-Modelle

Da die Item-Response-Modelle einer anderen Forschungstradition entstammen und kein Messfehler im faktoranalytischen Sinne herangezogen werden kann, können keine messfehlerzentrierten Ansätze zur Bestimmung der Reliabilität verwendet werden. Die Test-Retest-Korrelation kann trotzdem zur Bestimmung der Reliabilität verwendet werden, ebenso wie die Split-Half-Reliabilität, aber die Probleme in der Anwendung (Aufwand, Abstand der Messzeitpunkte, Testverkürzung) verbleiben.

Zentral für die Feststellung der Reliabilität von IRM ist die Feststellung der Präzision, mit der ein Testwert ermittelt wird. Diese differiert dabei je nach der Lage im Leistungskontinuum, die Genauigkeit der Messung ist im mittleren Bereich besser als im extremen Bereich (Rost, 1996, S. 353), was die Schätzung einer absoluten Messgenauigkeit unmöglich macht, sondern nur eine Schätzung für spezifische Fähigkeitsbereiche erlaubt. Diese Randunschärfen resultieren aus erhöhten Standardabweichungen der Standardfehler in den Extrembereichen, da diese durch die gegebene Normalverteilung der Personenparameter geringer besetzt sind. Dies ergibt sich daraus, dass die Varianz eines dichotomen Items x_i aus der Wahrscheinlichkeit des Vorkommens eines Wertes $X_i = 1$ und der Gegenwahrscheinlichkeit (Embretson & Reise, 2000; Rost, 1996) bestimmt wird, also gilt

$$Var(x_i) = p(x_i = 1) * p(x_i = 0).$$

Daraus lässt sich ableiten, dass die Varianz immer größer ist, wenn sich die Lösungswahrscheinlichkeit der mittleren Lösungswahrscheinlichkeit annähert (z. B. $0.5 * 0.5 = 0.25$, während $0.3 * 0.7 = 0.21$). Der Messwert einer Person ist demnach umso genauer, je weniger der Personenparameter vom wahren Parameter abweicht, also je kleiner die Varianz der Fehlervariable, des Kehrwertes der Itemparametervarianz, ist. Dasselbe gilt für mehrkategoriale Antwortformate. Die globale Reliabilität eines Tests ist somit definiert als die Relation zwischen dem Produkt der Varianz der wahren Messwerte $\bar{\theta}$ und dem Stichprobenumfang N und der Summe aller Personenmessfehler E_{θ_v}

[28] Hancock und Mueller (2011) stellen darüber hinaus ein Verfahren vor, bei dem Messmodell und Strukturmodell getrennt voneinander geprüft werden können.

$$Rel(\Theta) = \frac{\sum_{v=1}^{N} Var(E_{\theta_v})}{N * Var(\bar{\theta})}.$$

Diese ist aber stark durch die Anzahl der Items beeinflusst. Eine gute Passung kann zudem bedeuten, dass nicht etwa der Test perfekt ist, sondern eine mangelhafte Differenzierung stattfindet, also alle Personenparameter sehr nah beieinander liegen (vgl. Kap. 5.2.3). Eine perfekte Reliabilität von 1 kann beispielsweise erreicht werden, wenn angenommen wird, dass das Mittel der wahren Messwerte 0.5 und alle Personenmessfehler 0.5 seien, also de facto vollständige Zufälligkeit vorliegt. Diese mangelnde Differenzierung kann die Annahme zu der (Normal-)Verteilung der latenten Variable verletzen. Über die Hinzugabe von Items mit mittleren Schwierigkeiten oder die Reduktion von Items mit extremen Schwierigkeiten kann dieses Reliabilitätsmaß also stark beeinflusst werden, ohne dass eine tatsächliche Verbesserung stattfindet (Rost, 1996).

Aufgrund der vorgestellten Eigenschaften der Varianzverteilungen in IRM ist es üblich, anstatt eines absoluten Maßes der Reliabilität lokale Item- und Personenparameter zu prüfen, um auf die Reliabilität zu schließen (vgl. Kap. 5.2.3). Eine Inspektion der Item-Person Map oder Wright Map (Wilson, 2003), also der visuellen Prüfung der Verteilung der Schwierigkeitsparameter und parallel der Verteilung der Personenparameter, kann verwendet werden, um eine verteilungsbasierte Zusammenführung dieser lokalen Parameter vorzunehmen. In der Abbildung 3 ist eine Wright Map mit simulierten Beispieldaten eines eindimensionalen Raschmodells abgetragen. Beide Parameter liegen auf derselben Metrik, und die Verteilung der Personenparameter kann direkt mit der Verteilung der Schwierigkeitsparameter verglichen werden. Erwartet wird eine ähnliche Verteilung der Personenfähigkeit und der Schwierigkeitslagen der Items.

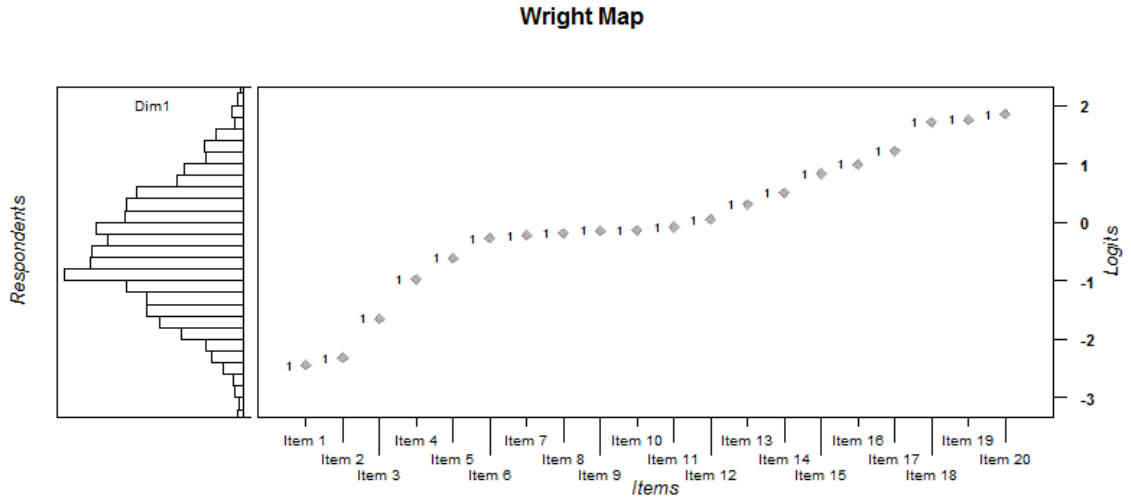


Abbildung 3: Wright Map[29]

[29] Die Abbildung wurde erstellt mit dem *R* Paket Wright Map (Torres Irribarra und Freund, 2014).

Latente Klassenmodelle

Die Messgenauigkeit ist für latente Klassenmodelle ebenfalls direkt mit dem Begriff der Präzision, hier der Exaktheit einer Zuordnung zu einer spezifischen Klasse, verbunden (American Educational Research Association et al., 2014). Die Präzision oder Zuordnungssicherheit ist allgemein durch die Wahrscheinlichkeit der Klassenzugehörigkeit bei gegebenem Antwortmuster eines Testobjekts determiniert. „Den Messfehler verringern heißt bei qualitativen Testmodellen die Zuordnungssicherheit erhöhen." (Rost, 1996, S. 361). Die Evaluation der Reliabilität erfolgt also über die Antwortmuster respektive die Antwortmusterwahrscheinlichkeiten innerhalb der abgeleiteten Klassen, die Klassenzugehörigkeitswahrscheinlichkeiten $p(v \in g|a_v)$ (vgl. Kap. 3.1) und dem Abgleich mit dem Zufall. Die mittlere Klassenzugehörigkeitswahrscheinlichkeit aller Klassen entspricht am ehesten einem klassischen Reliabilitätswert und ergibt sich als Mittelwert der Diagonalen auf einer Matrix der wahrscheinlichsten Klassenzugehörigkeiten und den latenten Klassen (ebd.).

Tabelle 3: Klassenzugehörigkeitswahrscheinlichkeiten

		Latente Klasse			
		1	2	3	∑
Mittlere	1	**.78**	.18	.04	1
Klassenzugehörigkeits-	2	.15	**.85**	.00	1
wahrscheinlichkeit	3	.19	.01	**.80**	1

Sowohl für dieses Maß als auch für die Reliabilitäsmaße der IRM, muss klargestellt werden, dass sie nur als Pseudoreliabilitätsmaße begriffen werden können. Für LCA/LPA entsprechen diese der Treffsicherheit in deterministischen Cluster- oder Klassifikationsverfahren, für IRM entsprechen sie der mittleren Personen- oder Itemschärfe. Die Ableitung dieser Reliabilitätsmaße zur Beschreibung des mittleren Fehlermaßes von latenten Klassenmodellebn kann in den Beiträgen von Schurig und Busch (2014) und Schurig et al. (2015) nachvollzogen werden.

4.3. VALIDITÄT

„Die Validität einer Messung bezieht sich auf die Frage, ob das gemessen wird, was gemessen werden sollte." (Friedrichs, 1982, S. 100). Sie stellt den Wahrheitsgehalt einer Schlussfolgerung dar (Shadish et al., 2002, S. 34) und ist das wichtigste der drei Gütemerkmale. Im Kontext quantitativer Tests bedeutet Validität, dass eine inhaltliche Erweiterung objektiver und reliabler Zahlenwerte erfolgen kann. Eine hohe Reliabilität und eine gegebene Objektivität ist dabei Voraussetzung, aber keine hinreichende Voraussetzung für eine hohe Validität. So kann ein Test, der nicht hinreichend reliabel ist, niemals valide sein. Die Validität bleibt immer durch die Mängel in der Reliabilität eingeschränkt. Dabei ist das Konzept der Validität im empirisch-sozialwissenschaftlichen Kontext schwer zu belegen, da im Gegensatz zu einem Feld wie der formalen Logik keine präzise technische Definition formuliert werden kann.[30] Dabei hat es

[30] Newton und Shaw (2013) empfehlen sogar, den Begriff der Validität im Kontext der Anwendung von Tests vollständig durch den der Qualität zu ersetzen.

umfassende Versuche gegeben, diese zu formulieren. Hier ist ein Rückblick auf die Genese des Validitätsbegriffs hilfreich.

Historisch betrachtet können auf Grundlage der Standards der American Educational Research Association (AERA), American Psychological Association (APA) und das National Council on Measurement in Education (NCME) verschiedene Etappen ausgemacht werden, auf deren Basis über Tests gedacht werden kann. AERA, APA und NCME erstellten in Anlehnung an Standards der APA (1954) in den Standards for Educational and Psychological testing[31] (1999, 2014; 1966, 1974; 1985) Orientierungshilfen und Regularien zum Umgang mit möglichen Kriterien. Bereits (1954) wurde festgehalten, dass Validität multikriterial ist und klar aufgezeichnet werden sollte, welches Kriterium adressiert wird (American Psychological Association et al., 1954, S. 18–19). Ergänzend wurden Standards für die Feststellung von Validität gegeben (American Psychological Association et al., 1954, S. 18–28). 1966 wurde ergänzt, dass generelle Validität nicht angenommen werden kann.

„C1.1. Statements in the manual about validity should refer to the validity of particular interpretations or of particular types of decision. ESSENTIAL [Comment: It is incorrect to use the unqualified phrase “the validity of the test.” No test is valid for all purposes or in all situations or for all groups of individuals.]” (American Psychological Association et al., 1966, S. 15)

Besonders hervorzuheben ist hier die unmissverständliche Aussage, dass es inkorrekt ist, von der Validität eines Tests zu sprechen. Dies hat in ähnlicher Formulierung in den Standards bis heute Bestand. Hingegen wurden partielle Annahmen zu einzelnen Dimensionen der Validität als wünschenswert angesehen, namentlich zu der Inhaltsvalidität, der Kriteriumsvalidität und der Konstruktvalidität (American Psychological Association et al., 1966; 1966; Lienert & Raatz, 1998). Zudem wurde festgehalten, dass die Inhaltsvalidität[32], also Trennschärfe, besonders Gewicht für Leistungstests hat, die Kriteriumsvalidität insbesondere für Eignungstests wichtig wäre und die Konstruktvalidität, also die theoretische Fundierung, für Persönlichkeitstests hervorzuheben sei (z.B. Rammstedt, 2010). Wie aber sollten diese zu prüfen sein?

Ein Ansatz zur Prüfung ist es, verschiedene Formen internaler und externaler Validität zu bestimmen, wie sie beispielsweise in psychologischer Tradition von Campbell vorgeschlagen und von Cook und Campbell (1979) ergänzt wurde. Eine umfangreichere Vorstellung der Teilaspekte, inklusive weiterer Bedrohungen derer, finden sich bei Shadish et al. (2002). Diese Differenzierungen sind dabei explizit nicht in Abgrenzung zu den Standards gedacht, sondern werden als Elaboration dieser verstanden (Cook & Campbell, 1979).

[31] In der Folge nur als Standards bezeichnet.

[32] Probleme bei der begrifflichen Trennung von Konstruktvalidität und Inhaltsvalidität sowie Probleme mit der Praxis theoretisch unreflektierter Anpassungen in Skalen allein auf der Basis von Trennschärfeparameter werden beispielsweise bei Rossiter (2008) diskutiert.

1. Die Konstruktvalidität beschreibt die Repräsentation der theoretischen Konstrukte durch die Messinstrumente und die Untersuchungsbedingungen. Dies meint vor allem die Funktionalität des Messmodells, welches inferenzstatistisch getestet werden kann (Hartig et al., 2008, S. 150). Die Konstruktvalidität wird vor allem durch die Theoriearbeit, die Konzeptualisierung und Operationalisierung gegeben. Bedrohungen sind beispielsweise Konfundierungseffekte, Verzerrungen durch fehlerhafte Operationalisierungen oder Antworttendenzen wie die Akquieszenz.
2. Die externe Validität beschreibt die Generalisierbarkeit von beobachteten Effekten. Dies ist vor allem vom Untersuchungsdesign und der Stichprobe abhängig. Bedrohungen sind beispielsweise unzureichende Stichproben, mangelhafte Prüfungen der Messinvarianz (vgl. Kap. 5.2.3) oder die mangelnde Berücksichtigung von Mediatoren.
3. Die Validität statistischer Schlüsse beschreibt die Qualität der Datenanalyse und der Messgenauigkeit. Bedrohungen sind beispielsweise zu geringe Teststärken, mangelnde Reliabilität von Instrumenten und Störeinflüsse.
4. Die interne Validität beschreibt, wie zweifelfrei Effekte belegt werden können. Dies hängt vor allem mit dem Untersuchungsdesign zusammen. Bedrohungen sind beispielsweise Selektionseffekte, Regressionseffekte „zur Mitte“ oder Panelmortalität.

Forschungslogisch ist die Konstruktvalidität dabei ein Teil der externe Validität und die statistische Validität ein Teil der internen Validität (Campbell, 1957). Alle diese Kriterien beschäftigen sich mit der methodischen Strenge quantitativer Forschung.

Diese Konzeption in den Standards und deren prominente Vertiefung führten zu verschiedenartigen Problemen. So kritisierte Cronbach (1982), dass die Schemata der Standards und Campbells zu wenig praxisorientiert seien und sich schlecht auf Evaluations- und Interventionsforschung übertragen lassen, also aufgrund der Herkunftstradition übermäßig auf experimentelle und quasi-experimentelle Studien zugeschnitten seien. Daher zentriert er auf das Konzept der Generalisierbarkeit in den vier Dimensionen der (1) Untersuchungseinheiten, der (2) untersuchten Effekte, der (3) Beobachtungen und des (4) Forschungssettings. Zudem stehen die Gewichtung und die Auswahl der zu prüfenden Kriterien dabei immer in direkter Verbindung zu der Tragweite der Forschungsbemühung. Beim high-stakes Testing (vgl. Ryan & Sapp, 2005) oder in der Diagnostik (vgl. Bühner, 2011) werden andere Maßstäbe und Anforderungen gültig, als beispielsweise bei dem Vergleich zweier Aggregate, welcher keine Generalisierbarkeit annimmt und keine Folgen für die Testobjekte hat.[33] Messick (1995) ergänzte und etablierte zu der Problematik einer möglichen ethischen Dimension des Tests die Kriterien der inhaltlichen (a) Relevanz, (b) Nützlichkeit, (c) Wertimplikationen und (d) sozialen Konsequenzen eines Tests.

Die Fragmentierung des Begriffs, ebenso wie die aufgeführten Anwendungsprobleme führten zu einer sprunghaften Entwicklung von einer großen Zahl von Maßeinheiten und einem regelhaften und unreflektierten Umgang mit einzelnen (punktuell sicherlich angemessenen) Maßen

[33] Dies wird z. B. auch unter dem Begriff der consequential validity diskutiert (Hartig, Frey und Jude , 2008, S. 158f) und betrifft direkt die wissenschaftsethische Frage nach der politischen Verantwortung von Forschenden.

(Messick, 1975; 1981) für Validität. In quantitativer und qualitativer Forschungstradition, ebenso wie beispielsweise in der Gesetzgebung und dem Management wurden eine Vielzahl von Konzepten entwickelt, welche von Newton und Shaw (2013) exemplarisch zusammengetragen wurden. Die Übersetzung einer erweiterten und annotierten Liste von Newton (2013) befindet sich in Anhang 1 und verdeutlicht ob ihres Umfangs die Problematik der fragmentierten Sichtweise.

In direkter Reaktion auf das inflationäre Auftreten dieser Maße sowie insbesondere der Kritik durch Messick (1981) wurde das Konzept der Validität in den folgenden Standards (1985) als einheitlich definiert. Das Konzept sollte durch die Heranführung von inhaltlicher, kriterialer und konstruktbezogener Evidenz (vgl. Kap. 3.3) belegt werden. De facto wurde damit die Konstruktvalidität zum einzig relevanten Konzept (Newton & Shaw, 2013, S. 303) und die Bezeichnung von Einzelkriterien als Dimensionen von Validität zurückgewiesen. 1999 wurde dies in den Standards dahingehend weitergeführt, dass die Validität (ebenso wie die Reliabilität) als Funktionen der Interpretation von Tests im speziellen Anwendungsfall zu verstehen sind und nicht dem Test zugeordnet werden sollten. „[…] all test scores are viewed as measures of some construct […] the validity argument establishes the construct validity of a test." (American Educational Research Association et al., 1999, S. 174). Zudem wurde betont, dass der Validierungsprozess ein wissenschaftlicher Prozess ist, da nicht ein Score, sondern die Interpretation des Score das Ergebnis darstellt (American Educational Research Association et al., 1999, S. 9).

Dies greift die mangelnde Verknüpfung mit theoretischen Annahmen auf. Die Standards liefern einen starken semantischen und konzeptionellen Rahmen für Testentwicklung und Testanwendung, aber die Positionierung der Theorie für die Bewertung der Validität bleibt auch dabei vage. Bereits Lord und Novick (1968) führten in ihrem richtungsweisenden Herausgeberband zu IRM aus, dass empirische und theoretische Validität zu trennen und jeweils hinreichend zu begründen seien. Im Rahmen der Standards ist diese aber eben nur als prozedurale Evidenz des Testinhalts gedacht (American Educational Research Association et al., 1999). Borsboom, Mellenbergh und van Heerden (2004) formulierten hierbei in Abkehr zu den Standards eine radikalere Betrachtungsweise, welche Validität ausschließlich an der technischen Evaluation des Tests, also der stochastisch-kausalen Verknüpfung der Testergebnisse und der zu messenden Attribute sowie der korrekten Spezifikation des Modells anknüpften. Wenn ein eindimensionales latentes Modell eine hinreichende Modellanpassung aufweist, also ein hinreichender Grad lokaler Unabhängigkeit nachgewiesen werden kann, kann dies als Evidenz für die Hypothese verstanden werden, dass eine latente Variable existiert, welche wiederum Variation in beobachteten, manifesten Variablen verursacht (Borsboom et al., 2003). Es wird induktive Unterstützung für die Hypothese eines hinreichenden gemeinsamen Grundes erbracht, was als Evidenz für Validität verstanden werden kann (Borsboom et al., 2004). Später argumentieren Markus und Borsboom (2013, S. 281) in direkter Bezugnahme auf die mangelnde Einbindung der Theorie in die Prüfungen der Validität, dass die Basis der Validitätsprüfung eines Tests voraussetzt, dass eine Theorie angenommen wird (egal wie minimal diese ist; ebd.) und diese nicht erst die Interpretation eines Test Scores ermöglicht (American Educational Research Association et al., 1999, S. 9). Erst die Annahme einer Theorie

legitimiert demnach die Herleitung und die Interpretation eines Testwerts in einer spezifischen Art. Diese Legitimierungen sind wiederum in Einklang mit den Standards, als relativ in Bezugnahme auf die Kommunikation mit Adressaten, den methodologischen Standards für Evidenz des Feldes und den verfügbaren Handlungsalternativen zu begreifen. Die Validität wird also stärker auf die Erkenntnis bezogen begriffen.

2014 wurden die Standards noch einmal um Rahmenkonzepte erweitert, nämlich die möglichen Formen und die Quellen von Evidenz für Validität. Die Formen können in empirische und prozedurale Evidenz unterschieden werden (vgl. Haladyna, 2006), wobei die erste Form die Post-hoc-Prüfungen und die zweite Form die Planung und das Vorgehen im Untersuchungsvorhaben meint. Als mögliche Quellen für die Evidenz werden der Testinhalt, die Internale Struktur, der Antwortprozess und das Verhalten gegenüber anderen Maßen genannt (American Educational Research Association et al., 2014, S. 14ff). Für alle Quellen werden Hauptbedrohungen identifiziert und prozedurale Vorgehensweisen empfohlen. Zusammenfassend wird in den aktuellen Standards festgehalten:

> „Validity refers to the degree to which evidence and theory support the interpretation of test scores for proposed uses of tests. […] The process of validation involves accumulating relevant evidence to provide a sound scientific basis for the proposed score interpretations." (American Educational Research Association et al., 2014, S. 11).

Trotzdem werden Begriffe wie Kriteriumsvalidität regelmäßig verwendet und in der Form von Punktschätzern in Reviewverfahren verlangt (Newton & Shaw, 2013, S. 308). Newton und Shaw (2013, S. 304) interpretieren das Verbleiben von traditionellen Gütekriterien als eine Strategie bewusst ungenauer Sprachweise, welche die Kommunikation über diese Kriterien erleichtert. Kline (2011) schreibt beispielsweise bewusst von Score Validity und entkoppelt den Begriff vom Test, ohne ihn aber vollständig an der Schlussfolgerung anzugliedern, da er in der Folge auf Maße wie die Kriteriumsvalidität von Testscores abhebt. Kritisch ausgedrückt erschwert diese Form der Umschiffung der Probleme um den Begriff und das Belegen der Validität - sowohl im wissenschaftlichen Kontext, als auch im Ergebnistransfer in die Praxis -, und es wäre hilfreich, klare und anwendbarere Konventionen für den sprachlichen Umgang mit der Evidenz für Validität zu formulieren.

Wie aber können verschiedene Formen der Evidenz für Validität nun sichergestellt werden? In der Tradition der Standards kann die Validität des Testinhalts über die Auswahl der Indikatoren sichergestellt werden und ist weitgehend analog zu einer strengen theorie-basierten Skalenkonstruktion (Rammstedt, 2010, S. 250). Die Kriteriumsvalidität wird durch den Abgleich mit externen Maßen (z. B. Leistungstestergebnisse mit einem Lehrerurteil) sichergestellt.[34] Die Konstruktvalidität kann wiederum durch Dimensionsanalysen, den Abgleich mit alternativen Instrumenten oder den Abgleich mit alternativen Urteilen (z. B. Selbst- vs. Fremdurteil oder Messwiederholung) sichergestellt werden. In der Tradition nach Campbell weicht dies leicht ab.

[34] Hierbei gilt, dass die Validität nie höher sein kann, als das geometrische Mittel der Reliabilitätskoeffizienten, da die Messunsicherheit mit verarbeitet werden muss (z. B. Rammstedt, 2010, S. 251).

Während beispielsweise die statistische und interne Validität gut über die technische Umsetzung und die Anwendungen von State-of-the-art Techniken erreicht werden können, sind Validierungsstrategien für die Konstruktvalidität und die externe Validität komplexer. Ein Ansatz zur technischen Konstruktvalidierung ist es, die Kriteriumsvalidität, die konvergente und/oder diskriminante Validität (Campbell & Fiske, 1959; Kline, 2011, S. 71) sowie die faktorielle Validität (Döring & Bortz, 2016, S. 446) zu belegen. Die Kriteriumsvalidität beschreibt in diesem Fall einen Abgleich mit einem oder mehreren externen Maßen, welche eine vergleichbare Größe bestimmen. Die konvergente Validität beschreibt, dass ein Messmodell nur ein spezifisches gegebenes Konstrukt widerspiegelt und kein anderes. Die diskriminante Validität beschreibt die Distinktheit des Konstruktes. Beide können bestimmt werden, indem Vergleiche mit alternativen Messmodellen vorgenommen werden. In Anlehnung an Skrondal und Rabe-Hesketh (2004) sind in der Abbildung 4 zwei hypothetische latente Strukturen abgetragen, anhand derer Hinweise auf konvergente und diskriminante Validität geprüft werden können. Im Fall des unteren Modells wurde beispielhaft festgestellt, dass ein Indikator ($x4'$) durch beide latente Variablen determiniert wird und die konvergente Validität somit nicht angenommen werden kann. Im gleichen Modell ist die Korrelation zwischen den latenten Dimensionen größer, als theoretisch angenommen werden kann, sodass auch die diskriminante Validität nicht angenommen werden kann.

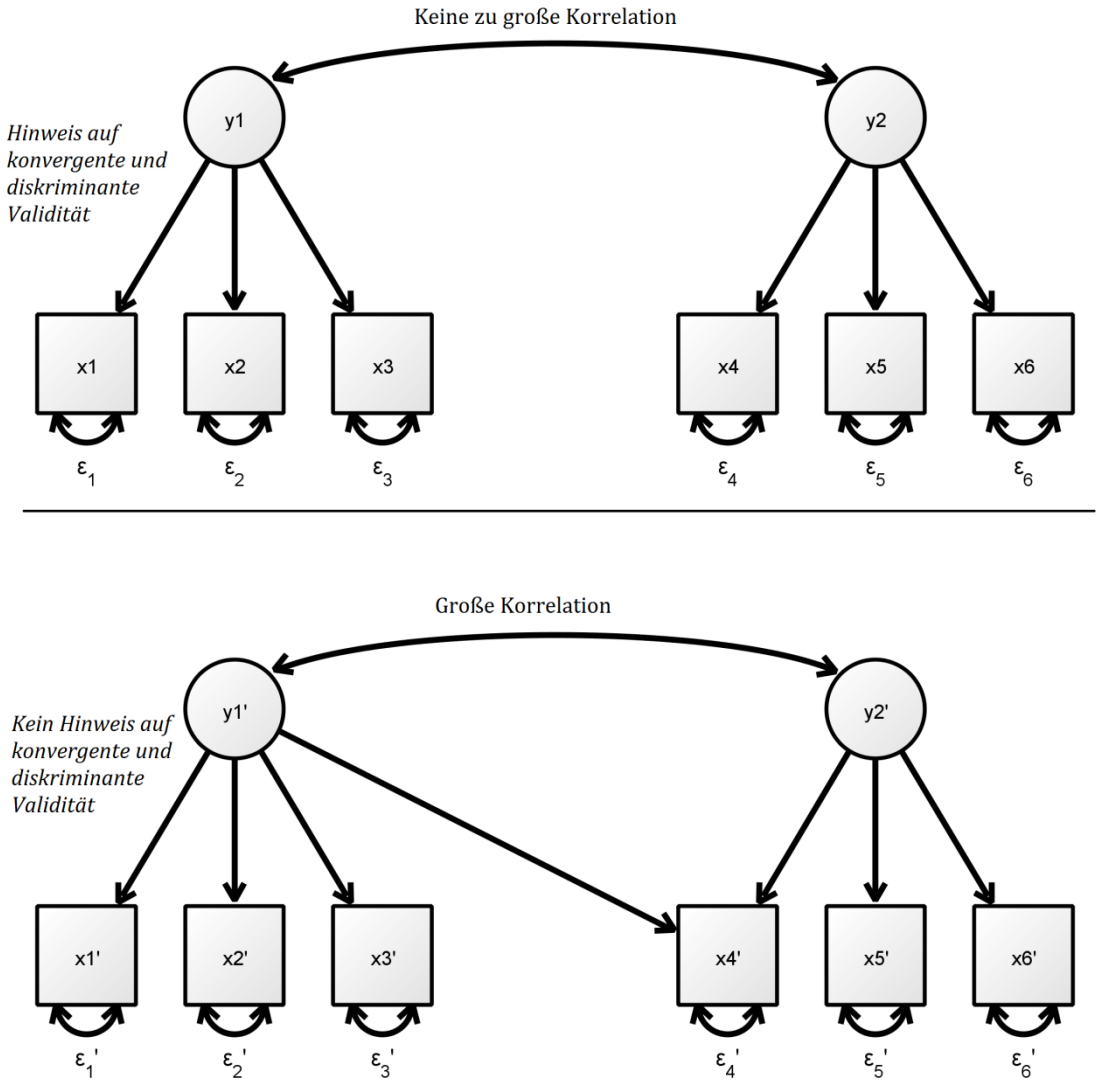


Abbildung 4: Kovergente und diskriminante Validität

Beide Maße sind somit vergleichbar mit der Paralleltestreliabilität. Zuletzt beschreibt in dieser Tradition die faktorielle Validität, ob eine angemessene Prüfung der Faktorstruktur einer Messung stattgefunden hat.

Es stellt sich nun die Frage, welche Grade für welche Form von Validität hinreichend sind. Dabei ist es in der Tradition nach Messick (1995) nötig, den interpretativen Rahmen zu vergegenwärtigen, in welchem eine Wahrheitsaussage validiert werden soll. Dies kann die Bewertung, die Verallgemeinerung, die Extrapolation, die Erklärung oder das Fällen weiterführender Entscheidungen sein (Hartig et al., 2008, S. 136). Vor dem Hintergrund der

notwendigen Tragfähigkeit der Wahrheitsaussage muss also entschieden werden, was hinreichend ist.

Zusammenfassend muss Testqualität, wie aufgezeigt wurde, differenziert und anhand unterschiedlicher Argumente sowie empirischer Befunde nachgewiesen werden, damit auf Validität verwiesen werden kann (American Educational Research Association et al., 2014). Diese sollten nicht als Validitätskriterien tituliert werden. Es ist zu leicht, diese im Sprachgebrauch falsch zu verwenden, etwa wenn man sagt, dass ein Experiment einer nicht experimentellen Studie hinsichtlich der Konstruktvalidität überlegen ist (Döring & Bortz, 2016, S. 93). Stattdessen sollte auch hier der Begriff der Evidenzen Verwendung finden. Evidenzen für die Annahme der Validität können verschiedene Quellen haben; APA, AERA und NCME (2014) nennen den Testinhalt, die Antwortprozesse, die interne Struktur, die Relation zu anderen Variablen sowie Nutzen und Konsequenzen des Tests. Dieser multikriteriale Ansatz steht dabei in direkter Konkurrenz zu der Denkweise, Tests auf der Basis von einzelnen Kriterien und Normen als (un-)brauchbar einzustufen (Newton & Shaw, 2013, S. 315). Dies ist zwar leichter zu versprachlichen, aber weder theoretische Grundlegungen, die methodologische Umsetzung der Testung, die statistischen Eigenschaften von Messmodellen, noch die angestrebte Tragweite, noch das inhaltliche Gewicht werden dabei hinreichend reflektiert. Die technische Annehmbarkeit eines Modells kann kein hinreichender Indikator für die Konstruktvalidität sein, ebenso wenig wie eine einzelne Korrelation zwischen zwei Instrumenten hinreichend ist, um die Konvergenzvalidität zweifelsfrei zu bestimmen. Festgehalten werden soll dabei, dass die Validität im spezifisch disziplinären Verständnis Eigenschaft eines Tests oder einer wissenschaftlichen Wahrheitsaussage sein kann. Es ist im Einzelfall zu prüfen, ob beispielsweise die Validität eines Items, eines Testscores, eines beobachteten strukturellen Zusammenhangs oder eine inhaltliche Weiterverwendung validiert werden soll. Es ist nachvollziehbar, dass sich beispielsweise das psychometrische Verständnis von Validität stark an der korrekten stochastisch kausalen Spezifikation der Modelle und logisch-formalen Kriterien orientiert (z.B. Borsboom et al., 2004), während es für die Bildungsadministration eine hohe Bedeutung hat, ob ein Leistungstest valide die Inhalte des Curriculums eines Faches wiederspiegelt. Dabei ist die rein technische Validierung Ausdruck eines eher schmalen wissenschaftlichen Konzeptes von Validität, z. B. bei Borsboom et al. (2004), während die Inklusion wissenschaftlicher und ethischer Dimensionen ein breiteres Konzept darstellt (Cronbach, 1982; Messick, 1995). Als Bestandteil einer Wahrheitsaussage kann die Evidenz für Validität unter unterschiedlichen Umständen in ihrer Stärke variieren (Shadish et al., 2002, S. 34). Beispielsweise könnte ein diagnostisches Instrument in einem Klassenraum nicht funktional sein. Dies ist nicht zwangsweise als zu generalisierendes negatives Merkmal zu verstehen: auch Tests, die in einer Situation nur eine geringe Validität aufweisen, können unter anderen Bedingungen nützlich sein. Eines der Probleme im Umgang mit dem Begriff der Validität ist sicherlich, dass nicht hinreichend betont wird, dass alle festgestellte Validität konditional ist.

Validität wird hier begriffen als einheitliches Konzept von Eigenschaft von Schlussfolgerungen, die auf einer Theorie und Testergebnissen beruhen. Forscher müssen also offenlegen (1), was sie annehmen und wie die Basis ihrer Annahme ist, (2) wie die Evidenzlage ist, (3) warum man die Validität auf Basis dieser annehmen und im logischen Schluss danach handeln sollte. Bei der

Heranführung von Evidenz ist dabei von entscheidender Bedeutung, die Konstruktvalidität zu prüfen. Hierbei sind latente Variablenmodelle besonders geeignet, um Evidenzen für die Konstruktvalidität zu prüfen, denn einige der Prüfstrategien (z. B. die Dimensionalität oder die Modellgüte) lassen sich unter den Annahmen der klassischen Testtheorie gar nicht prüfen (Baghaei & Tabatabaee Yazdi, 2016), und letztlich ist die Validität der beabsichtigten Interpretation, also der Wahrheitsaussage, von der technischen Umsetzung abhängig. Hierbei müssen die Dimensionen der Untersuchungseinheiten, der untersuchten Effekte, der Beobachtungen und des Settings mitgedacht werden, ebenso wie die Adressaten, also beispielsweise die Bildungsadministration oder pädagogisch tätige Personen. Normative Schwellen, wie sie in den Standards elaboriert wurden, werden als ein Rückschluss aus der unangemessenen Nutzung von einzelnen Kriterien (Newton & Shaw, 2013, S. 313) als Orientierungshilfen begriffen. „It would be the antithesis of science to require all scientists to work within a common paradigm.“ (Newton & Shaw, 2013, S. 313).

5. MODELLEVALUATION

Im vorangegangenen Kapitel wurden die generellen Konzepte der Güte von Tests vor dem Hintergrund latenter Messmodelle beschrieben. Es wurde festgehalten, dass die Validität in der inhaltlichen Komposition, dem statistischen Zusammenspiel der Komponenten und der theoretischen Verknüpfung eines Tests verankert ist. Zudem wurde aufgezeigt, dass die Reliabilität oder Präzision von Tests im Allgemeinen und insbesondere in latenten Variablenmodellen häufig nicht hinreichend durch einen einzelnen Zahlenwert bestimmt werden kann und die Objektivität latenter Variablenmodelle durch die hinreichende Identifikation des statistischen Modells, eine ausreichende Teststärke und die Funktionalität des Modells abgeleitet werden muss. Das folgende Kapitel beschäftigt sich mit den Regeln der Modellidentifikation sowie verschiedenen Maßen und Techniken der Modellbeurteilung, die verwendet werden können, um die Objektivität eines Tests nachzuweisen, damit über diese auf die Reliabilität und Evidenz für die Validität der entsprechenden Wahrheitsaussage geschlossen werden kann. Es werden dabei Maße und Techniken fokussiert, deren Anwendung in aktuellen bildungswissenschaftlichen Veröffentlichungen üblich sind.

5.1. MODELLIDENTIFIKATION

In statistischen Modellen werden zahlreiche Parameter simultan geschätzt, um die empirischen Zusammenhänge möglichst gut im Modell wiederzugeben. Dies funktioniert aber nur, wenn ausreichend viele beobachtete oder restringierte Informationen vorliegen (Döring & Bortz, 2016, S. 964). Eine Nicht-Identifikation bedeutet konkret, dass geschätzte Modellparameter nicht eindeutig sind, das Modell also nicht belastet werden kann.

Die Modellidentifikation ist im übertragenen Sinne ein Prüfmechanismus, um festzustellen, ob ausreichend viele Perspektiven und Beobachtungspositionen (vgl. Kap. 4.1) eingenommen wurden, um über ein Objekt zu berichten, also eine ausreichende Objektivität vorliegt. Dies ist besonders relevant, da Modellparameter geschätzt werden können (und werden), auch wenn Modelle empirisch nicht hinreichend identifiziert sind. Die Modellidentifikation ist technisch ein

einfaches Konzept. Ein Modell ist identifizierbar, wenn es theoretisch möglich ist, eine einzigartige „beste“ Lösung für alle Modellparameter zu finden (Kenny, Kashy & Bolger, 1998, S. 253). Dies ist in erster Instanz eine Frage, die auf die Modellcharakteristika und nicht auf die Datengrundlage zurückzuführen ist. Eine theoretische Modellidentifikation ist also unabhängig vom Stichprobenumfang (Kline, 2011, S. 93). Traditionell erfordert die Identifikation von Modellen eine formale mathematische Analyse (Kenny et al., 1998; Skrondal & Rabe-Hesketh, 2004, Kap. 5) des Verhältnisses von unbekannten zu bekannten Größen im Modell. Dabei sind (1) die Zahl der manifesten Variablen, (2) die Anzahl der Ausprägungen der manifesten Variablen bei kategorialer Ausprägung dieser und (3) in latenten Klassenmodellen die Anzahl der latenten Klassen als manifeste Größen anzusehen. Grundsätzlich sind moderne Analyseprogramme (z. B. Mplus, LatentGold, lavaan) in der Lage, anzuzeigen, ob ein Modell hinreichend identifiziert ist oder nicht – trotzdem ist es hilfreich, einige zentrale Bedingungen zur Identifikation zu kennen, um die Identifikation bereits in der Modellspezifikation zu bedenken. Kline (2011) bietet einen umfassenden Überblick zu der Modellidentifikation für Faktormodelle, dem in diesem Abschnitt stark gefolgt wird. Am Ende dieses Abschnitts findet eine Übertragung in IR- und LCA/LPA- Modelle statt. Einen formalen Überblick für generalisierte latente Variablenmodelle liefern Skrondal und Rabe-Hesketh (2004, S. 137).

Um eine Identifikation zu prüfen, existieren verschiedene Bedingungen. Zwei zentrale Bedingungen sind

- die Skala der latenten Variablen[35]
- sowie die Anzahl der Freiheitsgrade df, also das Verhältnis der bekannten (oder restringierten) und der zu schätzenden Modellparameter.

Als Parameter zählen generell (1) die Varianzen abhängiger Variablen, (2) die Kovarianzen unabhängiger Variablen, (3) Faktorladungen und (4) Regressionskoeffizienten. Diese können frei variieren, also geschätzt, fixiert oder beobachtet werden. Die Pfadkoeffizienten von Residuen sind immer auf 1 fixiert, sodass diese nicht gezählt werden. Varianzen von beobachteten und Kovarianzen zwischen abhängigen und zwischen unabhängigen und abhängigen Variablen sind keine Parameter. k sei die Anzahl der beobachteten Varianzen und Kovarianzen. Die Anzahl der bekannten Parameter n wird bestimmt aus k und muss gleich oder größer als die Anzahl der zu schätzenden, also freien, Parameter u sein. Diese Regel ist als die t Regel (Reinecke, 2014, S. 57) oder Counting Rule (Kline, 2011, S. 125) bekannt.

Es gilt:

$$n = (\frac{k(k+1)}{2})$$

und

$$df = n - u.$$

[35] Zur Festlegung der Metrik der latenten Variablen vergleiche Kapitel 3.1.

Modelle können unteridentifiziert ($df_M < 0$), gerade identifiziert ($df_M = 0$) oder überidentifiziert sein ($df_M > 0$). Unteridentifizierte Modelle ergeben keine eindeutigen Modellparameter, also keine eindeutige Modelllösung. Gerade identifizierte Modelle sind in ihrer Güte nicht bestimmbar, da die Freiheitsgrade gleich 0 sind und keine Modelltests vorgenommen werden können (z. B. Bühner, 2011, S. 406; Kline, 2011, S. 125ff; vgl. Kap. 5.2). Modelle ohne latente Variablen oder latente Modelle ohne jedwede Restriktionen besitzen keine zu schätzenden Parameter oder sind gerade identifiziert; somit können diese nicht auf ihre Modellgüte getestet werden (Bollen, 1989, S. 95ff).[36] Nur für überidentifizierte Modelle gilt, dass einzigartige Parameterschätzungen vorliegen und das Modell auf seine Güte getestet werden kann. Eine hohe Zahl von df ist dabei ein Indikator für Modellsparsamkeit.

Aus dieser Bedingung lässt sich ableiten, dass ein latentes Messmodell mit einer kontinuierlichen latenten Variable gerade identifiziert ist, wenn mindestens drei Variablen verwendet werden. Denn in diesem Fall existieren drei beobachtete Werte (n), deren Varianzen bekannt sind, woraus sich ergibt, dass

$$(\frac{3*(3+1)}{2} = 6 = n),$$

da drei Residuen und drei Faktorladungen zu schätzen sind ($6 = u$). Es gilt $df = 0$. Wenn nur zwei manifeste Variablen vorliegen, muss das latente Konstrukt mit zumindest einem weiteren latenten Konstrukt korreliert sein (Goodman, 1974; Bollen, 1989), um Identifizierbarkeit zu erreichen. Die Zahl der Freiheitsgrade ist aber lediglich eine notwendige und keine suffiziente Bedingung für eine Identifikation (Rigdon, 1995), denn es kann zu einer empirischen Unteridentifikation kommen. Kenny (1979) beschreibt beispielhaft einen Fall mit zwei korrelierten Faktoren mit jeweils nur zwei Indikatoren. Wenn nun die Korrelation zwischen den Variablen, welche zur Identifikation nötig ist, nahe 0 liegt, ist das Modell nicht-identifiziert, wenn also empirische Parameter nahe 0 liegen, tragen diese nicht zur Modellidentifikation bei.

Um eine Heuristik der Modellidentifikation in Anlehnung an Kline (2011, S. 132) vorzustellen, muss das Konzept der Rekursivität eingeführt werden. Der Begriff der rekursiven Modelle hat insbesondere in der Pfadanalyse, also im Strukturgleichungsmodellen ohne latente Variablen, Bedeutung. Rekursive Modelle sind Modelle, bei denen keine Rückkopplung auf sich selbst, beispielsweise über Kovarianzen oder Mediatoren, stattfindet. Es kommen lediglich indirekte Effekte einer Variablen auf sich selbst vor. In der Abbildung 5 sind verschiedene Formen nicht rekursiver Modelle dargestellt. Rekursive Pfadmodelle sind immer hinreichend identifiziert (z.B. Bollen, 1989, S. 95).

[36] Für Modelle ohne latente Variablen wird nicht die Güte einer Schätzung bestimmt, sondern nur, ob Restriktionen haltbar sind.

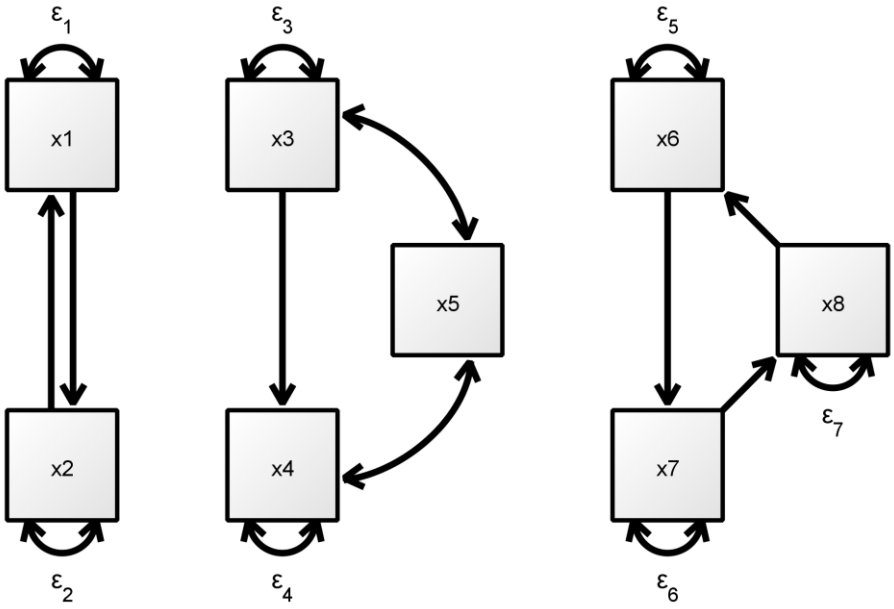


Abbildung 5: Nicht rekursive Pfadmodelle

Zur eindeutigen Identifikation von nicht rekursiven Modellen müssen weitere Bedingungen erfüllt werden, die Ordnungs- und die Rangregel. Beide beziehen sich auf die parallelen Gleichungen in Strukturanalysen. Die Ordnungsregel besagt, dass die Zahl der exogenen (unabhängigen) Variablen gleich oder größer der Zahl der endogenen (abhängigen) Variablen minus 1 sein muss. Die Rangregel besagt, dass jede endogene Variable, welche einen Teil einer rekursiven Schleife darstellt, von zumindest einer unabhängigen Variablen beeinflusst werden muss, die außerhalb der Schleife liegt. Die Erfüllung der Rangregel kann mittels einer algebraischen Prüfung sichergestellt werden (Rigdon, 1995). Die Rangregel ist suffizient für die Identifikation. Eine detailliertere Darstellung der Rangregel und der Ordnungsregel findet sich bei Kline (2011, S. 133ff).

Für komplexe Modelle, wie zum Beispiel Mehrgruppenmodelle, Modelle mit Kreuzladungen und Korrelationen zwischen den Fehlertermen verschiedener Faktoren und Modellen mit latenten Regressionen gelten Übertragungen dieser Bedingungen in jedes einzelne Messmodell und in die jeweiligen Strukturmodelle. Zusammenfassend kann eine Heuristik vorgestellt werden, die Identifikationsbedingungen für verschiedene Modelltypen zusammenfasst. Diese wurde von Kenny, Kashy und Bolger (1998, S. 253ff) formuliert und von Kline (2011) erweitert.

Tabelle 4: Verkürzte Regeln zur Identifikation von Strukturgleichungsmodellen

Regel	
	Rekursive Modelle
1	Rekursive Pfadmodelle sind immer identifiziert.
	Nicht- rekursive Modelle
2	Nicht- rekursive Pfadmodelle, die die Ordnungs-Bedingung erfüllen, sind identifiziert.
3	Nicht- rekursive Pfadmodelle, die die Rang-Bedingung erfüllen, sind suffizient identifiziert.
	Kongenerische Messmodelle
4	Ein kongenerisches konfirmatorisches Messmodell mit einem Faktor ist identifiziert, wenn der Faktor mindestens drei Indikatoren hat, deren Fehler nicht korreliert sind. Wenn mehr Indikatoren vorliegen, können Restriktionen entfallen.
	Multiple Messmodelle
5	Für Modelle mit mehreren Messmodellen gilt, also mehreren korrelierten latenten Variablen, dass für jeden Faktor mindestens zwei Indikatoren vorliegen, deren Fehlerterme nicht korrelieren. Wenn mehr Indikatoren vorliegen, können Restriktionen entfallen.
	Multiple Messmodelle mit korrelierten Residuen
6	Für Modelle mit mehreren Messmodellen und Residualkorrelationen gilt, dass für jeden Faktor mindestens drei Indikatoren vorliegen, deren Residuen nicht korrelieren, oder wenigstens zwei Indikatoren unkorreliert sind und entweder die Residuen beider Indikatoren nicht mit dem Residuum eines dritten Indikators eines anderen Faktors korrelieren oder eine Gleichheitsrestriktion auf die Ladungen zweier Faktoren gelegt wird.
6a	Zudem muss für jedes Faktorenpaar mindestens jeweils ein Residuum unkorreliert mit den anderen bleiben.
6b	Und für jeden Indikator muss zumindest ein anderer Indikator bestehen, der nicht mit dessen Residuum korreliert.
	Multiple Messmodelle mit Kreuzladungen
7	Für Modelle mit mehreren Messmodellen und komplexen Indikatoren, also Indikatoren, die auf mehrere Faktoren laden, gilt, dass die Bedingungen 6 und 6a erfüllt sein müssen.
8	Wenn zudem korrelierte Messfehler vorliegen, muss die Bedingung 6b erfüllt sein und für jeden Faktor auf den eine Kreuzladung erfolgt, muss wenigstens ein Indikator vorliegen, dessen Fehler nicht mit dem komplexen Indikator korreliert.
	Modelle mit latenten Regressionen
9	Alle einzelnen Messmodelle erfüllen die Bedingungen für einzelne Messmodelle.
10	Für das Strukturmodell gelten die Bedingungen 1 bis 3.

Wie gesehen werden kann, gibt es je nach Modellkomplexität unterschiedliche Mindestanzahlen von Indikatoren pro Faktor. In komplexeren Modellen ist es möglich, dass weniger Indikatoren verwendet werden, da die Identifikation über „statistische Anker" (Kline, 2011, S. 135) ermöglicht wird. Mulaik und Millsap (2000, S. 40) argumentieren, dass Faktormodelle immer über mindestens vier Indikatoren verfügen sollten, während Hayduk und Glaser (2000b, S. 8f) den Standpunkt vertreten, dass auch Modelle mit nur einem Indikator angemessen sein können, insofern diese in einen größeren Kontext eingebunden sind (vgl. Hayduk & Glaser, 2000a).[37] Als Daumenregel kann ein Merksatz von Kenny (1979, S. 143) gelten: „Two *might* be fine, three is better, four is best, and anything more is gravy." Wenn in Faktormodellen wenigstens vier Variablen Indikatoren eines Faktors sind und das Faktormodell konsistent ist, unterstützt dies die Annahme der Objektivität des Faktors (Mulaik, 2004, S. 443). Bei Messmodellen mit weniger als vier Indikatoren müssen hingegen weitere Bedingungen gegeben sein, damit die Objektivität angenommen werden kann. Beispielhaft wird im Beitrag von Jaekel et al. (im Erscheinen [2017]; Beitrag 6) eine latente Variable zeitlich versetzt nur auf der Basis zweier Indikatoren abgeleitet. Die Schätzung der Modellgüte eines einzelnen Modells pro Messzeitpunkt ist nicht möglich, da dieses Modell negative Freiheitsgrade aufweist ($df = -1$). Die Indikatoren sind allerdings in IRM geschätzte Kompetenzscores zu zwei Messzeitpunkten in den Domänen Englisch Leseverstehen und Englisch Hörverstehen, so dass inhaltlich angenommen wird, dass deren Kovarianz einen Schluss auf die Englischkompetenz erlaubt und die Messungen kausal verknüpft sind. Die Schätzung eines verknüpften Modells erscheint somit angemessen und ist technisch möglich ($df = 6$).

Eine Erhöhung des Komplexitätsgrades der Bestimmung der Identifizierbarkeit liegt vor, wenn multiple manifeste Items mit Schwellenwerten, also polytome kategoriale manifeste Variablen oder auch manifeste Variablen mit Schwellenwerten in Mischverteilungsmodellen, vorliegen. Hier können einzigartige Kombinationen von Werten auftreten, die eine Nicht-Identifizierbarkeit auslösen, obwohl die Zahl der Freiheitsgrade größer als 1 ist (McCutcheon, 2002). Derartige Sonderfälle können über die Einführung von Gleichheitsrestriktionen oder die Aggregation von gering besetzten Kategorien gelöst werden.

Es kann festgehalten werden, dass Identifikation bei Pfadmodellen, einfachen konfirmatorischen Modellen und rekursiven Modellen über die zwei Grundbedingungen einfach festzustellen ist und sich dieses Konzept auf latente Klassenmodelle und IRM ausweiten lässt. Auch für IRM muss zur Identifikation das Skalenniveau der latenten Variable fixiert werden. Dies geschieht wieder über den Einsatz von Techniken ähnlich der UVI oder ULI, wobei das Erkenntnisinteresse für die zu fixierende Größe leitend ist. Wenn die Varianz der latenten Variable irrelevant ist und Personenwerte abgeleitet werden sollen, können der Mittelwert und die Varianz der latenten Variable fixiert werden, üblicherweise auf einen Mittelwert von 0 und einer Standardabweichung von 1. Wenn hingegen der Mittelwert und die Standardabweichung der latenten Variable geschätzt werden und im Erkenntnisinteresse stehen, müssen andere Modellparameter fixiert werden. Insofern dies gegeben ist, gelten die gleichen

[37] Die Diskussion um diese Frage steht in direktem Zusammenhang zu konkurrierenden Modellierungsstrategien. Eine Ausgabe von *Structural Equation Modelling: A Multidisciplinary Journal* (2000, Volume 7 – Issue 1) trägt diese Diskussion zusammen.

Identifikationsregeln wie für kongenerische Modelle. Dies ist aber nur im seltensten Fall relevant, da die meisten IRM über mehr Items verfügen als Strukturgleichungsmodelle, um das Fähigkeitsspektrum, also üblicherweise eine Normalverteilung, hinreichend abbilden zu können.

Für LCA- und LPA-Modelle gilt, ebenso wie für Faktormodelle, dass drei Items hinreichend für eine Identifikation sind (Goodman, 1974) und die Metrik der latenten Variable durch die Zahl der zu ermittelnden Klassen gegeben ist. Beispielhaft gibt es bei vier dichotomen Items 16 mögliche Antwortmuster und 15 Freiheitsgrade ($Anzahl\ der\ Antwortmuster - 1$), wenn alle Muster vorkommen. Ein Modell mit nur drei latenten Klassen benötigt 13 unabhängig geschätzte Parameter, um identifiziert zu sein, sowie die Klassenanzahl und für jede der Klassen vier bedingte Antwortwahrscheinlichkeiten in diesem Beispiel. Die Zahl der benötigten Parameter ergibt sich also aus einem Klassenanzahlsparameter (1) zuzüglich der Antwortwahrscheinlichkeiten für jedes Item innerhalb jeder Klasse ($4; 1 + [3 * 4] = 13$). Die Identifikation wäre in diesem Beispiel also gegeben, aber bereits ein Modell mit vier Klassen wäre unteridentifiziert, da 17 Parameter benötigt würden. Für LPA-Modelle mit polytomen Items erhöht sich das Problem drastisch (McCutcheon, 2002) und das Problem der empirischen Unteridentifikation ist schwerwiegender, da die Zahl der Freiheitsgrade nicht die nötigen Zellbelegungen berücksichtigt. Auch hier müssen gegebenenfalls Gleichheitsrestriktionen oder Aggregate eingeführt werden, um das Problem aufzulösen. Neben komplexeren Tests ist es praktikabel, die Zellhäufigkeiten zu prüfen und die Analyse mit variierenden Startwerten zu wiederholen, um zu überprüfen, ob die Ergebnisse lokale oder globale Maxima darstellen. Identifizierte Modelle sollten mit mehreren Startwerten deckungsgleiche Ergebnisse aufweisen.[38]

5.2. Modellbeurteilung

Die Möglichkeit, theoretische Annahmen anhand ihrer Nähe zu empirischen Daten zu prüfen und zu diagnostizieren, stellt den entscheidenden Vorteil statistischer Modelle dar. Für wenigstens 30 Jahre wird der beste Weg diskutiert, Hypothesen zu testen und Modellgültigkeit zu beurteilen. Es ist außerdem wahrscheinlich, dass sich dies fortsetzen wird, da es kein einzelnes statistisches Rahmenmodell gibt, welches wahre und falsche Hypothesen trennen kann (Kline, 2011, S. 190). Auf der Basis fehlerbehafteter Messungen und Stichprobenrealisierungen in den Bildungs- wie auch in den Sozialwissenschaften im Allgemeinen sowie der philosophischen Frage, ob ein „perfektes“ Modell überhaupt existiert, muss ein Modell immer als eine Ideenapproximation, eine pragmatische Wirklichkeitsreduktion verstanden werden. Das Ziel ist dabei, ein plausibles, sparsames und substanziell bedeutsames Modell zu erstellen, welches gut zu den Daten passt. Die konkrete Modellgenese, also das Vorgehen im Modellierungsprozess, kann hier nicht bearbeitet werden, und dazu findet sich eine große Zahl von modelltypübergreifenden und modelltypspezifischen Lehrbüchern (Bacher et al., 2010; Backhaus, Erichson & Weiber, 2013; Bühner, 2011; Kaplan, 2009; Kline, 2011; Moosbrugger & Kelava, 2008; Reinecke, 2014; Rost, 1996; Sedlmeier & Renkewitz, 2013). Seltener hingegen findet sich eine modelltypübergreifende Zusammenfassung möglicher Strategien der Prüfung

[38] Daneben stehen komplexere Verfahren zur Verfügung, die z. B. von Goodman (1974) eingeführt wurden.

und Evaluation von Gesamt- und Teilmodellen (Bühner, 2011; Rost, 1996), denn diese ist entscheidend abhängig von dem Modelltyp. Beispielsweise liegen einfach darstellbare globale absolute Fit-Indizes für probabilistische Modelle, also IRM und LCA/LPA, nicht oder nur eingeschränkt vor. Zudem existieren keine absolut gültigen Standards; auch nicht für arrivierte Methoden der Modellbeurteilung. Daher werden in der Regel eine Anzahl verschiedenartiger Korrespondenzanalysen zwischen dem Modell oder Teilen des Modells und den Daten oder alternativen Modellen vorgenommen, um die Passung des Modells auf die Daten zu beurteilen.

Ein genereller Ansatz zur Prüfung der Gültigkeit des abgeleiteten Modells sind Techniken der Stichprobenwiederholung (*resampling*; Skrondal & Rabe-Hesketh, 2004, S. 272). Um eine Prüfverteilung für die anhand der echten Werte errechneten Fit-Maße zu erhalten, resimuliert man viele künstliche Datensätze, berechnet für jeden die betreffende Prüfstatistik und beurteilt, ob der Wert der echten Daten noch im Schwankungsbereich der simulierten Werte liegt (Rost, 1996, S. 338). Eine Einführung zu sogenannten *Bootstrap-Resampling*-Techniken, zum Zwecke der Modellbeurteilung, der Herstellung von Konfidenzintervallen und dem verbesserten Umgang mit fehlenden Werten, findet sich beispielsweise bei Shikano (2010).

In der Folge soll der Versuch unternommen werden, eine Systematisierung verschiedener Strategien und Techniken aufzuzeigen die verwendet werden können, um Modelle auf ihre Passung zu den Daten hin zu beurteilen. Dabei wird das Augenmerk nicht auf die technische Herleitung gelegt, welche an jeweils aufgeführten Stellen nachvollzogen werden kann, sondern auf die impliziten Annahmen und Bedingungen sowie bekannte Charakteristika der Kriterien, also auf notwendige Kenntnisse um verwendbare Maße in deren forschungspraktischen Beurteilungen.

5.2.1. GLOBALER ABSOLUTER FIT

Der absolute Fit quantifiziert die Diskrepanz zwischen dem Modell und den Daten in einem einzelnen Wert. Wenn diese Diskrepanz geringer ist, als zufällig angenommen werden kann, kann dies als Unterstützung für das Modell gewertet werden. Eine Voraussetzung für den Einsatz absoluter Teststatistiken ist gegeben durch die modelltheoretische Annahme, dass die spezifizierte Kovarianzmatrix Σ gleich der interessierenden Populationskovarianzmatrix $\Sigma(\Theta)$ ist

$$\Sigma = \Sigma(\Theta).$$

Da die Annahme aber in der Regel nicht prüfbar ist, wird die Abweichung gegenüber einer Stichprobenkovarianzmatrix *S* bestimmt, deren Repräsentativität durch die Stichprobenziehung gewährleistet sein muss. Im Rahmen von Strukturgleichungsmodellen bedeutet dies, dass Σ, welche durch die Modellvorgaben spezifiziert worden ist, *S* ähnlich genug ist, um die Differenz als Stichprobenfehler zu begreifen (Kline, 2011, S. 193). *S* wird dabei durch alternierende möglichst perfekte Modelle oder Nullmodelle geschätzt. Die Evaluation der Abweichung kann über Signifikanzprüfungen oder approximativ erfolgen. Ursachen für globale Verletzungen der Modellgüte können in der Heterogenität oder Qualität der Items und/oder der Heterogenität der Personen und/oder der Verletzung der lokalen stochastischen Unabhängigkeit liegen (Rost, 1996, S. 340). Modelle können in diesem Sinne nur verglichen werden, wenn diese in einer hierarchischen Abfolge (*nested Models*) zueinander stehen (Rost, 1996, S. 229ff). Das zu

prüfende Modell muss sich durch die Hinzugabe von Restriktionen im Prüfmodell darstellen lassen und die Hinzugabe der Restriktionen darf sich nicht auf das Fixieren von Parametern auf 0 beschränken.

5.2.1.1. Hypothesenprüfung

Bei Prüfungen über Signifikanztests wird die mögliche Annahme oder Ablehnung über die Wahl des Fehlerniveaus α festgelegt (üblicherweise .05 oder .01). Zumeist wird hierbei eine „badness-of-fit"-Logik angelegt, so dass ein statistisch signifikantes Ergebnis eine problematische Modellanpassung aufzeigt. Die Nullhypothese besagt also, dass das Modell nicht zur Datenstruktur passt (z.B. Rost, 1996, S. 331f). Der verbreitetste Test ist der Pearson χ^2-Test für Modellgüte (Hu & Bentler, 1998, S. 426). Der χ^2-Test basiert auf dem Mittel aus den beobachteten Diskrepanzen durch eine Diskrepanzfunktion d. Diese werden durch eine Diskrepanzfunktion *F* gewichtet und minimiert (Bollen, 1989, S. 257). Der Testwert ergibt sich dann generalisiert als

$$\chi_d^2 = (n-1)F_d.$$

Dieser kann, da es sich um eine Approximation einer χ^2-Verteilung handelt, auf Signifikanz gegenüber einem saturierten Modell geprüft werden (Raykov & Marcoulides, 2006, S. 41) und folgt dabei einer Likelihood-ratio-Logik (Maydeu-Olivares & Garcia-Forero, 2010, S. 190).

Diskrepanzfunktionen

Um die Diskrepanzen zu bestimmen und zu minimieren, können verschiedene Funktionen verwendet werden, zum Beispiel über die Methodenfamilien der gewichteten kleinsten Quadrate (*weighted least squares*; WLS), ungewichteten kleinsten Quadrate (*unweighted least squares*; ULS oder *ordinary least squares*; OLS) oder die maximierte Wahrscheinlichkeit (*maximum liklihood*; ML). Vertiefungen zu den Unterschieden finden sich beispielsweise bei Browne (1984) sowie Muthén und Kaplan (1985) und Asparouhov und Muthén (2010). Überblickshafte Zusammenfassungen finden sich bei Bühner (2011, S. 407)und Kline (2011, S. 195). Die Wahl der Diskrepanzfunktion wird auch als die Wahl des Schätzers bezeichnet. Diese Auswahl ist dabei nicht trivial und steht in direkter Abhängigkeit zu den Verteilungsannahmen in den beobachteten Variablen.

Während der ML- Ansatz zumeist robust ist und am häufigsten verwendet wird (Kline, 2011, S. 176), gibt es verschiedene implizite Modellannahmen, welche verletzt sein können und verzerrte χ^2 Werte und Standardfehler zur Folge haben (Reinecke, 2014, S. 113ff). Der ML-Ansatz setzt voraus, dass die Werte weder schief verteilt sind noch eine hohe Kurtosis haben. Ersteres führt zu einer Über- und Letzteres zu einer Unterschätzung der Parameter (Hoogland & Boomsma, 1998). Da auf der Basis des zentralen Grenzwertsatzes Verteilungen mit steigendem Stichprobenumfang immer eher einer Normalverteilung folgen, ist der ML- Ansatz insbesondere für große Stichproben geeignet. Zudem setzt der Einsatz von ML-Diskrepanzfunktionen voraus, dass es sich um eine Kovarianzmatrix handelt. Grundsätzlich sollte zwar mit Rohdaten gearbeitet werden, aber für sekundäranalytische Zwecke stehen teils nur Korrelationsmatritzen zur Verfügung. Hier gilt, dass der ML-Schätzer nur dann verwendet werden kann, wenn die Skalen invariat sind (vgl. Kline, 2011, S. 175). Üblich ist inzwischen eine robuste Korrektur der

ML-basierten χ^2-Statistik (z.B. Satorra & Bentler, 1994, MLM oder Yuan, Chan & Bentler, 2000, MLR).[39]

Die Verwendung von WLS-Schätzern (Browne, 1984) ist unempfindlich gegenüber Verletzungen der Normalverteilungsannahme. Bei der Verwendung von WLS-Schätzern wird es zudem möglich, ordinal-skalierte Indikatoren zu verwenden und damit eine hohe Ähnlichkeit zu IRM herzustellen (Muthén, 2007). Die übliche Konnotation von Itemschwierigkeiten in IRM wird dabei als Schwellenparameter begriffen. Während also robuste ML-Schätzer auch mit Nicht-Normalität umgehen können, also Fit-Statistiken und Standardfehler adjustieren (Satorra & Bentler, 1994), werden die latenten Dimensionen noch immer linear vorhergesagt. Unter der Verwendung von WLS-Schätzern wird es hingegen im Rahmen von Strukturgleichungsmodellen möglich, stetige aber nicht-lineare Verbindungen zwischen manifesten und latenten Variablen herzustellen. Außerdem können erweiterte Fit-Statistiken verwendet werden, die hohe Flexibilität des Strukturgleichungsansatzes kann genutzt werden, gemischt skalierte Indikatoren können verwendet werden, und eine breite Palette leicht verständlicher Softwareanwendungen ist verfügbar.[40]

Ableitungen des χ^2-Wertes

Der χ^2-Test kann einer wahrscheinlichkeitsbasierten oder einer deterministischen Logik folgen (vgl. Rost, 1996; Kap. 5). Die wahrscheinlichkeitsbasierte Ableitung erfolgt über die Likelihood L, die allgemein bestimmt ist als Produkt der unkonditionalen Musterwahrscheinlichkeiten $p(a_v)$ über alle Personen im Modell.

$$L = \prod_{v=1}^{N} p(a_v)$$

Die beobachteten Musterwahrscheinlichkeiten werden dabei schlicht abgezählt. Wenn bei 10 Testteilnehmerinnen und Testteilnehmern und drei Items die Ausprägungen {1,2,4} dreimal vorgekommen, ist $p(a_{\{1,2,4\}}) = .3$. Alle vorkommenden Wahrscheinlichkeiten summieren sich zu 1. Die Antwortmusterwahrscheinlichkeit im Modell ist unbekannt und wird dabei auf Basis der Modellannahmen zu L_1 maximiert. Die χ^2-verteilte Teststatistik ergibt sich dann aus der logarithmisierten Likelihood-Ratio multipliziert mit -2.

$$L = -2 * log(\frac{L_0}{L_1}) \approx \chi^2$$

Der deterministische oder klassische (Rost, 1996, S. 336) Ansatz verwendet anstatt von Wahrscheinlichkeiten die beobachteten *b* und erwarteten *e* Musterhäufigkeiten.

[39] Die Korrekturen sind in den meisten gängigen Softwarepaketen implementiert, hier muss gegebenenfalls geprüft werden, welche Korrektur angewendet wird.

[40] Die Modelle sind allerdings rechenintensiv, anfällig für computationale Probleme aufgrund geringer Zellbesetzungen, und es existieren noch keine Umsetzungen für z. B. Mehrebenenenmodelle. In aktuellen Vergleichen zu anderen robusten Schätzern, wie bei Li (2014) wurde WLS-Schätzer positiv bewertet.

$$\chi^2 = \sum_{a_v} \frac{(b_{a_v} - e_{a_v})^2}{e_{a_v}}$$

Praktisch sind beide Ansätze annähernd äquivalent (Rost, 1996, S. 336), aber die Zahl der Modellparameter wird insbesondere im wahrscheinlichkeitsbasierten Ansatz sehr hoch, weswegen der Test für latente Klassifikationsverfahren und IRM eher ungeeignet ist und nur bei geringer Zahl von Items und Schwellen innerhalb der Items eingesetzt werden kann. Generell müssen immer die Zellbesetzungen hinreichend sein, denn die χ^2-Verteilung ist kontinuierlich und kann über diskrete Häufigkeiten schlecht angenähert werden.[41]

Praktische Probleme

Eine zentrale Voraussetzung für die Verwendung von Hypothesenprüfungen der Modellgüte sind hinreichende Stichprobenumfänge (Reinecke, 2014, S. 114). Bei kleinen Stichprobenumfängen (<100) werden die χ^2-Werte zu groß und müssen durch eine hohe theoretische Absicherung und Reliabilität der Messmodelle kompensiert werden (Boomsma & Hoogland, 2001). Mindestens sollten für jeden df mehrere Versuchspersonen vorliegen (Bollen, 1989, S. 268), Hoogland und Boomsma (1998) empfehlen ein Verhältnis von fünf Personen auf einen df für Modelle mit einer ML-Diskrepanzfunktion. Für andere Diskrepanzfunktionen gelten wiederum andere Empfehlungen, zum Beispiel wird für die WLS-Funktion ein Verhältnis von 20:1 nahegelegt (Hoogland & Boomsma, 1998, S. 363). Um dieses Problem aufzugreifen wurden robuste WLS-Schätzer entwickelt, die bei einem geringeren Verhältnis funktional sind. Dies umfasst beispielsweise den "M*eans-and-variance adjusted weighted least squares"*-Ansatz in MPlus (WLSMV; Muthén & Muthén, 1998-2015). Dieser findet beispielsweise in den Beiträgen von Schurig et al. (2012) und Busch et al. (2015) Anwendung.

Eine weitere Problematik der χ^2 kann bereits aus der Herleitung als Produkt des F-Wertes abgelesen werden. Mit steigendem Stichprobenumfang steigt auch die Wahrscheinlichkeit, eine falsche Nullhypothese zurückzuweisen (Bollen, 1989, S. 268). Auch geringe Differenzen zwischen S und Σ haben bei großen Stichproben dramatische Auswirkungen, sodass nahezu jedes Modell „widerlegt" wird (Reinecke, 2014, S. 115). Hier werden in der Literatur unterschiedliche und teils widersprüchliche Angaben zu der optimalen Anzahl der Werte, also zumeist der Fälle gemacht, welche häufig aus der jeweiligen Erfahrung heraus formuliert wurden. Kline (2011, S. 201) gibt beispielsweise 200 bis 300 Fälle an. Alternativ zu einem Schwellenwert wird anteilig das Verhältnis χ^2/df verwendet, aber auch hier variieren die Angaben zu einem angemessenen Verhältnis massiv. Während Backhaus et al. (2013, S. 149) ein Verhältnis von 2.5:1 und Tabachnick und Fidell (2007) sogar 2:1 empfehlen, geben ältere Publikationen (z.B. Bentler & Chou, 1987) zumeist höhere Verhältnisse (5:1) an. Beispielsweise kann in den Beiträgen von Busch et al. (2015) und Jaekel et al. (im Erscheinen [2017]; Beitrag 6) beobachtet werden, dass die χ^2-Werte in den Prüfungen der Modellanpassung keine Signifikanz erreichten, obwohl Modelle für hinreichend gut zu den Daten passend befunden wurden.

[41] Eine Daumenregel besagt, dass eine Zellbesetzung von >5 dabei hinreichend ist (Maydeu-Olivares und Garcia-Forero, 2010, S. 191).

Die Größe von Korrelationen zwischen beobachteten Variablen im Modell führt naturgemäß zu einer inflationären Erhöhung des χ^2-Wertes, wenn diese unterdrückt werden (Kline, 2011, S. 201). Dies ist im sozialwissenschaftlichen Bereich von hoher Relevanz, da viele Variablen in einem natürlichen, wenn auch häufig geringem Zusammenhang stehen und eine forschungspragmatische Reduktion demnach bei der Bewertung des χ^2 in Konkurrenz zu einer übermäßigen Vereinfachung steht. Darüber hinaus können unberücksichtigte Zusammenhänge auf dimensionale Probleme hindeuten.

Eine hohe einzigartige oder nicht mit dem Modell zusammenhängende Varianz einzelner Variablen ist im Verständnis der Messung ein Zeichen geringer Reliabilität und resultiert im Verlust statistischer Power. Die Eigenschaften des χ^2-Wertes in seiner Verwendung als Schwellenwert für eine Annahme oder eine Ablehnung machen es aber einfacher für das Modell, mit den Daten übereinzustimmen, wenn bedeutungslose Hypothesen mitgetestet werden, das geprüfte Modell also wenige Freiheitsgrade aufweist. Aufgrund der aufgeführten Probleme exakter Hypothesentests wurden approximative Gütemaßen entwickelt.

5.2.1.2. Approximative Modellgütemaße

Es gibt inzwischen eine sehr große Zahl von approximativen Gütemaßen. Eine Sonderausgabe von Personality and Individual Differences in 2007 (z. B. Barrett, 2007; Millsap, 2007; Steiger, 2007) beschäftigt sich ausschließlich hiermit. Nahezu alle Maße sind dabei Umformulierungen des χ^2-Wertes, welche einen Strafterm für Modellkomplexität einführen. Die Logik hinter diesen Straftermen ist, dass Regressionsmodelle mit vielen wissenschaftlich unbedeutenden unabhängigen Variablen immer auch mehr Varianz aufklären, ebenso wie in reflektiven Faktormodellen mehr Faktoren immer eine bessere Vorhersage einzelner Indikatoren erlauben und in latenten Klassenanalysen mehr Klassen immer mehr Varianz in den Indikatoren aufklären bis hin zu einer perfekten Aufklärung, wenn jeder Fall seine eigene Klasse darstellt. Strafterme sollen also zu einer Balance zwischen der Aufklärung und Parsimonität beitragen. Für alle Indizes gilt, dass diese keine linearen Maße mit normativen Schwellen darstellen, sondern eine qualitative und deskriptive Information über die Modellanpassung an die Daten geben. Alle darüber hinaus gehenden Interpretationen eines einzelnen Wertes überbelasten dessen tatsächliche Aussagekraft (Hu & Bentler, 1999). Da in den vergangenen Jahrzehnten Dutzende von Maßen entwickelt wurden, beschränkt sich diese Arbeit auf vier Indizes, welche in der Literatur, insbesondere bei der Beurteilung von Strukturgleichungsmodellen, weite Verbreitung gefunden haben (Bühner, 2011) sowie einen Ausblick auf die Beurteilung von IRM durch approximative Gütemaße.

Approximative Gütemaße im Rahmen von Strukturgleichungsmodellen

Die vier Indizes sind an verschiedenen Stellen umfassend vorgestellt worden (z. B. Jöreskog et al., 1979; Hu & Bentler, 1998; Hu & Bentler, 1999; Steiger, 2007), sodass keine Konzentration auf die technischen Herleitungen, sondern auf implizite Charakteristika der Maße vorgenommen wird.

Der *Tucker-Lewis-Index* (TLI) oder *Non-Normed-Fit-Index* ist bestimmt als

$$TLI = \left(\frac{\chi^2{}_0}{df_0} - \frac{\chi^2{}_1}{df_1}\right) \Big/ \left(\frac{\chi^2{}_0}{df_0} - 1\right),$$

wobei das Verhältnis χ^2/df den Strafterm und die Indizierung 0 das Nullmodell und 1 das zu prüfende Modell darstellt. Insofern sich dieses Verhältnis nicht ändert, bleibt der TLI gleich. Der Wert ist nicht normiert, so dass Werte außerhalb einer Spannweite von 0 bis 1 liegen können.

Der *Comparative-Fit-Index* (CFI) weist eine hohe Ähnlichkeit zum TLI auf und wird bestimmt als

$$CFI = \left(\left(\chi^2{}_0 - df_0\right) - \left(\chi^2{}_1 - df_1\right)\right) / \left(\chi^2{}_0 - df_0\right).$$

Der CFI bestraft jeden einzelnen Parameter und nicht das Verhältnis der Parameter zu dem Testwert. Für den CFI und den TLI gilt, dass ein Wert nahe 1 einen guten Fit bedeutet. Ein gängiger Schwellenwert für die Annahme einer guten Passung des Modells auf die Daten ist .95 (Hu & Bentler, 1999). Werte größer 1 werden dabei auf 1 gesetzt. Ein Wert von 1 tritt für den CFI und TLI auf, wenn $\chi^2{}_m < df_m$, womit eine Unteridentifizierung vorliegt. Der CFI ist, wenn er kleiner 1 ist, immer etwas größer, also liberaler als der TLI. Meade, Johnson und Braddy (2008) stellen in Anlehnung an Cheung und Rensvold (2002) fest, dass der CFI besonders geeignet ist, um Messinvarianz (vgl. Kap. 5.2.3) zu testen, wenn die Differenz der χ^2-Werte aufgrund des Stichprobenumfangs nicht angemessen ist. Eine praktische Anwendung der Differenzen des CFI findet sich bei Schurig, Glesemann und Schröder (2016).

Der TLI und der CFI berücksichtigen direkt die Höhe der Zusammenhänge in den Daten. Geringe modellweite mittlere Zusammenhänge bedeuten einen geringeren Fit. Wenn beispielsweise ein funktionales CFA-Modell mit einer Anzahl von unzusammenhängenden Drittvariablen ergänzt wird, welche ihrerseits nur geringe Ladungsgewichte aufweisen, beispielsweise um Zusammen–hangshypothesen zu testen, werden TLI und CFI sinken. Da der TLI das Verhältnis und keine Differenz wertet, kann dies dazu führen, dass dieser Wert in Modellen mit geringer Zahl von df bei niedrigen modellierten Korrelationen besonders scharf abfällt. Dies kann aber nicht als Aussage zu der eigentlichen Modellgüte gewertet werden. Dies gilt auch anders herum: Ein unpassendes Modell kann auf Basis dieser Indizes „passend“ gemacht werden, wenn unnötig Drittvariablen mit hohen Zusammenhängen hinzugegeben, also Variablen ergänzt werden, die keinen oder nur einen geringen Zusammenhang zu den leitenden Fragestellungen aufweisen. Für Modelle die auf ordinale manifeste Variablen zurückgreifen, also WLS- Schätzer verwenden (vgl. Kap. 5.2.1.1), werden diese Maße erfahrungsgemäß schnell übermäßig kritisch, da durch die zu schätzenden Schwellenparameter die Modelle komplexer werden. Anstatt das eine einzige Faktorladung pro Item geschätzt wird, werden Ladungen für jede Schwelle der Items benötigt. Es muss angenommen werden, dass beide Maße in diesem Fall übermäßig kritisch sind. Aufgrund dieser Eigenschaften wurde dieses Maß beispielsweise in den Beiträgen von Busch et al. (2015) und Jaekel et al. (im Erscheinen [2017]; Beitrag 6) nicht berichtet.

Der *Root Means Square Error of Approximation* (RMSEA) ist ein fehlerbasierter Index, bei dem geringe Werte nahe 0 eine gute Modellpassung anzeigen. Dieser wird abgeleitet als:

$$RMSEA = \sqrt{(\chi^2{}_1 - df_1)/(df_1(n-1))}.$$

Einer der Gründe, warum sich der RMSEA einer besonders großen Beliebtheit erfreut, ist, dass ein Konfidenzintervall für den Wert erstellt werden kann (Kline, 2011, S. 206). Auf Basis dieses Intervalls kann ein Hypothesentest durchgeführt werden, der in der Nullhypothese besagt, dass der RMSEA kleiner oder gleich .05 ist. Dieser Wert basiert auf den von Browne und Cudeck (1992) vorgeschlagenen Schwellenwerten ($.05 = guter\ Fit$; $.08 = knapper\ Fit$). Zusätzlich kann die obere Schwelle des Konfidenzintervalls deskriptiv verwendet werden, um die Sicherheit zu bewerten, mit der der RMSEA innerhalb der Range annehmbarer Werte liegt. Dadurch, dass im Nenner die Freiheitsgrade und die Stichprobengröße verwendet werden, wird der RMSEA kleiner, wenn mehr df_m oder eine größere Stichprobe vorliegen. Dies bedeutet aber nicht, dass per Definition sparsamere Modelle bevorzugt werden (Kline, 2011, S. 205). Für den RMSEA gilt, dass $RMSEA = 0$, wenn $\chi^2{}_m \leq df_m$. Dies bedeutet jeweils keinen perfekten Fit, sondern eine Unteridentifizierung oder eine gerade Identifizierung des Modells (vgl. Kap. 5.1). Dazu ist der RMSEA unzulänglich bei Fällen mit besonders kleinen χ^2-Werten im Verhältnis zu den Freiheitsgraden. Der RMSEA sollte in solchen Fällen nicht berichtet werden, um Fehlinterpretationen zu vermeiden. Kenny, Kaniskan und McCoach (2015) argumentieren, dass der RMSEA auch für Modelle mit sehr geringer Zahl von Freiheitsgraden (nahe 1) gar nicht bestimmt werden sollte, da der Multiplikator zu gering wird. Anstelle dessen sollte direkt nach lokalen Fehlspezifikationen gesucht werden.

Das letzte Modellgütemaß, welches hier vorgestellt wird, ist der *Standardized Root Mean Residual* (SRMR). Dieser beruht auf Abweichungen in den Differenzen der mittleren absoluten Residuen der manifesten Modellvariablen. r steht dabei für die Korrelation der Items j und k und n für die Anzahl der Items. Es ist anzumerken, dass es kleine Unterschiede in verschiedenen computationalen Umsetzungen des SRMR gibt (Maydeu-Olivares & Garcia-Forero, 2010, S. 193). Eine formale Darstellung ist

$$SRMR = \sqrt{\sum_j \sum_{k<j} \left(\frac{r_{jk}^2}{\frac{n(n+1)}{2}}\right)}.$$

Der SRMR ist der einzige der hier vorgestellten Indizes, der sich nicht auf den χ^2-Wert bezieht. Da die Kovarianzmatrizen auf unstandardisierten Variablen beruhen, werden diese in Korrelationsmatrizen überführt. Hu und Bentler (1999) schlagen einen Schwellenwert von SRMR $\leq .08$ für die Feststellung eines guten Modellfits vor. Dieser wird aber bei Kline (2011, S. 209) als zu liberal eingestuft, da, wenn der Wert nahe der Schwelle liegt, mehrere einzelne Paare diesen Wert überschreiten könnten, was auf eine verringerte lokale Aussagekraft hindeuten kann. Demgegenüber wird vorgeschlagen, dass die Differenzen in den Residuen direkt inspiziert werden sollten. Der Wert ist dahingehend verzerrt, dass bei kleinen Stichproben und geringen df_m tendenziell übermäßig positiv zu interpretierende Werte wiedergegeben werden

(Kenny et al., 2015), und verhält sich dabei ähnlich wie der RMSEA. Der SRMR hat keine Korrektur für die Modellkomplexität und kann demnach nicht verwendet werden, um die Parsimonität des Modells zu bewerten.

Approximative Gütemaße im Rahmen von IRM

Für den Bereich der Strukturgleichungsmodelle gibt es umfangreiche Forschungen und Erfahrungswerte zu Maßen der Modellevaluation, welche in Handlungsempfehlungen und Orientierungsgrößen umgearbeitet wurden. Deutlich weniger Ansätze und Erfahrungswerte gibt es für IRM. Die Kernproblematik liegt in den Eigenschaften der Varianzverteilungen zwischen den Personen, den Items und gegebenenfalls den Rateparametern (vgl. Kap. 4.2). Da die Zusammenhangsannahme zwischen den Indikatoren und den latenten Variablen nicht linear ist, ist es informativer, lokale Abweichungsmaße zu inspizieren.

Die approximativen Gütemaße für IRM zentrieren alle die Aggregation der standardisierten Residuen. So ist es möglich, eine Anpassung des SRMR auch für die Evaluation von IRM zu verwenden (Maydeu-Olivares, 2013). Eine Alternative ist die Q3-Statistik (Yen, 1984). Diese basiert auf der Korrelationsmatrix aller Itemresiduale, nachdem das Modell gefittet worden ist. Alternativ verwendet eine Teststatistik von McDonald und Mok (1995) direkt die Kovarianzmatrix. Jeweils können die Mittelwerte oder die Maxima der Matrizen verwendet werden, um insbesondere Verletzungen der lokalen stochastischen Unabhängigkeit zu kontrollieren.

5.2.1.3. Überlappung als Gütemaß

Insbesondere die approximativen Modellgütemaße wurden explizit für Faktormodelle entwickelt. Für latente Klassifikationsverfahren kann hingegen die wahrscheinlichkeitsbasierte Überlappung verschiedener Cluster als Gütemaß angeführt werden. Es wurde bereits erwähnt, dass der Anteil der Unschärfe in latenten Klassenanalysen als Pseudo-Reliabilitätsmaß begriffen werden kann (Rost, 1996, S. 361). Die wahrscheinlichkeitsbasierte Überlappung der Cluster in latenten Klassenanalysen steht der deterministischen Klassenzuordnung in klassischen Clusteranalysen wie Clusterzentrenanalysen oder hierarchischen Clusteranalysen gegenüber (Bacher et al., 2010). Für Clusteranalysen, die neben den Anwendungen in Sozial- und Wirtschaftswissenschaften eine breite Verbreitung im Data Mining, der Epidemiologie und im Maschinenlernen gefunden haben, existiert eine breite Palette von sogenannten Validitätsmaßen[42], um die externe und interne Güte[43] der Klassenlösungen zu bestimmen. Die externe Güte beschreibt dabei den Abgleich mit alternativen Zuordnungen, die interne Güte die Trennung zwischen den Clustern respektive die Dichte innerhalb der Cluster. Für die externe Güte können die Kenngrößen Sensitivität (*sensitivity, true positive, recall*), Spezifität (*specificity, true negative rate, correct rejection*), Ausfallrate (*fallout, false positive*) und negative Fehlklassifikationsrate (*miss-rate, false negative*) bestimmt und in Wahrscheinlichkeiten überführt werden, wodurch Vorhersagewerte und Klassifikationsraten erstellt werden können,

[42] Die Validierung ist an dieser Stelle nicht mit der eingeführten Begrifflichkeit der Validität in den Sozialwissenschaften zu vergleichen, sondern meint nur, dass ein System die Praxisanforderungen erfüllt.
[43] Um Verwirrung zu vermeiden, wird hier auf die allgemeine Formulierung der Güte zurückgegriffen – in der Fachliteratur wird an nahezu jeder Stelle von der Validität der Clusterlösung gesprochen.

insofern eine Vergleichsverteilung vorliegt. Dies wird in Kapitel 5.2.2 erneut aufgegriffen. Die interne Güte wird bestimmt, um die optimale Zahl der Klassen zu bestimmen. Dies wird normalerweise über die Maximierung oder Minimierung eines oder mehrerer Qualitätskriterien hergestellt, beispielsweise einer geometrischen Nähe- oder Distanzfunktion. Dabei existieren dutzende Maße, die verwendet werden können, um die Güte von Clusterlösungen zu bestimmen, was die Beurteilung von Klassifikationen ohne vertiefende Kenntnisse häufig erschwert.

"The validation of clustering structures is the most difficult and frustrating part of cluster analysis. Without a strong effort in this direction, cluster analysis will remain a black art accessible only to those true believers who have experience and great courage" (Jain & Dubes, 1988, S. 222).

Beispielsweise gibt es die Silhouettenweite, den Calinski-Harabasz-Index, die Punkt-Biseriale-Korrelation sowie die Streuung innerhalb und zwischen den Clustern als interne Gütemaße. Das *F*-Maß, der Hubert $\hat{\Gamma}$, der Kontingenzkoeffizient, der korrigierte Rand-Index und eine große Zahl korrelativer Maße können zur Bestimmung der externen Güte herangezogen werden. Die Maße werden an verschiedenen Stellen umfassend vorgestellt (z. B. Bacher et al., 2010; Jain & Dubes, 1988; Kaufman & Rousseeuw, 1990). Allerdings ist ein Großteil dieser Maße an deterministische Klassenzugehörigkeiten gebunden. Während es zwar möglich ist, probabilistische Klassenzugehörigkeiten in deterministische Klassenzugehörigkeiten zu überführen, muss klar sein, welches Ausmaß an Informationen verloren gehen kann, wenn die Zuordnungsunsicherheit nicht verrechnet wird. In der Tabelle 5 sind exemplarisch eine probabilistische und eine deterministische Klassenzugehörigkeit, welche aus der probabilistischen abgleitet wurde, abgetragen.

Tabelle 5: Exemplarische probabilistische und deterministische Klassenzugehörigkeit

	Probabilistische Klassenzugehörigkeit		Determinsitische Klassenzugehörigkeit	
	p(Klasse 1)	p(Klasse 2)	Klasse 1	Klasse 2
Objekt 1	.19	.81	0	1
Objekt 2	.51	.49	1	0

Wie dargestellt wurde, determiniert die Höhe der Zuordnungssicherheit die Zuverlässigkeit der Zuordnung, und es gibt auch hier verschiedene Möglichkeiten, diese zu operationalisieren; beispielsweise den Dunn-Index (Kaufman & Rousseeuw, 1990, S. 171) oder der Backer-Index (Jain & Dubes, 1988, S. 132f), welche auch bei Bacher et al. (2010, S. 369) vorgestellt werden. Diese Übersicht ist aber keinesfalls erschöpfend. Bei Zhang, Ji, Yang, Zhang und Xie (2014) werden zehn weitere Indizes vorgestellt, welche zwischen 1974 und 2010 eingeführt worden sind, und es existieren auch hier diverse weitere. Keiner dieser Indizes, ebenso wenig wie der Dunn-Index und der Backer-Index, war in der Lage, unter allen Umständen eine optimale Clusterzahl anzuzeigen. Auch liegen für keinen dieser Indizes abgesicherte Schwellenwerte, sondern häufig nur Erfahrungswerte vor. Generell muss empfohlen werden, diese nur ergänzend einzusetzen, da die Eigenschaften und Charakteristika nicht hinreichend erforscht worden sind (Bacher et al., 2010). Das bedeutet, dass primär relative Indizes und lokale Gütemaße zur internen und externen Gütebestimmung bei latenten Klassifikationsverfahren herangezogen

werden müssen, da deterministische Maße gegenüber der Datenstruktur nicht angemessen sind und Maße für eine probabilistische Datenstruktur nicht hinreichend gut abgesichert sind.

5.2.2. *Globaler relativer Fit*

Ein erweitertes Vorgehen zur Prüfung eines Modells ist es, dieses mit alternativen Modellen zu vergleichen. Globale absolute Modellevaluationen folgen dieser Logik, indem gegen ein saturiertes oder ein Nullmodell getestet wird, aber auch wenn ein Modell diesbezüglich für hinreichend gut befunden wurde, heißt dies nicht, dass es das beste denkbare Modell ist. Es heißt ebenso wenig, dass es keine weiteren Modelle gibt, die die Daten ebenso gut erklären. Diese Aussage würde erfordern, dass die mögliche Existenz alternierender Modelle (streng gesehen aller theoretisch denkbaren) berücksichtigt werden muss. Eine Alternative, die Modellvergleiche nicht nur gegenüber saturierten Obermodellen oder gegenüber Nullmodellen ermöglichen, sind relative Fit-Indizes oder informationstheoretische Maße (IC). Diese haben besonders für wahrscheinlichkeitsbasierte Modelle (LCA, LPA, IRM) und Dimensionalitätsanalysen in FA Bedeutung. Diese Maße beruhen auf der Likelihood *L* oder dem χ^2-Wert[44] (vgl. Abschnitt 5.2.1.1) und enthalten ebenso wie die meisten approximativen Maße Strafterme für die Modellkomplexität. Formal wird durch alle Informationskriterien die Distanz der wahren Information zu einem Modell geschätzt. Aber interpretiert wird nicht diese Distanz, sondern das Verhältnis der Distanz zu den Distanzen alternativer Modelle.

Auch hier gibt es eine große Zahl verschiedener Maße, bei dem an dieser Stelle eine Beschränkung auf zwei gängige Maße, das Akaike Information Criterion (AIC) und das Bayes Information Criterion (BIC; z. B. Maydeu-Olivares & Garcia-Forero, 2010, S. 191) sowie deren Korrekturen als das Constistent AIC[45] (CAIC) und das Sample Size Adjusted BIC (saBIC; z.B. Dziak, Coffman, Lanza & Li, 2015) erfolgt[46]. Dabei beschreibt n_p die Anzahl der unabhängigen Modellparameter.

Der AIC wird dabei in keiner Weise vom Stichprobenumfang *N* berührt:

$$AIC = -2\log L + 2n_p$$

Der Stichprobenumfang wird im cAIC aufgegriffen, welcher das logarithmierte Stichprobenvolumen multipliziert mit den abhängigen Parametern mal zwei zusätzlich aufschlägt. Damit sind die Werte des cAIC immer größer als die Werte des AIC.

$$cAIC = -2\log L + (\log N) * 2n_p$$

[44] Grundlegend kann beobachtet werden, dass je nachdem welcher Methode eine Veröffentlichung zentral gewidmet ist, entweder der χ^2 Wert, bei Veröffentlichungen zu Faktormodellen (z.B. Kline, 2011), oder die *L* bei Veröffentlichungen zu probabilistischen Modellen (z.B. bei Rost, 1996) herangezogen wird, um diese Maße zu bestimmen.

[45] Es existieren weitere Korrekturen des AIC und BIC, welche beispielsweise als AIC_c oder ${}_{adj}BIC$ notiert sind.

[46] Für alle Indizes existieren Formulierungen, die nicht die Anzahl der Parameter dazuzählen, sondern die Freiheitsgrade abziehen. Das ergibt unterschiedliche Werte, aber die relevanten Relationalen bleiben erhalten (Kline, 2011, S. 220).

Der BIC ergibt ebenfalls einen deutlich größeren Koeffizienten als der AIC, allerdings werden im Gegensatz zum cAIC die n_p nur einfach gewichtet:

$$BIC = -2\log L + (\log N) * n_p$$

Im adjustierten BIC wird hingegen der Stichprobenumfang umformuliert:

$$saBIC = -2\log L + (\log((N+2)/24)) * n_p$$

Für alle Maße gilt, dass ein kleinerer Wert, also eine geringere Distanz, eine bessere Passung bedeutet, aber die unterschiedlichen Gewichte lassen erahnen, dass sich abgeleitete Bevorzugungen des einen oder des anderen Modells bei der Verwendung unterschiedlicher Maße umkehren können. Zwar lassen sich über die Informationskriterien relative Werte der Modellanpassung einfach miteinander verknüpfen, aber es zeigt sich ebenso, dass eine gewisse Beliebigkeit bei der Auswahl eines Index herrscht (Rost, 1996, S. 329). Als generelle Empfehlung formuliert Rost, dass der AIC bei kleiner Itemanzahl mit großen Musterhäufigkeiten und der BIC bei großer Itemanzahl mit geringen Musterhäufigkeiten verwendet werden sollte. Dziak et al. (2015) untersuchten die Indizes für Faktoranalysen und Klassenanalysen und stellten analog fest, dass der AIC und seine Abwandlungen häufig große Modelle, während der BIC und seine Abweichungen vornehmlich kleine Modelle bevorzugen. Tofighi und Enders (2008) stellten in einer Untersuchung zu der Performanz der Kriterien fest, dass der saBIC von diesen Werten am besten unter verschiedenen Rahmenbedingungen in der Lage war, die korrekte Klassenzahl von Mischverteilungsmodellen mit Wachstumskomponenten konservativ, also unter der Bevorzugung eines sparsameren Modells, zu schätzen. Da für große Stichproben die Gefahr größer ist, ein Modell zu überspezifizieren, wird hier der BIC empfohlen und vice versa. Alle weiteren Bedingungen sind vom wahren Modell abhängig, welches in Simulationsstudien definiert werden konnte, aber in der Realität normalerweise nicht vorliegt. Die Autorinnen und Autoren fassten zusammen, dass wenn der AIC anzeigt, dass ein Modell zu komplex ist, dies vermutlich die Daten nicht gut repräsentiert. Auf der anderen Seite kann angenommen werden, dass, wenn der BIC ein zu kleines Modell anzeigt, es vermutlich zu klein ist oder schlecht zu den Daten passt (ebd., S. 16). Mulaik (2001) diskutiert hingegen, dass, da die Strafterme nicht für alle Modelltypen funktional sind, diese gar nicht für Modellvergleiche verwendet werden sollten, da die Werte auf reinen Kurvenanpassungen beruhen und kein Kriterium besitzen, wann Regelmäßigkeiten in den Daten durch das Modell angemessen abgebildet werden. Damit wiederspricht er Rost (1996) und lehnt die Annahme ab, dass Informationskriterien nützlich sein können, um die Gültigkeit eines Modells zu bewerten. Der Einwand kann bedingt nachvollzogen werden, da diese Werte als harte Kriterien nur begrenzt einsetzbar sind, aber trotzdem ist deren Verwendung als konditionale Orientierungsmaße in Paarvergleichen von Modellen üblich und nützlich.

Eingesetzt werden relative Maße, zum Beispiel bei der Exploration der optimalen Klassenzahl in LCA/LPA. Dabei ist es üblich, mehrere Modelle mit steigender Anzahl von Klassen sukzessiv gegeneinander zu prüfen. In der Nomenklatur der Clusteranalyse können internale oder externale Maße dazu verwendet werden (Bacher et al., 2010). Die Nutzung der $logL$, der IC und ergänzender Maße ist dazu etabliert. Bei derartigen Prüfungen ist das Idealszenario, dass alle

Indizes gleichzeitig ein Optimum anzeigen, sodass die Klassenzahl multikriterial bestimmt werden kann. Üblicher sind allerdings Szenarien, in denen die Indizes unterschiedliche Optima bestimmen oder beständig weiter absinken, bis die Zahl der Klassen so groß wird, dass entweder sehr kleine Klassengrößen auftreten oder die Klassen sich nicht mehr inhaltlich voneinander unterscheiden lassen (z.B. Schurig et al., 2015). Hier gibt es neben den IC verschiedene Techniken, um trotzdem auf eine optimale Klassenzahl zu schließen.

Zum einen kann die prozentuale Verbesserung PV gegenüber dem Nullmodell, hier der 1-Klassenlösung, abgeleitet werden. Dabei werden der absolute Wert der $logL$ der entsprechenden Klassenlösung K und der $logL$ der 1-Klassenlösung miteinander verrechnet (Bacher et al., 2010, S. 363).

$$PV_{1,K} = 1 - \frac{|logL_K|}{|logL_1|}$$

Parallel kann zum anderen die Verbesserung gegenüber dem vorangegangenen Modell errechnet werden (ebd.).

$$PV_{k-1,K} = 1 - \frac{|logL_K|}{|logL_{K-1}|}$$

Hier ist es sinnvoll, a priori Schwellen (beispielsweise 0.1%, 1% oder 5%) zu bestimmen, welche angewendet werden sollen (ebd.), damit die Klassenauswahl keinen Beliebigkeitscharakter annimmt.

In der Tabelle 6 ist ein beispielhafter Modellvergleich abgetragen. In dem Beispiel werden latente Klassifikationen von Leistungsprofilen miteinander verglichen, um eine optimale Zahl von Klassen zu identifizieren (Schurig et al., 2015). Der Tabelle kann entnommen werden, dass die berichteten Kriterien immer weiter sinken.[47] Die Argumentationslinie für die Wahl einer Lösung mit sieben Profilen bezieht sich zum einen auf die Änderung der $logL$ und aller Informationskriterien von unter 0.1 Prozent und zum zweiten darauf, dass die Zellbesetzungen in den beiden Randprofilen bei einer Lösung mit acht Klassen auf unter ein Prozent der Grundgesamtheit gefallen ist, so dass diese nicht hinreichend belastbar wären. Ein vergleichbares Vorgehen findet sich bei dem Beitrag von Schurig und Busch (2014), wobei zusätzlich ein längsschnittlicher Abgleich vorgenommen wird.

[47] Die alternativen Werte des AIC und des saBIC verhielten sich analog und wurden daher nicht mit aufgeführt.

Tabelle 6: Modellvergleiche für ein europäisches Referenzmodell von Leistungsprofilen (N = 74868; Schurig et al., 2015, S. 43)

Modell	Anzahl der Leistungsprofile K	Mittlere $-2logL$	BIC	cAIC	Anzahl der Parameter n_p
I	3	-1243507	2487170	2487041	14
II	4	-1231360	2462922	2462756	18
III	5	-1224201	2448649	2448446	22
IV	6	-1220405	2441102	2440862	26
V	7	-1218364	2437065	2436789	30
VI	8	-1217217	2434816	2434503	34

Alternativ können für relative Modellvergleiche Likelihood-Verhältnisstatistiken (Likelihood-Quotienten-Test oder auch Likelihood-Ratio-Tests) verwendet werden, um signifikante Veränderungen zwischen zu vergleichenden Modellen zu identifizieren, wenn eine hierarchische Beziehung zwischen den Modellen besteht (Rost, 1996, S. 332). Diese Tests werden aber nicht mehr empfohlen, da die Likelihood-Differenzen keiner χ^2-Verteilung folgen (Bacher et al., 2010, S. 365; McLachlan & Peel, 2000, S. 185ff; Rost, 1996, S. 332). Eine Adjustierung existiert in Form des Vuong-Lo-Mendell Rubin-Tests (Lo, Mendell & Rubin, 2001), welcher in verschiedenen Anwendungen (z. B. MPlus) implementiert ist. Eine gangbare Alternative, um Signifikanz–aussagen machen zu können, sind Bootstrap-Verfahren (vgl. Kap. 5.2.1.2). Likelihood-Ratio-Tests auf Basis von Bootstraps scheinen die gewünschten Eigenschaften aufzuweisen[48]. Ein parametrisches Bootstrap Verfahren ist bei McLachlan und Peel (2000) vorgestellt und wird bei Nylund, Asparouhov und Muthén (2007) positiv evaluiert. Es verbleiben zwar die bereits diskutierten Probleme von χ^2-Tests, also dass diese für große Stichproben keine belastbaren Ergebnisse erbringen, aber da die Relation von Werten zueinander bewertet wird und nicht die jeweilige Schwelle zur Signifikanz, ist das Problem weniger relevant.

5.2.3. LOKALE GÜTEMAßE

Globale Fit-Statistiken kombinieren eine große Zahl von Diskrepanzen in einem einzelnen Wert, sodass lokale Diskrepanzen unter Umständen verdeckt bleiben (Steiger, 2007). Daher ist es bedeutsam, nicht nur Fit-Kriterien zu evaluieren, sondern alle Modellparameter für eine Prüfung des Gesamtmodells hinzuzuziehen. Lokale Prüfungen können differenziert darüber Aufschluss geben, ob und in welchem Umfang Abweichungen aufgrund von beobachteten Variablen,

[48] Diese weisen erhöhte Anforderungen an die Rechenleistung auf, welche aber auf aktuellen Computersystemen zumeist problemlos erreicht werden.

unbeobachteten Variablen oder sogar Testpersonen auftreten. Dabei ist die Beobachtung von differenziertem Antwortverhalten von besonderer Bedeutung.

Personenfit

Differenziertes Antwortverhalten kann sowohl als Eigenschaft der Items als auch als Eigenschaft der Personen(gruppen) betrachtet werden. Während in der Folge zentral die Eigenschaften von Items fokussiert werden, beschäftigt sich dieser Abschnitt in der Folge mit der Anpassung von Personen an ein Modell. Der Person-Fit wird über die Wahrscheinlichkeit des Zustandekommens eines Antwortmusters, für IRM bei gegebenem Summenscore r_v der Person $p(a_v|r_v)$[49], abgeleitet und kann mittels der Annahme des Modells verteilter Prüfgrößen (z. B. einer Normalverteilung für IRM) auf Signifikanz getestet werden (vgl. Meijer & Sijtsma, 2001). Der „Fit" von Personen auf ein statistisches Modell muss dabei kritisch gesehen werden, widerspricht er doch gängigen Annahmen der Modelltestung in den Sozialwissenschaften, in denen die Antworten der Personen theoretisch die Realität wiedergeben. Statistisch-mathematisch sind Personen und Items aber exakt das Gleiche: Es handelt sich um variierende oder nicht variierende Modellparameter. Nützlich ist das Vorgehen beispielsweise als Reflexion auf die Trennschärfe eines IRM insbesondere im obersten und untersten Testbereich. Rost (1996, S. 352) führt aus, dass der Messwert einer Person umso exakter ist, umso weniger er vom wahren Parameter abweicht. Der Anteil, den jedes Item beiträgt, ist aber unterschiedlich groß, da unterschiedlich viel Varianz beigesteuert wird. Leichte und schwere Items haben weniger Varianz, da sie von mehr respektive weniger Personen gelöst werden. Unter der Normalverteilungsannahme der wahren Verteilung der Personenparameter tragen Items mit mittleren Lösungswahrscheinlichkeiten am meisten Varianz bei. Da ein Test durch Verlängerung immer verbessert werden kann (ebd.), werden die Messeigenschaften also erhöht, wenn Items in entsprechenden Schwierigkeitslagen (vgl. Kap. 4.2) hinzugegeben werden. Für die Evaluation von Personen ist das angeführte Maß also für die Feststellung auffälliger Antwortmuster nützlich. Das Prüfverfahren ist dabei dasselbe wie für die Analyse von Items (s. u.); die Frage ist nur, inwieweit die Fähigkeitsverteilung oder die Verteilung der Schwierigkeitsparameter als Prüfgröße gewählt wird.

Personenheterogenität

Die Beobachtung von Personenheterogenität ist eine weitere Möglichkeit, Abweichungen des Modells von der Realität zu erklären. Diese hat insbesondere vor dem Hintergrund von internationalen Vergleichsstudien an Bedeutung gewonnen und kann als Teilbereich von Messinvarianzanalysen (M*easurement Invariance*; MI) gesehen werden. Diese beschäftigen sich mit der Frage, ob Operationalisierungen unter verschiedenen Bedingungen funktional, also vergleichbar sind (Cheung & Rensvold, 2002; Meade & Lautenschlager, 2004). Diese Prüfungen können über den Vergleich konfirmatorischer Faktormodelle hierarchisch für verschiedene Gruppen von Parametern erfolgen. Dabei werden die Intercepts der unstandardisierten Faktorladungen, die Faktorladungen, die Faktormittelwerte und gegebenenfalls auch die Residualvarianzen (Vandenberg & Lance, 2000) zwischen den interessierenden Gruppen fixiert

[49] Diese Formel gilt nur für dichotome IRM, eine polytome Ableitung findet sich z. B. bei Glas und Dagohoy (2007).

und die Abweichungen über den Vergleich von Gütemaßen bewertet (vgl. Kap. 5.2.1.2). Spezifische Abweichungen für einzelne Items zwischen Personengruppen können zudem über die Prüfung auf partielle Messinvarianz festgestellt werden (Byrne, Shavelson & Muthén, 1989). Eine exemplarische Anwendung von Messinvarianzanalysen zwischen vier Schulfächern und über vier Schuljahre hinweg findet sich bei Schurig et al. (2016). Auf Basis der festgestellten Grade der Invarianz konnte eine hohe Generalisierbarkeit der angenommenen mehrdimensionalen latenten Struktur festgestellt werden, welche als Evidenz für externe Konstruktvalidität interpretiert werden kann. Somit könnten auf Basis dieser Ergebnisse Vergleiche der latenten Strukturen zwischen Klassenstufen und sogar Fächern vorgenommen werden.

Im Rahmen von IRM gibt es die Möglichkeit, personen- und itembasierte Verzerrungen auf der Basis der Item-Response-Funktionen (IRF) festzustellen. So können Verzerrungen auch für Personen und einzelne Antwortkategorien oder verschiedene Itemstatistiken wie der Schwierigkeit, der Ratewahrscheinlichkeit oder der Diskrimination festgestellt werden (Thissen, Steinberg & Gerrard, 1986), ebenso wie der Itemfit zwischen Gruppen verglichen werden kann (*Differential Item Functioning;* DIF). Die Bedeutung zeigt sich in der in den letzten Jahren angestiegenen Forderung nach der Prüfung auf Messinvarianz (z. B. Schulte, Nonte & Schwippert, 2013; van de Schoot, Schmidt, Beuckelaer, Lek & Zondervan-Zwijnenburg, 2015; Schurig et al., 2016) und DIF in IRM (z.B. Klieme & Baumert, 2001; McElvany & Schwabe, 2013).

Itemqualität

In der Folge soll vertiefend auf Itemanalysen eingegangen werden, da diese normalerweise durch die Forschenden leichter manipulierbar sind und im Sinne der eigentlichen Modellevaluation eine höhere Relevanz haben. Es gilt grundlegend, dass die Anzahl und die Qualität der verwendeten Items eine hohe Relevanz haben, welche über die notwendige Zahl für die Modellidentifikation hinausgeht. Bei der Vorstellung der verschiedenen Evaluations–strategien wird häufig unterstellt, dass Indikatoren mit hoher Qualität ausgewählt wurden. Die Qualität der Items wird dabei über die Trennschärfe, also die Faktorladungen, bestimmt. In IRM und Faktormodellen sollte diese hoch und homogen sein, und in LCA/LPA sollten diese idealtypisch je nach Klasse nahe 1 oder 0 sein (Wurpts & Geiser, 2014). Grundsätzlich wird angenommen, dass die Verwendung von Indikatoren mit hoher Qualität vorteilhaft ist (Hayduk & Littvay, 2012; Marsh, Hau, Balla & Grayson, 1998; Wurpts & Geiser, 2014). Doch während Marsh et al. (1998) im Sinne maximaler Varianzaufklärung dafür argumentieren, im Rahmen von konfirmatorischen Faktormodellen möglichst viele Indikatoren zu verwenden, vertreten Hayduk und Littvay (2012) die Position, im selben Analyserahmen nur die wenigen besten Indikatoren zu benutzen, um Konfundierungen besser entgegentreten zu können. Beide Argumentationslinien fokussieren allerdings auf die Bearbeitung der a posteriori Definitionen der latenten Variablen, und die theoretische Fundierung der a priori Definition spielt eine unzureichende Rolle. In der Strategie nach Marsh und Kolleginnen und Kollegen könnten nahezu beliebige Items eingesetzt werden. Solange diese zumindest marginal mit allen anderen Items korrelieren, würde die Messung „verbessert“, und bei Hayduk und Littvay muss angenommen werden, dass die Items inhaltlich austauschbar oder redundant sind. Diese Annahme wiederum wurde auch testtheoretisch mit dem Schritt von τ-äquivalanten zum kongenerischen Modell

ausdrücklich fallen gelassen, sodass jeder Schritt in Richtung einer Aufblähung oder künstliche Homogenisierung letztlich im schlimmsten Fall das zu testende oder messende Theorem verknappen, also etwas Irrelevantes präzise getestet oder gemessen wird. Festgehalten werden soll also, dass mindestens eine zur Modellidentifikation notwendige Zahl von Items ausgewählt werden sollte, die messtheoretisch qualitativ mindestens so gut sein müssen, dass das Modell in globalen Modellbeurteilungen angenommen werden kann, ohne dabei das inhaltliche Theorem zu berühren. Sollte dies nicht möglich sein, müssen entweder die Indikatoren oder auch das Theorem hinterfragt werden.

Itemfit

In Item- und DIF-Analysen wird üblicherweise versucht „schlechte" oder besonders heterogene Items zu isolieren. Dazu ist es zentral, das Ausmaß, also die besagte Trennschärfe, festzustellen, in dem die Beantwortung eines einzelnen Items mit der Personenfähigkeit zusammenhängt. Insofern alternativ Itemheterogenität vorliegt, können die Items im Rahmen von SEM oder IRM mehrdimensional sein, ohne dass dies modellbasiert hinreichend berücksichtigt wurde; dem kann faktoranalytisch nachgegangen werden. In Faktormodellen kann die Trennschärfe über die Ladungsstrukturen abgeleitet werden und in Klassenanalysen über die summierte Differenz der Items zwischen den Klassen. Der Itemfit hat in den IRM eine besonders gewichtige Bedeutung, da häufig keine hinreichenden globalen Modelltests zur Verfügung stehen. In IRM kann die Trennschärfe beispielsweise über den Q-Index (Tarnai & Rost, 1990) oder die mittleren Abweichungsquadrate (MNSQ; Bond & Fox, 2001) nachgezeichnet werden; beide sind konzeptionell ähnlich. In gängigen Softwarelösungen hat sich die Verwendung des MNSQ etabliert. Eine umfassende Darstellung der Herleitung und Interpretation findet sich bei Wright und Masters (1982, 84f). Der MNSQ kann als Maß der Zufälligkeit oder (Über)Determiniertheit von Items betrachtet werden. Der Erwartungswert ist dabei 1 und Werte kleiner als 1 weisen einen *Overfit* auf, was bedeutet, dass Werte „zu gut" vorhergesagt werden können. Werte größer 1 zeigen eine unvorhersehbare Varianz oder einen *Underfit* an. Da sich die Werte auf einer Verhältnisskala bewegen, zeigt beispielsweise ein MNSQ von 1.2 20 Prozent nicht aufgeklärte Varianz an. Eine standardisierte MNSQ-Form kann als t-Werte begriffen und direkt verwendet werden, um die Signifikanz einer Abweichung der Funktion des Items zum Modell zu bestimmen. Dabei können zwei Formen des MNSQ zum Einsatz kommen, deren Angemessenheit sich nach dem Verhältnis der Schwierigkeit des Items und der Personenfähigkeit richtet. Statistisch entsprechen mittlere Abweichungsquadrate χ^2-Werten, welche durch die Freiheitsgrade geteilt werden. Der *Outfit* basiert auf dem Verhältnis der Summe der quadrierten standardisierten Abweichungen sowie den Freiheitsgraden und kann als Maß verstanden werden, welches insbesondere die Güte des Items für Personen beschreibt, deren Fähigkeiten zu der Schwierigkeit der Items passen. Der *Infit* ist sensitiver für Ausreißer, da hier eine Gewichtung für unzutreffende Antworten stattfindet. Eine vertiefende Darstellung der Anwendung der Werte sowie von Schwellenparametern 0.5 und 2.0, findet sich bei Linacre (2002).

Generell werden die Schwellenparameter für die harte „Annahme" und „Ablehnung" eines Items kritisch diskutiert, denn die Werte beruhen stark auf dem vorliegenden Stichprobenumfang und sind im Bildungskontext beispielsweise vor dem Hintergrund curricularer Angemessenheit

eines Tests inhaltlich nicht austauschbar. Smith, Schumacker und Bush (1998) argumentieren mit Hinweis auf Wright, dass die Verwendung von harten Schwellenwerten unangemessen ist und stattdessen stichprobenspezifische Schwellenwerte abgeleitet werden könnten, welche den Stichprobenumfang berücksichtigen und schärfere Schwellen für kleine Stichproben oder Itemzahlen x vorgeben, dabei kann als x entweder das eine oder das andere eingesetzt werden.

$$Schwellenparameter\ fuer\ gewichteten\ MNSQ\ (infit) = 1 \pm \frac{2}{\sqrt{x}}$$

$$Schwellenparameter\ fuer\ ungewichteten\ MNSQ\ (outfit) = 1 \pm \frac{6}{\sqrt{x}}$$

In internationalen Schulleistungsstudien wurde übereinstimmend der MNSQ mit Schwellenwerten von .8 und 1.2 verwendet (z. B. IGLU, PISA und TIMSS; z.B. Bond & Fox, 2001; OECD, 2012). Zusammenfassend müssen die Schwellenparameter aber relativ für den Inhalt des Items, den Stichprobenumfang und die gewünschte Tragfähigkeit der Ergebnisse gewählt werden. Beispielsweise muss sichergestellt werden, dass ausreichend viele leichte und schwere Items enthalten sind.

Misfit in IRM

Eine nützliche Übersicht zur Analyse von Item- oder Personenmisfit findet sich auf der Website der Software WinSteps (Linacre, 2016). Diese sollte nicht überbelastet werden, kann aber hilfreich sein, um aus der Kombination der Informationen zu In- und Outfit, sowie Under- und Overfit Rückschlüsse auf mögliche Erklärungen zu finden. Diese sind in der Tabelle 7 abgetragen. So kann beispielsweise aus der Kombination einer übermäßigen Itemanpassung im Infit, also für die Extremgruppe der Testteilnehmerinnen und Testteilnehmer, also dem Testteil, der zugleich per Definition am wenigsten präzise ist, und einem unauffälligen Outfit abgeleitet werden, dass eventuell eine Redundanz des Items in dessen Schwierigkeitslage vorliegt. Das Item misst übermäßig gut für Personen, die sich in der entsprechenden Schwierigkeitslage bewegen und ist für Personen außerhalb der üblichen Wertespannweite unauffällig.

Tabelle 7: Beurteilung von Personen- oder Itemfit in IRM (Linacre, 2016)

Analyseeinheit	Infit	Outfit	Erklärung
Generell	Underfit	Underfit	Keine Konvergenz/mangelnde Testpräzision
Item	Underfit	Underfit	schlechtes/unpassendes/unausgewogenes Item
	Underfit	Overfit	unbeobachtete Interaktion
	Overfit	Overfit	Einstimmigkeit/überverwendung des mittleren Bereichs
	Overfit	-	Redundanz
Person	Underfit	Underfit	Überverwendung extremer Kategorien
	Underfit	-	Verarbeitungsfehler/seltsames Verhalten
	Overfit	Overfit	Einstimmigkeit/überverwendung des mittleren Bereichs
Person mit hoher Fähigkeit	-	Underfit	Unaufmerksamkeit
Person mit niedriger Fähigkeit	-	Underfit	Spezialwissen/Raten
	Overfit	-	Besondere Aufmerksamkeit

Extreme Maße

Neben der Verwendung der Antworten der Testteilnehmerinnen und Testteilnehmer und der Itemeigenschaften zur lokalen Gütebestimmung gibt es die Möglichkeit, die Verteilung der latenten Variablen zu überprüfen. Eine Prüfung, welche die Logik der Prüfung des Modells auf die Daten umkehrt, ist die Analyse von Ausreißern. Dabei werden Objekte oder Cluster identifiziert, welche nicht zu dem Modell passend erscheinen. Das entspricht der formalen Prüfung der Passung von Personen zu dem Modell über den Personenfit in IRM und kann nützlich sein, um unerwünschtes Antwortverhalten zu identifizieren. Dies kann bei Modellen mit kontinuierlichen latenten Variablen vorgenommen werden, indem über das abgeleitete Maß Extremfälle identifiziert und gegebenenfalls ausgeschlossen werden. Beispielsweise sind Ausreißer in der grafischen Darstellung als Boxplots (Tukey, 1977, S. 39ff) üblicherweise als Datenpunkte definiert, die weiter als das 1.5-fache des Interquartilsabstandes zwischen dem 25 Prozent- und dem 75 Prozent-Quartil von der Box entfernt sind, also außerhalb der Whisker liegen.

Für Klassenmodelle können wiederum schlecht zugeordnete Personen oder Objekte oder schlecht definierte Cluster direkt über die Klassenzugehörigkeitswahrscheinlichkeiten und deren Mittel innerhalb der Cluster abgeleitet werden. Es ist möglich, hier Schwellenwerte für mindeste Zuordnungssicherheiten zu setzen, sodass beispielsweise zugeordnete Objekte mit einer im Verhältnis zur Klassenzahl geringen Zuordnungswahrscheinlichkeit als Ausreißer oder

Rauschen begriffen werden, das Vorgehen ist dabei allerdings normativ. Darüber hinaus wird durch dieses Vorgehen die Annahme exhautiver Klassen [50] verletzt, welche latenten Klassifikationsmodellen zugrunde liegen. Es obliegt der oder dem Forschenden, zu definieren, ob Ausreißer existieren und in der Folge zu entscheiden, ob diese wegen Eigenschaften des Modells, also Messungenauigkeiten, oder wegen Eigenschaften der Messobjekte, also zum Beispiel wegen absichtlich verzerrtem Antwortverhalten, auftraten. Dementsprechend muss entschieden werden, ob diese auszuschließen sind.

Modifikationsindizes

Zuletzt soll eine Möglichkeit eingeführt werden, Modelle post hoc zu verbessern: die Inspektion der sogenannten Modifikationsindizes (MacCallum, 1986). Zwar sollten Modelle idealtypisch a priori hinreichend spezifiziert sein, aber dies ist häufig in der Praxis nicht möglich, entweder weil die Theorie nicht ausreichend klar (Raykov & Marcoulides, 2006, S. 49ff), oder weil die Performanz der Indikatoren noch unbekannt ist. Modifikationsindizes liegen in verschiedenen Formen vor, beispielsweise als *Expected Parameter Change* (EPC) oder erwartete Änderung im Model Fit (Wald-Index oder Lagrange-Multiplikatoren-Test). Im Grundsatz soll geschätzt werden, welche Änderungen sich ergeben würden, wenn Parameter, beispielsweise Kreuzladungen oder Residualkovarianzen, freigegeben würden. Dabei muss betont werden, dass alle Modifikationen, die dann tatsächlich umgesetzt werden, eine theoretische Begründung benötigen (z.B. Raykov & Marcoulides, 2006, S. 51). MacCallum, Roznowski und Necowitz (1992) geben eine kritische Zusammenfassung zum Einsatz der Indizes und zeigen auf, welche Einschränkungen in der Generalisierbarkeit entstehen, wenn eine Modellmodifikation oder sogar eine Modellgenese zu datengetrieben betrieben wird. Marsh et al. (2009) empfehlen, modifizierte Modelle, welche nicht kreuzvalidiert werden konnten, nur als explorative Modellprüfungen zu verstehen.

Zusammenfassend sollten lokale Abweichungen relativ zu ihrem Gewicht und ihrer Anzahl beurteilt werden. Auch mehrere kleine lokale Abweichungen führen nicht zwangsweise zu einer Ablehnung des Modells, aber auch eine geringe Zahl von großen lokalen Abweichungen muss zu Modifikationen oder einem Verwerfen des Modells führen (Bühner, 2011, S. 430).

[50] Dies meint, dass alle Objekte einer Klasse zugehören müssen, die Klassen für die Stichprobe also erschöpfend sind.

5.3. Zusammenfassung

Ein Rückblick auf den forschungslogischen Ablauf bei Friedrichs (1982, S. 51) erlaubt es, die Strukturierung und Herausstellung der vorgestellten Schritte zur Evaluation statistischer Modelle einzuordnen.

- Der *Entdeckungszusammenhang* beschreibt dabei den Anlass einer Forschungsarbeit und kann beispielsweise in Problemen der Theoriebildung, vorliegenden Befunden oder sozialen Problemen begründet liegen.
- Der *Begründungszusammenhang* stellt sich durch das methodologische Vorgehen vor dem Hintergrund bestehender Theorien dar und hat die Aufgabe eine möglichst exakte, nachprüfbare und objektive (Friedrichs, 1982, S. 53) Prüfung formulierter Hypothesen sicherzustellen.
- Die Ergebnisse werden bei erfolgreicher Prüfung der Hypothesen in den *Verwertungs- und Wirkzusammenhang* übergeben, der sich seinerseits direkt auf den ursprünglichen Entdeckungszusammenhang beziehen muss. Alternativ setzt ein rekursiver Prozess ein, der die vorliegenden Ergebnisse in direkten Kontrast mit den in der Studie vorliegenden Annahmen und Hypothesen bringt.

Schwerpunktmäßig wird in dieser Arbeit der Abschnitt der Operationalisierung, also der Verbindung von den theoretisch angenommenen Regeln und der Beobachtungsebene im Begründungszusammenhang bearbeitet (ebd, 1982, S. 77). Der interessierende wissenschaftliche Begriff wird im Bereich der Operationalisierung dabei auf logisch verbundene Indikatoren reduziert, es greift die pragmatische Erkenntnis, dass „Wissenschaftler gezwungen sein können Begriffe von hoher Allgemeinheit zu definieren, um nicht die Begriffe der Wissenschaft ständig neu definieren zu müssen." (ebd., S. 80). Die Indikatoren werden ihrerseits aus der Theorie des Objektbereichs, dem Begründungszusammenhang, abgeleitet. Dies entspricht dem Vorgehen bei der Formulierung von Modellannahmen im statistischen Sinne. Das Ziel der Operationalisierung muss im statistischen Sinne dabei eine zeit- und gelegenheitsstabile Realitätsabbildung sein, so dass weitere Prüfungen zu einem späteren Zeitpunkt, an einer anderen Population oder mit einer anderen Methode möglich werden, ohne dass dabei der wissenschaftliche Begriff inhaltlich berührt wird. Die Operationalisierung im forschungslogischen Ablauf wird Prüfungen unterworfen und nach hinreicheden Falsifikationsversuchen in den Verwertungszusammenhang überführt oder an die zugrunde legende Theorie zurück gespiegelt. Diese rekursive Denkweise bei einer Zurückweisung der Hypothesen liegt auch der Logik der statistischen Modellanpassungen, also der Spezifikation und Re-Spezifikation von Modellen und deren Prüfung, zugrunde. Dies gilt für Vorhersagemodelle, also Strukturmodelle, ebenso wie für Erklärungsmodelle, also Messmodelle.

Im vorangegangenen Abschnitt wurden Regeln der Modellidentifikation, Strategien und Techniken der Modellbeurteilung vorgestellt. Da die Modellidentifikation ein grundlegendes Konzept ist, sollte diese bereits im Modellierungsprozess mitgedacht werden. Auf das Problem der empirischen Unteridentifikation hin kann kontrolliert und korrigiert werden. Hingegen ist die Evaluation der Güte eines Modells komplexer. Es wurde ein Raster von global absoluten,

global relativen und lokalen Modellprüfungen und Strategien der Modellbeurteilung vorgestellt, und innerhalb jeder Klasse wurden gängige Maße für verschiedene Modelltypen expliziert.

Es konnte festgestellt werden, dass es in der Modellbeurteilung viele *„best practice“* Ansätze, aber wenig absolute Sicherheit gibt. Während es wünschenswert wäre, dass Modelle mittels eines einzelnen Tests sicher zu bestätigen oder abzulehnen wären, muss festgestellt werden, dass verschiedene Probleme beim Einsatz von Hypothesentests zur Beurteilung der Passung eines Modells zu den Daten vorliegen, wie zum Beispiel der Stichprobenumfang oder die mittlere Korrelation zwischen den beobachteten Variablen. Aber auch wenn der Test technisch angemessen ist, muss hinterfragt werden, ob ein Hypothesentest in einer deterministischen Logik zu der Frage „Kann die Abweichung zu einem erwarteten Wert, welcher durch die Nullhypothese definiert wurde, als zufällig betrachtet werden?“ in so einem komplexen Setting nützlich ist (Sedlmeier, 2009, S. 2). Eine Alternative sind approximative Güte- und Überlappungsmaße. Dabei existiert neben den hier vorgestellten Maßen eine große Zahl weiter gut erforschter Maße, die in dieser Arbeit nicht behandelt werden konnten sowie eine noch größere Zahl von Evaluationsmaßen über deren Verhaltensweisen in spezifischen Testsituationen noch keine hinreichende Sicherheit besteht.

Es gibt umfangreiche Kontroversen, über die Verwendung insbesondere approximativer Maße und verschiedene Autorinnen und Autoren argumentieren, dass im Falle von SEM nur der χ^2-Wert interpretiert werden sollte (z.B. Barrett, 2007). Zumeist wird aber angenommen, dass die Maße einen hohen Wert haben, obwohl es sich um qualitative Maße und nicht etwa um harte Kriterien handelt. Dies sollte nicht als „unwissenschaftlich“ betrachtet werden, denn die Evaluation jedweder wissenschaftlichen Erkenntnis unterliegt einigen Graden von Subjektivität (Kline, 2011, S. 191). Für approximative absolute Gütemaße und die Überlappung von Clusterlösungen wurde festgehalten, dass diese Indizien für die Angemessenheit eines Modells liefern können, wenn Hypothesentests nicht angemessen eingesetzt werden können. Aber die Schwellen, welche herangezogen werden können, um eine Passung zu bewerten, sind umstritten (Barrett, 2007; Millsap, 2007). Zudem weisen diese unterschiedliche Grade von Sensitivität für spezifische Situationen auf, so dass deren Schwellenwerte immer unter der Berücksichtigung der Rahmenbedingungen zu betrachten sind (Kenny et al., 1998). Entscheidungsregeln sind nützlich, aber eine strikte Orientierung an diesen kann in einigen Situationen unpassend sein, somit sollten sie als Daumenregeln betrachtet werden (Hu & Bentler, 1999, S. 4). Skrondal und Rabe-Hesketh (2004, S. 280) schreiben dazu: „Research on diagnostics for latent variable models still appears to be in its infacy, especially for models with noncontinous responses“. Dies bedeutet, dass der Forschende häufig in der Verantwortung ist, zu erläutern, warum eine akzeptable Misspezifikation angenommen wurde. Ein vielzitierter Satz, der von George Box geprägt wurde, wurde bei Box und Draper (1987, S. 424) wie folgt wiederholt: „Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful.“ Millsap (2007) argumentiert, dass dafür a priori definiert werden muss, wann es sich um eine akzeptable Fehlpezifikation handelt, was aber weitere Forschung im Bereich der Funktionsweise von Fit Werten, umfassende statistische Kenntnisse und exakte Kenntnisse hinsichtlich der erwarteten Ergebnisse bei Anwenderinnen und Anwendern erfordert. Damit ist diese Forderung häufig unrealistisch. Beim Einsatz und bei der Bewertung der verschiedenen Indizes ist es daher

von entscheidender Bedeutung, multikriterial vorzugehen und sich zu vergegenwärtigen, welche impliziten Eigenschaften die einzelnen Maße aufweisen. Auf diese Weise kann begründet entschieden werden, welche Indizes besonders geeignet sind, um eine Modellpassung zu bewerten, und welche Maße situativ weniger geeignet oder sogar ungeeignet sind.

Dasselbe Schema muss bei der Beurteilung von Modellvergleichen auf der Basis relativer Indizes herangezogen werden, wobei zudem mitgedacht werden muss, welche Modelle aus einem hypothetischen Modelluniversum mit dem angenommenen Modell verglichen werden sollten, um die Angemessenheit zu erörtern. Während es bei Klassenmodellen naheliegend ist, eine Spannweite von theoretisch angemessenen alternativen Anzahlen von latenten Klassen zu verarbeiten, so sollten bei Vergleichen von IRT und SEM neben dem Nullmodell theoretisch plausible Alternativmodelle und alternative Dimensionierungen geprüft werden. Auch und insbesondere weil aufgezeigt wurde, dass dies eine empirische Möglichkeit ist, um Evidenz für Validität zu erbringen (vgl. Kap. 4.3).

Absolute Modellgütekriterien ermöglichen einen zusammenfassenden Blick auf Modelle, die aus Dutzenden von Parametern zusammengesetzt sein können, aber die Existenz dieser Kriterien befreit die Forschenden nicht von der Notwendigkeit, auch lokale Maße zu inspizieren. Letztlich sind alle absoluten Kriterien Aggregate lokaler Abweichungen des Modells zu den Daten und eine Aggregation kann immer zu Informationsverlust führen. Die vollständige Prüfung eines komplexen Modells kann demnach nicht allein über eine Inspektion kriterial abhängiger Indizes erfolgen, sondern ist ein Prozess, der sich mit jeder einzelnen Komponente des Modells, der Daten, den Zusammenhängen zwischen den Daten und dem Modell und den Zusammenhängen im Modell befasst. Dies kann selbstverständlich in Veröffentlichungen nicht vollständig nachgezeichnet werden, aber im Sinne guter wissenschaftlicher Praxis muss in Veröffentlichungen zumindest erwähnt werden, welche Schritte der Modellevaluation neben der Inspektion von wenigen Maßen, welche alle manipulierbare Charakteristika aufweisen, vorgenommen wurden.

In den bisherigen Ausführungen und Beispielen wurde ein Schwerpunkt auf eindimensionale Modelle gelegt, die eine Mehrebenenstruktur nicht mitaufgreifen. Insofern aber verschiedene Modelle kombiniert werden, sollte gegeben sein, dass jedes einzelne funktional ist. Dies gilt für jede einzelne Dimension eines Modells mit mehreren latenten Variablen ebenso wie für Modelle mit mehreren Ebenen. Jede Komponente eines umfassenden Modells, welche ihrerseits ein Messmodell darstellt, ist dabei zu berücksichtigen. Bollen und Long (1993, S. 7) schreiben dazu: „It is important to remember that even a model with excellent overall fit indices can be unacceptable because of the components of the model. “ Es ist technisch möglich, dass der Gesamtfit Probleme mit Teilmodellen überschattet, was eine Überschätzung der Modellanpassung zur Folge hätte und in verzerrten Ergebnissen resultieren könnte.

Im Folgenden werden drei Heuristiken der Modellprüfung vorgestellt, welche die in dieser Arbeit aufgeführten Indizes, Tests und Maße nach ihrer Funktionsweise kategorisieren. Jeweils eine für FA, IRM und LCA/LPA, wobei mehrere Maße in verschiedenen Modelltypen verwendet werden können. Zum Umgang mit diesen Maßen, gegebenenfalls Problemen mit derer Anwendung und entsprechenden Quellangaben wird auf das Kapitel 5.2 verwiesen.

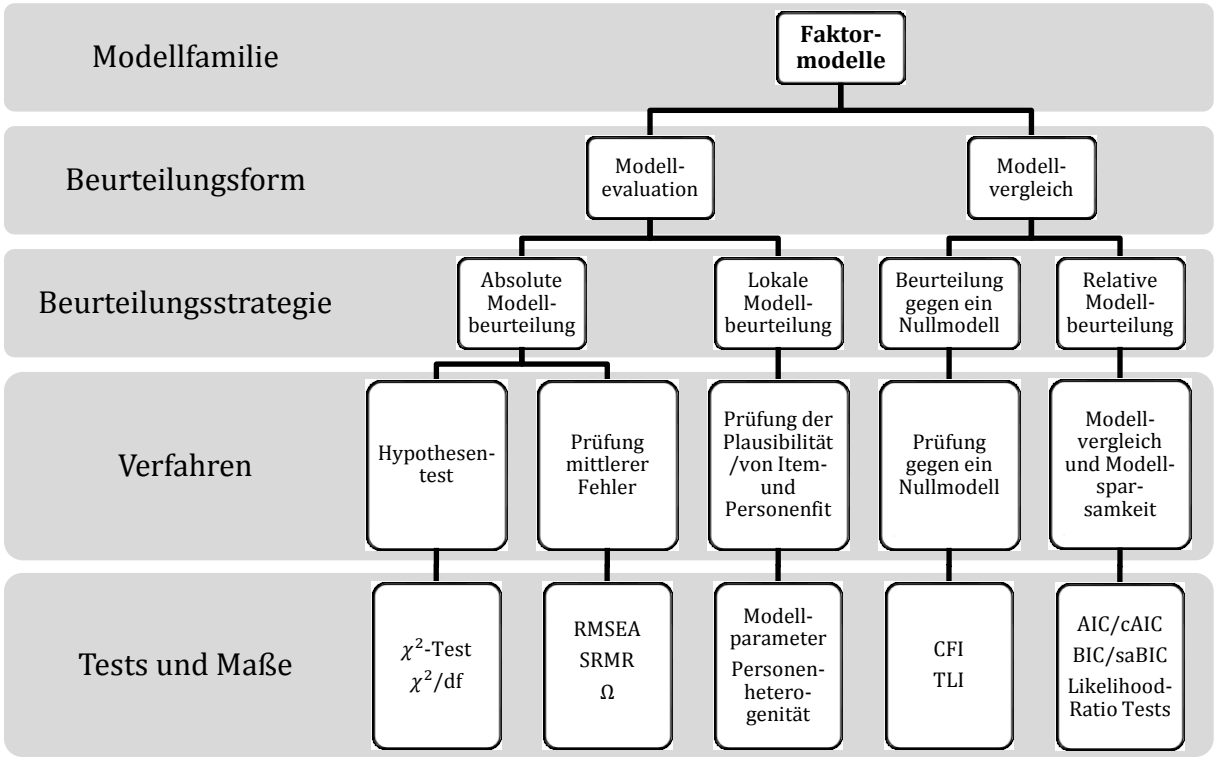


Abbildung 6: Modellbeurteilungsheuristik für Faktoranalytische Modelle

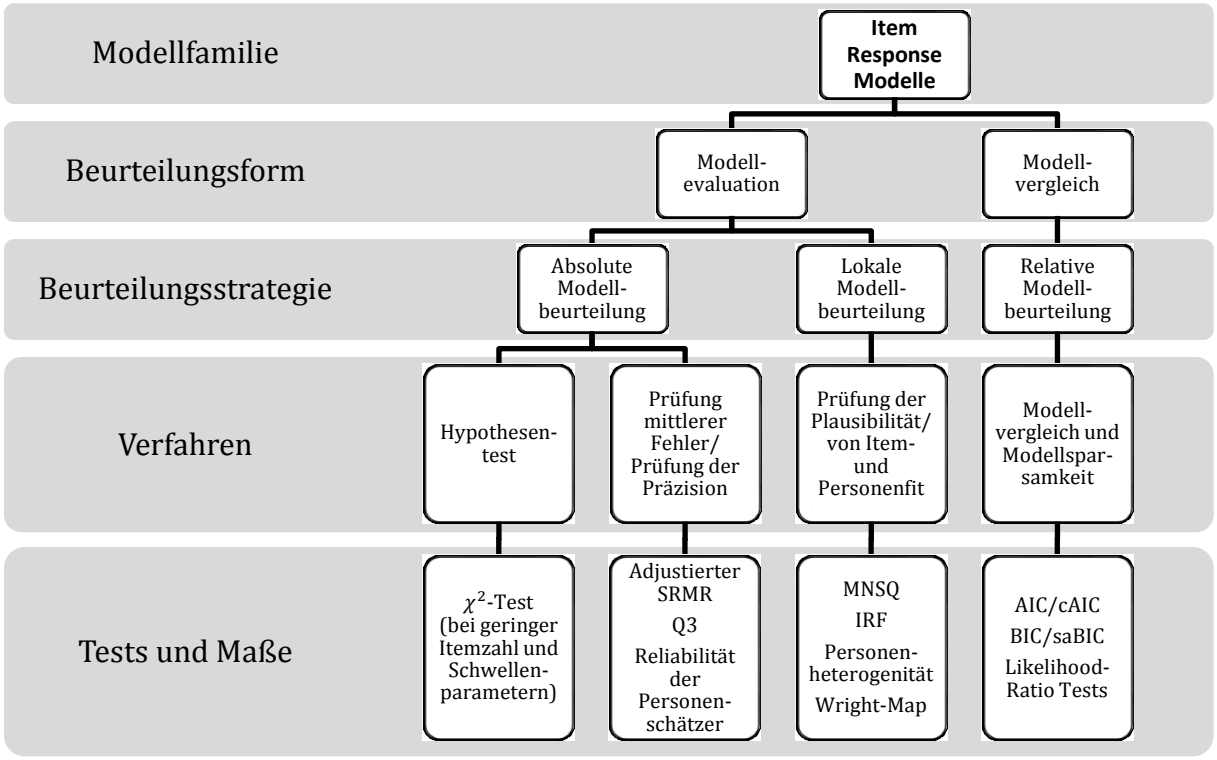


Abbildung 7: Modellbeurteilungsheurtisik für Item Response Modelle

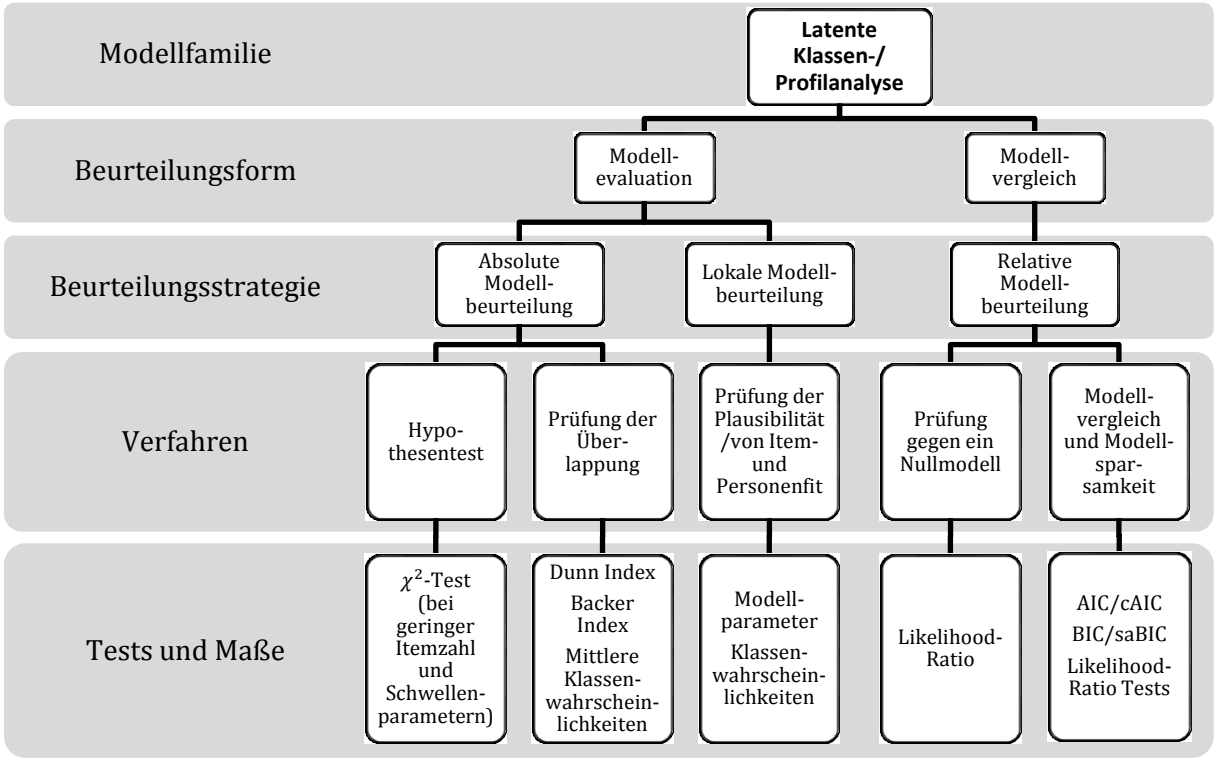


Abbildung 8: Modellbeurteilungsheuristik für Latente Klassenmodelle und Latente Profilmodelle

Diese Aufführungen stellen keine vollständige Auflistung der Modellprüfungsstrategien dar, sondern bieten ein Raster, in das weitere Strategien implementiert werden. Für alle Modellprüfungen wird empfohlen, dass mindestens Maße der Beurteilungsformen der Modellevaluation und des Modellvergleiches aufgeführt werden, besser ist es, wenn mindestens Maße jeder Beurteilungsstrategie herangeführt werden. Eine Zuordnung der Tests und Maße gegenüber den Hauptgütekriterien ist nur bedingt möglich, da gegebenenfalls Überschneidungen vorliegen, aber es kann festgehalten werden, dass Maße der Absoluten Modellbeurteilung maßgeblich geeignet sind, um neben der Modellidentifikation die Objektivität der Modelle sicher zu stellen. Fehlerzentrierte Maße sowie Überlappungs- oder Präzisionsmaße, Prüfungen der lokalen Abweichungen von Modellannahmen und Prüfungen gegenüber einem Nullmodell können verwendet werden, um auf die Reliabilität eines Modells zu schließen. Dies kann bedingt auch als Evidenz für interne Konstruktvalidität verstanden werden. Relative Modellbeurteilungen gegenüber alternativen Modellen oder Maßen können Evidenz für vorliegende externe Konstruktvalidität erbringen.

„Fit“ ist häufig das Kriterium auf das Forschende hinarbeiten. Diesen herzustellen ist aber nicht besonders beeindruckend, da jedes Modell dazu gebracht werden kann, hinreichenden Fit aufzuweisen, sobald beispielsweise ausreichend viele Freiheitsgrade entfernt, also die Restriktionen gelockert werden oder, je nach Bewertungsmaßestab, „nutzlose“ Hypothesen

mitgetestet werden[51]. Solche Modelle haben aber kaum wissenschaftlichen Nutzen (Kline, 2011, S. 189). Eine nachgewiesene Modellgültigkeit macht außerdem keinerlei Aussage zu dem substanziellen Gewicht der latenten Variablen. Die Fehleranteile in Vorhersagemodellen können auch bei gut passenden Modellen groß sein, sodass die faktische Varianzaufklärung gering bleibt. In diesem Fall beschreibt das Modell präzise seinen eigenen Mangel an prädiktiver Relevanz (ebd.). Das Ziel kann nicht sein, Modelle zum „Funktionieren" zu bringen, dies zeugt nur von technischer Expertise, sondern es ist Theorien zu testen, indem ein statistisches Modell spezifiziert wird. Sollte ein Modell nicht passen, obwohl es angemessen spezifiziert wurde, fordert es seinerseits die Theorie heraus, unabhängig davon, ob es verwendet oder zurückgewiesen wurde. Der Erfolg einer Modellierung hängt davon ab ob sich die Analyse auf substanzielle theoretische Fragen bezieht und nicht davon, ob ein Modell als gut befunden wurde oder nicht. Ob ein wissenschaftlich nutzloses Modell auf die Daten passt oder nicht, ist irrelevant (Millsap, 2007).

6. DISKUSSION

Die Nutzung von Modellen mit latenten Variablen hat in den vergangenen Jahren und Jahrzehnten in den Bildungswissenschaften einen deutlichen Aufschwung erfahren (Glymour et al., 2014). Dies kann auf den gestiegenen politischen Bedarf an evidenzbasiertem empirischen Wissen, der hohen Flexibilität und erleichterten Zugänglichkeit anspruchsvoller statistischer Methoden durch computerbasierte Umsetzungen zurückgeführt werden. Die verwendeten Modelle werden in verschiedenen Disziplinen verwendet und parallel weiterentwickelt, wodurch es zu Fragmentierungen im Sprachgebrauch und einer übermäßigen Untergliederung in der Betrachtungsweise der Modelle gab. Während es aus methodologischer (Muthén, 2002; Skrondal & Rabe-Hesketh, 2004) und wissenschaftstheoretischer (vgl. Borsboom, 2008) Sicht üblich ist, latente Modelle verstärkt in einem generalisierten Rahmen zu begreifen, werden die Modelle in Anwendersicht häufig noch immer als unterschiedlich und voneinander abgegrenzt verstanden. Die hohe Zahl anwendungsorientierter Veröffentlichungen zu einzelnen Modelltypen (z. B. Bacher et al., 2010; Kline, 2011; Raykov & Marcoulides, 2006; Reinecke, 2014) kann als Indikator dafür gesehen werden.[52] Gemeinsamkeiten innerhalb der zugrunde liegenden theoretischen Annahmen, der Bewertungen der Modellgüte nach wissenschaftlichen Kriterien, der Modellidentifikation und der Modellevaluation können bei fragmentierten Darstellungen aber nicht klar herausgestellt werden. Zugleich haben forschungspraktische Veröffentlichungen, die von einem generalisierten Analyserahmen ausgehen, häufig eine erschwerte Einstiegsschwelle, da erweiterte Grundkenntnisse, beispielsweise statistisch-mathematische Kenntnisse, vorausgesetzt werden müssen, welche in einer geisteswissenschaftlichen

[51] Zum Beispiel können in IRM Items im mittleren Fähigkeitsbereich ergänzt werden, welche aber nicht zur Trennschärfe beitragen, oder in Strukturmodellen können Prädiktoren ergänzt werden, die die abhängige Variable zwar erklären, aber nicht der zentralen Fragestellung zugehörig sind.

[52] Insbesondere zum Thema Strukturgleichungsmodelle gibt es, auch und insbesondere wegen derer Flexibilität, inzwischen eine große Menge Einführungsliteratur, ohne dass die konzeptionelle Verwandtschaft zu anderen Modellen betont wird.

Ausbildung nur bedingt vermittelbar sind (z. B. Rost, 1996; Skrondal & Rabe-Hesketh, 2004; Steyer & Eid, 2001).

In dieser Arbeit wurde der Versuch unternommen, Einblicke in den Umgang mit den erkenntnistheoretischen und messtheoretischen Fundamenten von latenten Größen in einer generalisierten Sichtweise zu geben und deren Tragfähigkeit vor den Anforderungen an die moderne Bildungsforschung zu bewerten. Zudem wurden Gütekriterien von quantitativen Tests und Messverfahren im Hinblick auf diese Modelle diskutiert und Strategien der Modellevaluation soweit möglich modellfamilienübergreifend zusammengeführt. Dabei wurde ein Schwerpunkt darauf gelegt, formale Prinzipien zu versprachlichen, um implizite Annahmen bei der Verwendung von Modellen, Konzepten und Maßen herausstellen zu können, welche bei einem ritualhaften Umgang mit diesen nicht immer klar sind (Gigerenzer, 2004). Insbesondere wurde dabei das Gewicht der theoretischen Konstrukte in der Abgrenzung zu operationalen latenten Variablen vor dem Hintergrund der impliziten Grenzen von Evaluationsstrategien betont. Allgemeine formale Modellherleitungen und -formulierungen wurden nicht dargestellt, da dies in Lehrbüchern und Fachliteratur bereits hinreichend geschehen ist. Konzepte wurden somit nur wenn nötig und nur oberflächlich eingeführt. Dies muss als Einschränkung verstanden werden, da weitere technische Annahmen und Bedinungen somit unterschlagen werden. Ebenso musste aufgrund der hohen Zahl von Maßen und Tests zur Modellevaluation eine Beschränkung auf einige wenige Beispiele vorgenommen werden. Zwar wurde versucht, dabei auf forschungs–praktisch aktuelle Entwicklungen Rücksicht zu nehmen, es kann aber nicht ausgeschlossen werden, dass dies an einzelnen Stellen misslang und aktuelle Entwicklungen, insbesondere in der statistischen Fachliteratur, nicht aufgezeigt wurden. Ergänzend sei angemerkt, dass auch wenn hier schwerpunktartig von Modellen mit latenten Variablen berichtet wurde, ein Großteil der quantiativen Forschungsarbeiten im bildungs- und sozialwissenschaftlichen Kontext noch immer auf der Basis direkt beobachteter Größen vorgenommen wird und viele der hier vorgestellten Konzepte, insbeondere der Wissenschafts- und Erkenntnistheoretischen Annahmen, diesen Arbeiten ebenso zugrunde liegt. Modelle mit latenten Variablen sind lediglich eine Möglichkeit der Verarbeitung beobachteter Größen und auch ohne diese wurden umfassende und stark belastbare Forschungsergebnisse erarbeitet.

Eingangs wurde das Höhlengleichnis Platons herangeführt. Latente Variablenmodelle sind sicherlich kein Schritt aus der metaphorischen Höhle Platons heraus, sie sind kein Hilfsmittel, mit dem die Fesseln der Gefangenen gelöst werden können. Sie können zum einen verstanden werden als eine möglichst objektbezogene Beschreibung der Schatten an der Wand durch die Gefangenen. In dieser Lesart wären sie Teil eines explorativen Methodenrepertoires, mittels dessen unbekannte Gegenstände begreifbar gemacht werden sollen. Zum anderen können sie ebenso als Mittel verstanden werden, mit dem die Gefangenen ihre Einschränkung konditional lockern können. Denn so ein Gefangener laut Platon die Höhle wenn auch nur hypothetisch verlässt und die (theoretische) Wahrheit sowie die wahre Beschaffenheit der Objekte erblickt und in die Höhle zurückkehrt, so werden ihm die übrigen Gefangenen nicht glauben, wenn er von seinen Beobachtungen zu der Struktur und dem Wesen der Objekte berichtet, da sie in ihrer Wahrnehmung eingeschränkt sind. Wenn er aber die Struktur der Dinge möglichst detailliert beschreibt, so können die in der Höhle zurückgebliebenen diese Beschreibungen anhand der

durch sie wahrgenommenen Beschaffenheit der Schatten prüfen. Sie werden nicht in der Lage sein, die Behauptung zu verifizieren, aber sie können eine gelungene Beschreibung anhand der ihnen zur Verfügung stehenden Beobachtung der Schatten auch nicht widerlegen. Dazu muss der Gefangene, der die Höhle hypothetisch verlassen hat, allerdings auch beschreiben, in welcher Art die Gegenstände ihre Schatten werfen, also Annahmen zu der physikalischen Repräsentation der Gegenstände als Schatten an der Wand unterbreiten. So kann einbezogen werden, dass die Gegenstände bewegt werden und sich gegebenenfalls unterschiedlich weit vom Feuer entfernt befinden. Dies entspricht expliziten Annahmen und der Kommunikation dieser Annahmen zum vorliegenden Repräsentationsmechanismus, welcher für latente Variablenmodelle durch die repräsentative Messtheorie und die Testtheorie gegeben werden. In dieser Lesart wären latente Variablenmodelle konfirmatorischer Natur.[53] Die gleiche Möglichkeit besteht auch im Rahmen der „Wissenschaft der Schatten", wenn also die theoretische Wahrheit nicht erblickt wurde, sondern auf diese nur geschlussfolgert werden soll. Dies befindet sich auch im Rahmen der Annahme, dass in der Politeia keinesfalls die Grenzen des Erkennbaren aufgezeigt werden sollen, sondern die Grenzen des Vermittelbaren respektive des beim Gesprächspartner erreichbaren (Szlezák, 2003). Um Konsens über die Gestalt eines Gegenstands zu erreichen, muss dessen Struktur anderen Gefangenen möglichst gut beschrieben werden, damit diese die objektivierten Annahmen im Rahmen ihrer Möglichkeiten prüfen können. Übertragen auf den hier bearbeiteten Gegenstandsbereich kann die Verwendung formalisierter Sprache, unter Rückbezug auf die Messtheorie und eine Testtheorie den Rückschluss auf theoretische Konstrukte erleichtern.

Latente Variablen sind theoretisch begründete, aber erkenntnisbezogen nicht direkt verfügbare Entitäten (vgl. Borsboom, 2008). Damit die theoretischen Konstrukte beobachtbar gemacht werden können, müssen statistische latente Variablenmodelle formuliert werden. Es wurde herausgestellt, dass die Überbrückung der Definitionen des Forschungsgegenstandes als theoretisches Konstrukt und der formalen Betrachtung als latente Variable dabei als Problem der Theorie ebenso wie der Empirie begriffen werden muss. Während in der Empirie Anstrengungen unternommen werden müssen, dem Untersuchungs–gegenstand gegenüber angemessene Formen der Repräsentation, der Operationalisierung und des Forschungsdesigns zu wählen, müssen Theorien konsistent und überprüfbar sein. Häufig mangelt es aber gerade an nachvollziehbaren, prüfbaren Konzepten zu der Existenz und der Gestalt der theoretischen Konstrukte selbst (Raykov & Marcoulides, 2006, S. 49). In reflektiven konfirmatorischen Modellen wird dann bei der Formulierung von Messmodellen angenommen, dass die beobachtbaren Variablen durch die zugehörigen latenten Variablen determiniert werden. Als Basis dienen dabei die Zusammenhangsstrukturen von Beobachtungen (FA), ähnliche Reaktionsmuster (LCA/LPA) oder die Verhältnisse von Variablen zu einer angenommenen Verteilung (IRM). Statistische Modelle können genutzt werden, um die Operationalisierungen eines Forschungsgegenstandes zu entwickeln oder Datenstrukturen zu explorieren, wenn noch keine hinreichenden Kenntnisse über eine mögliche Strukturierung oder Repräsentation

[53] In Platons Höhlengleichnis wird angenommen, dass wenn ein Sehender zurückkehrt, er die Gebundenen nicht überzeugen kann, sondern von diesen getötet werden würde, wenn sie könnten – es soll aber unterstellt werden, dass dies den Anwendern konfirmatorischer latenter Modelle auch im berechtigten wissenschaftlichen Diskurs nicht widerfährt.

vorliegen[54]. Aber in der konfirmatorischen Prüfung von hypothetischen Annahmen anhand des Abgleichs eines Modells mit vorliegenden Daten unter der Berücksichtigung von Messfehlern respektive mangelnder Präzision liegt der maßgebliche Nutzen von latenten Variablenmodellen. Diese Befunde müssen modern und interdisziplinär angemessen geprüft werden.

„An angemessenen wissenschaftlichen Standards, die auch disziplinübergreifend anzuwenden sind, führt kein Weg vorbei“ (Schwippert, 2016, S. 36). Es wurde festgehalten, dass die Güte von Tests über die

- die Objektivität durch die Identifizierbarkeit und Passung des Modells sowie eine hinreichende Teststärke hergestellt wird,
- die Reliabilität durch die systematische Inspektion der Modellparameter sowie verschiedener aggregierter fehlerzentrierter Maße sichergestellt wird und
- die Validität nicht datengetrieben festgestellt werden kann. Evidenzen für vorliegende Validität können aber über die Konstruktvalidität, also die Passung und Herausstellung der Angemessenheit von Teilmodellen und gegebenenfalls dem Gesamtmodell, erbracht werden, in Ermangelung anderer Vergleichsgegenstände insbesondere gegenüber alternativen Modellen.

Die Evaluation und inhaltliche Bewertung eines Modells muss demnach auf der Basis verschiedener Maßstäbe erfolgen. Erfolgreiche internale Modellprüfungen und Abgleiche gegenüber einem Nullmodell oder einem saturierten Modell können als Evidenz dafür verstanden werden, dass eine ausreichende Objektivität gegeben ist, und die Reliabilität hinreichend sowie internale Evidenzen für vorhandene Konstruktvalidität vorliegen. Darauf aufbauend können externale Vergleiche gegenüber alternativ

- dimensionierten (z. B. eine Prüfung der Modellanpassungen einer eindimensionalen Struktur gegenüber einer mehrdimensionalen Struktur in CFA oder IRM) oder
- strukturierten (z. B. die Prüfung der Modellanpassungen bei Umkehrung der Wirkrichtung einer latenten Regression in Strukturmodellen) oder
- metrisierten (z. B. die Prüfung der Modellanpassungen bei verschiedenen Klassenzahlen in LCA/LPA)

Modellen erfolgen. Diese Vergleiche erlauben einen Schluss auf externe Evidenz für vorliegende Konstruktvalidität. Die Einschätzung des notwendigen Umfangs der Prüfungen der Reliabilität und Validität wird dabei durch gängige wissenschaftliche Standards, durch die notwendige Tragfähigkeit der Ergebnisse und das Forschungsdesign gegeben. Beispielsweise ist es nachvollziehbar, dass akzeptable Fehleranteile in der Medizin geringer sind als in den Bildungswissenschaften, ebenso ist es aber technisch leichter, Konfundierungen designbasiert auszuschließen. Generell muss festgehalten werden, dass die Maße der Modellevaluation keinesfalls erschöpfend beforscht worden sind und absolute kriteriale Maße, insbesondere Forschungsgegenständen in den Sozialwissenschaften gegenüber, kaum angemessen sind. Es existieren keine „goldenen Schwellenmaße“ und Werte, die in der Fachliteratur als

[54] Häufig müssen sie im Prozess einer Modellentwicklung derart genutzt werden.

wünschenswert bezeichnet werden, sind maßgeblich als Orientierungshilfen zu verstehen (Hu & Bentler, 1999). Deutliche Lücken für belastbare externale Gütemaße existieren insbesondere für latente Klassifikationen und generell auch für Clusteranalysen mit Fuzzy-Logik (Wurpts & Geiser, 2014). Ebenso existiert keine einheitliche Meinung über die Bewertung der Funktionalität (und Nützlichkeit) von globalen Gütemaßen für IRM (Rost, 1996).

Die Verwendung von Maßen der Modellgüte können Forschenden also keinesfalls Entscheidungen abnehmen. Reflektiertes Denken steht hier mechanischen Ritualen gegenüber, wie zum Beispiel dem Testen auf Signifikanzen ohne eine Berücksichtigung der Teststärke oder des Effektgewichts gegenüber (Sedlmeier, 2009), und eine verstärkte Betonung von Effektstärken, wie sie in vielen Journals zu finden ist, war bereits von Cohen (1977) gefordert worden.[55] Auch das mechanische „fitten" von Modellen lässt sich im Sinne des kritisierten Signifikanzrituals begreifen, und von Gigerenzer (2004, S. 18) wurde festgestellt, dass derartige Rituale in den Sozialwissenschaften noch weit verbreitet sind. Auch von Merkens (2003, S. 50) wird bemängelt, dass sich die Annahme eines kritisch, nach den Regeln der Kunst geprüften Modells häufig in der Kontrolle des methodischen Vorgehens erschöpfe. Die Annahme der Vergegenständlichung der Realität durch geprüfte Modelle darf sich also nicht in der mathematischen Prüfung erschöpfen, sondern bedarf weiterhin des wissenschaftlichen Diskurses (Merkens, 2003, S. 51). Gigerenzer (2004, S. 18) stellt hierzu das statistische Denken in Abgrenzung zum statistischen Ritual heraus. Statistisches Denken sei selbstreflektiv; es beinhaltet die Abwägung, welche Methode oder welches Modell für eine Situation die beste ist und unter welchen Annahmen das gilt. Letztlich obliegen Annahmen über die Angemessenheit und Funktionalität von latenten Variablenmodellen und den daraus abgeleiteten handlungsleitenden Forderungen subjektiv den Forschenden (Kline, 2011, S. 191). Diese Urteile sind qualitativer Natur und basieren auf domänenspezifischem Wissen sowie auch persönlichen Werten und gesellschaftlichen Anliegen (vgl. Popper, 1994) vor dem Hintergrund quantifizierter Unsicherheit.

Damit Strukturen oder Repräsentationen konfirmatorisch geprüft werden können,

- muss eine überprüfbare theoretische Annahme formuliert werden,
- müssen sprachliche Unschärfen in der Kommunikation über das Objekt abgebaut werden,
- müssen möglichst viele verschiedene Perspektiven einbezogen werden, und
- es muss eine Offenheit bezüglich den Einschränkungen der Betrachtungsform, also den mechanischen Annahmen zur Repräsentationsform, hergestellt werden, um die Angemessenheit der Repräsentation zu unterstreichen.

Mit der Verwendung latenter Variablenmodelle können dann Möglichkeiten geschaffen werden, Hypothesen Falsifizierungsversuchen zu unterziehen und Annahmen innerhalb von Stichproben zu prüfen. Diese Annahmen können bei angemessener Stichprobenziehung und Quantifikation der Modellanpassung, also der Benennung der Unsicherheit der Wertzuschreibung, generalisiert

[55] Wolf (2008) hatte zur besseren Bewertung der Effektmaße beispielsweise 1986 die These aufgestellt, dass Effekte von $cohens\ d < .25$ keine bildungswissenschaftliche Bedeutung hätten.

werden. Quantitative sozialwissenschaftliche Untersuchungen sind, wie alle wissenschaftlichen Tätigkeiten aus verschiedenen Gründen mit Unsicherheit belastet. Popper (1994, S. 223) schreibt im Sinne des kritischen Realismus: „Unsere Wissenschaft ist kein System von gesicherten Sätzen, auch kein System, das in stetem Fortschritt einem Zustand der Endgültigkeit zustrebt. Unsere Wissenschaft ist kein Wissen [*epistēmē*]: weder Wahrheit noch Wahrscheinlichkeit kann sie erreichen.“ Viele Faktoren wie die Wahl der Indikatoren, die Modellspezifikation und vor allem die Ziehung von Stichproben erzeugen Unsicherheit. Der Einsatz von Modellen mit latenten Variablen in der empirischen Bildungsforschung ist eine Möglichkeit des Umgangs mit dieser Unsicherheit, da diese über die vorgestellten Maße der Modellbeurteilung konditional quantifiziert wird. Es wurde aufgezeigt, dass latente Variablenmodelle mathematisch, aber auch erkenntnistheoretisch nicht wohl formuliert, also nicht ohne Messungenauigkeiten, auskommen können, womit empirische Erkenntnisse per Definition probabilistischer und nicht deterministischer Natur sind (Manski, 2003). Jedoch bietet der Einsatz dieser Modelle auch eine Abschätzung der Fehleranfälligkeit, einen expliziten Einbezug des Messfehlers oder der (Un)Präzision eines Tests, sodass eine Einordnung der Stärke der vorhandenen Evidenz für das Vorhandensein stochastischer Kausalbezüge vorgenommen werden kann.[56]

Dieses Problem indirekter Beweisführung und der vorläufige Charakter aller bildungswissenschaftlichen Erkenntnisse erschweren allerdings einen Transfer der Ergebnisse in die Öffentlichkeit (Bromme & Prenzel, 2014). Der Einfluss der Bildungsforschung ist in hohem Maße von der Professionalität in der Forschung und der Präsentation der Ergebnisse abhängig (Fend, 2009, S. 25). Es wurde ausgeführt, dass die Bildungswissenschaften in Theorien und Befunden handlungsleitendes Wissen erzeugen sollen (Bromme et al., 2014; Schwippert, 2016, S. 36), aber das Kerngeschäft von Wissenschaftlern ist die Anwendung von Qualitätskriterien für die Ableitung von Wahrheitsaussagen, welche als die Dimensionen der (1) Relevanz, der (2) methodischen und ethischen Strenge sowie der (3) angemessenen Präsentation (Döring & Bortz, 2016, S. 90) vorgestellt wurden (vgl. Kap. 2). Analog wird bei Markus und Borsboom (2013, S. 282) methodenspezifischer formuliert, dass Forscher in einem Prozess der Handlungsableitung die Aufgabe hätten (a) klar zu machen, was die verwendeten technischen Variablen repräsentieren sollen, (b) welcher Gestalt die Evidenz für oder wider die Annahme der Angemessenheit der Repräsentation ist und (c) warum die empirische Evidenz für Validität angenommen werden sollte. Forscher sind Experten in (1) und (2) sowie (a) und (b), sie haben eine geringere Expertise in (3) und (c), denn dies sind Domänen, welche unter Umständen philosophisches und/oder methodisches Spezialwissen erfordern und Anknüpfungsfähigkeit in der Öffentlichkeit, also in die Praxis, beinhaltet (Markus & Borsboom, 2013, S. 282). Gleichzeitig ist, wenn steuerungs- oder handlungsleitendes Wissen erzeugt werden soll, der zwingende letzte Schritt die Transferleistung. Inwieweit diese final von der Administration oder praktisch Tätigen oder von wissenschaftlicher Seite aus vorgenommen werden sollte, wird kritisch diskutiert (Bellmann & Müller, 2011; z.B. Bromme et al., 2014) und kann an dieser Stelle nicht beantwortet werden.

[56] Gleichzeitig ist es eben wegen dieser Unsicherheit sinnvoll, möglichst große Effekte zu suchen, da diese zumeist unabhängig des Designs beobachtet werden können Rubin (1974, S. 700).

Kritisch wird zur quantitativen empirischen Bildungsforschung und deren objektivem Wissenschaftsverständnis häufig formuliert, dass die Wirkung von pädagogischem Handeln durch die Unmöglichkeit der Trennung von Erkenntnissubjekt und Erkenntnisobjekt an die Messung gebunden sei, womit die Essenz einer objektiver Verfahrensweisen, nämlich der Unabhängigkeit des Handelns von der Messung, verletzt sei (vgl. Bellmann & Müller, 2011). Bereits dadurch, dass das Wissenschaftssystem an bestehenden Verhältnissen partizipiert, würde der Blickwinkel verzerrt (Herzog, 2010). Ebenso muss festgehalten werden, dass die quantitiative Methodologie in ihrer Gänze darauf aufbaut, dass die Forschenden eine hinreichende Vorstellung von ihrem jeweiligen Forschungsbereich haben, denn insoweit über einen Sachverhalt ein relevanter Wissensbestand fehlt oder eine Repräsentation nicht verfügbar ist, so kann dieser in den Hypothesen nicht auftauchen (z. B. Kelle, 2009, S. 34). Hier steht der forschungspragmatische Gedanke der notwendigen Reduktion (Friedrichs, 1982, S. 80) direkt dem eng verknüpften Wirkungsgefüge in bildungswissenschaftlichen Kontexten gegenüber.

Auch um dies aufzulösen wird seit Längerem eine Methodenunspezifik (Burzan, 2016; Friedrichs, 1982, S. 81; Kelle, 2009) gefordert, und es kann deutlich beobachtet werden, dass seit Mitte der 1980er Jahre eine deutlicher Aufschwung in der Erweiterung der methodologischen Ansätze stattfand, indem Forschungskonzepte üblicher wurden, die eine Integration von qualitativen und quantitativen Daten und Methoden vornahmen (Burzan, 2016; Kelle, 2009; Kuckartz, 2014). Probleme für integrative Ansätze wurden zum einen in der (Un)Haltbarkeit der Übereinstimmung der Weltanschauungen, im Rahmen des forschungslogischen Ablaufs, also insbesondere bei der Formulierung wissenschaftlicher Begriffe, gesehen. Zum anderen findet in einem pragmatischen Ansatz eine (Über)Betonung der technischen Aspekte integrativer Designs statt (Tashakkori & Teddlie, 2010). Einigkeit besteht jedoch über die umfassenden pragmatischen Potenziale reflektierter Methodenverknüpfung (z.B. Burzan, 2016). Ein beispielhafter Ansatz eines Designs der Methodenverknüpfung, bei dem also in einem Mixed-Method Ansatz erst eine quantiative Untersuchung erfolgt und dann deren Ergebnisse vertiefend in einem qualitiativen Ansatz bearbeitet werden (Kuckartz, 2014), findet sich bei Busch et al. (2015). Die Falsifikation von Hypothesen in einem statistischen Modell und die Beobachtung unerwarteter Einflüsse, konnten in dieser Forschungsarbeit über ergänzendes Interviewmaterial erklärt werden. Die Bedeutung der Methodenintegration kann beispielhaft auch über die Verankerung qualitativer und quantitativer Methoden in den Empfehlungen der Deutschen Gesellschaft für Erziehungswissenschaft (DGfE) für das Kerncurriculum des Studiums der Erziehungs–wissenschaft gesehen werden (DGfE, 2010). Es verbleiben bei der Integration verschiedener Forschungsansätze, egal ob diese mehrere qualitative, mehrere quantitative oder gemischte Methoden beinhalten, unvermeidlich zwischenmethodische Unschärfen. Daher muss es insbesondere gelingen, eine theoretische Verknüpfung der unterschiedlichen logischen und semantischen Konstruktionsweisen ihrer Aussagen zu finden (Friedrichs, 1982, S. 81).

Die Gleichsetzung von Methodenrepertoires mit ideologischen Grundvorstellungen ist und war häufig die Grundlage der Kritik am quantitativen Paradigma (Brügelmann, 2015). Dies kann als eine Renaissance des Werturteilsstreits respektive des Positivismusstreits verstanden werden (z.B. Kelle, 2009, S. 41ff). Im Sinne Max Webers ist die Objektivität von sozialwissenschaftlicher Erkenntnis zwar auf die Wertideen ausgerichtet, die ihr Erkenntniswert verleihen und die

ebenso wie die Bedeutung der Erkenntnis aus der Wertidee verstanden wird, aber trotzdem kann aus der empirischen Erkenntnis heraus niemals ein Nachweis ihrer Geltung erfolgen (Weber, Winckelmann & Weber, 1988, S. 213). In einem ähnlichen Sinne wie im nachfolgenden kritischen Rationalismus werden auch bei Weber Wertfragen und Fragen der Gültigkeit getrennt. Objektivität bedeutet im Sinne in der Tradition des kritischen Rationalismus Nachvollziehbarkeit und kritische Auseinandersetzung mit Theorien und Erkenntnissen gleichermaßen. Moralische oder politische Positionen entscheiden nicht über die Annahme oder Ablehnung von Hypothesen. Dies setzt voraus, dass das Wissenschaftssystem Plural genug ist, um nicht einer allseits geteilten impliziten Ideologie zu folgen (Döring & Bortz, 2016, S. 60). Unterschiedliche Auffassungen darüber, ob Objektivität überhaupt erreichbar und ethisch angemessen ist, bleibt bis heute und über die Bestrebungen von Forschungssynthesen hinweg, ein in den Diskussionen zwischen einem quantitativen und qualitativen Paradigma umstrittener Punkt, trotz des Vorliegens klarer Gemeinsamkeiten in epistemologischer und axiomatischer Hinsicht (Saldern, 1992). Das Ziel jeder Methode muss eine datengestützte Fehlerreduktion, also eine „Wahrheitsfindung“ sein, wobei die Daten als Gegner der eigenen Annahmen begriffen werden (Fend, 2009, S. 27). Für die weitere Etablierung interdisziplinärer Bildungsforschung wäre es fatal, wenn methodische oder methodologische Ansätze unreflektiert kritisiert werden, um dem Blickfeld der eigenen tradierten Forschungsdisziplin zusätzliches Gewicht zu verleihen (Schwippert, 2016, S. 31). Zugleich müssen aber methodische und metatheoretische Schwächen benannt, reflektiert und aufgehoben werden. Damit Evidenz erzeugt werden kann, ist die Weiterentwicklung empirischer Methoden erforderlich, damit in einer kritischen Zusammenschau entscheidungsrelevante und starke Evidenz erzeugt werden kann (Bromme et al., 2014, S. 14). Die interdisziplinäre Kooperation hat dabei zur Weiterentwicklung der gemeinsam verwendeten Methoden geführt (ebd.). Dieser Prozess verlangt eine nachvollziehbare und vollständige Dokumentation der Methoden und Befunde. Neuere Entwicklungen bei dem Aufbau von Forschungsdatenzentren und deren kooperativer Verknüpfung, zum Beispiel durch den *Verbund Forschungsdaten Bildung*, welcher im Auftrag des BMBF eingerichtet wurde, zeigen die Wahrnehmung des Problems und dessen Stellenwerts auf.

Die Erfolge der Forschungsansätze von Studien, wie sie nach der Teilnahme an der internationalen *Reading Literacy* Studie (vgl. Schwippert & Goy, 2008) häufiger vorkamen, wegen ihrer methodischen und methodologischen Wurzeln generell zu kritisieren, ist für die Zukunft der Bildungsforschung fahrlässig (Schwippert, 2016, S. 34). Es ist nicht in Abrede zu stellen, wie sehr die Forschung und die Theorieentwickung in der Schul- und Unterrichtsforschung, wie auch in der Schulpädagogik in Fachdidaktiken und der Lehr-Lern Psychologie durch quantitative Ansätze angeregt wurden (Schwippert, 2016, S. 34).

Damit an diese Erfolge auch zukünftig angeknüpft werden kann, ist es hilfreich und notwendig, nicht nur die Integration neuer Theorien aus der Soziologie, Psychologie oder Ökonomie anzustreben, sondern auch die Integration und Anwendung verschiedener empirischer Methoden (Schwippert, 2016, S. 34). Dazu ist es notwendig, die Anwendung von Methoden in anderen auch fachfremden Disziplinen zu betrachten und nützliche Methoden und Betrachtungsweisen in die eigenen Disziplin zu übertragen (Skrondal & Rabe-Hesketh, 2004, S. ix). Dies erfordert eine interdisziplinäre Übersetzungsleistung ebenso wie eine technische

Umsetzung, welche für Bildungsforscher anknüpfungsfähig ist, diese Verknüpfung wird nicht immer durch Diziplinen geleistet werden, welche die Methoden entwickeln. Methodische Lösungen für forschungspraktische Probleme in den Bildungswissenschaften werden nicht immer in die Disziplinen in diesem Feld getragen, sondern müssen aktiv eingeholt werden.

LITERATURVERZEICHNIS

Adorno, T. W., Albert, H., Dahrendorf, R., Habermas, J., Pilot, H. & Popper, K. R. (1993). *Der Positivismusstreit in der deutschen Soziologie*. München: Dt. Taschenbuch-Verl.

American Educational Research Association; American Psychological Association; National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, D.C.: Autor.

American Educational Research Association; American Psychological Association; National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, D.C.: Autor.

American Psychological Association, American Educational Research Association & National Council on Measurement in Education. (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin, 51* (2), 1–38. https://doi.org/10.1037/h0053479

American Psychological Association; American Educational Research Association; National Council on Measurement in Education. (1966). *Standards for Educational and Psychological Tests and Manuals*. Washington, D.C.: APA.

American Psychological Association; American Educational Research Association; National Council on Measurement in Education. (1974). *Standards for Educational and Psychological Tests and Manuals*. Washington, D.C.: Autor.

American Psychological Association; American Educational Research Association; National Council on Measurement in Education; Joint Committee on Standards for Educational and Psychological Testing (U.S.). (1985). *Standards for educational and psychological testing*. Washington, D.C.: Autor.

Anderson, C. A. (1961). Methodology of Comparative Education. *International Review of Education* (7), 1–23.

Asparouhov, T. & Muthén, B. (2009). Exploratory Structural Equation Modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 16* (3), 397–438. https://doi.org/10.1080/10705510903008204

Asparouhov, T. & Muthén, B. (2010). *Weighted Least Squares Estimation with Missing Data.* Zugriff am 20.12.2016. Verfügbar unter http://ww.statmodel2.com/download/GstrucMissingRevision.pdf

Atteslander, P. (2010). *Methoden der empirischen Sozialforschung* (13. Aufl.). Berlin: Erich Schmidt Verlag.

Autorengruppe Bildungsberichterstattung. (2008). *Bildung in Deutschland 2008. Ein indikatorengestützter Bericht mit einer Analyse zu Übergängen im Anschluss an den Sekundarbereich I*. Bielefeld: Bertelsmann.

Bacher, J., Pöge, A. & Wenzig, K. (2010). *Clusteranalyse. Anwendungsorientierte Einführung in Klassifikationsverfahren* (3. Aufl.). München: Oldenbourg.

Bacher, J. & Vermunt, J. K. (2010). Analyse latenter Klassen. In C. Wolf & H. Best (Hrsg.), *Handbuch der sozialwissenschaftlichen Datenanalyse* (S. 553–574). Wiesbaden: VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-92038-2_22

Backhaus, K., Erichson, B., Plinke, W. & Weiber, R. (2016). *Multivariate Analysemethoden*. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-662-46076-4

Backhaus, K., Erichson, B. & Weiber, R. (2013). *Fortgeschrittene multivariate Analysemethoden* (2. Aufl.). Berlin u.a.: Springer Gabler.

Baghaei, P. & Tabatabaee Yazdi, M. (2016). The Logic of Latent Variable Analysis as Validity Evidence in Psychological Measurement. *The Open Psychology Journal, 9* (1), 168–175. https://doi.org/10.2174/1874350101609010168

Barrett, P. (2007). Structural equation modelling. Adjudging model fit. *Personality and Individual Differences, 42* (5), 815–824. https://doi.org/10.1016/j.paid.2006.09.018

Bartholomew, D. J. (2004). *Measuring intelligence. Facts and fallacies*. Cambridge, UK: Cambridge University Press.

Baumert, J., Bos, W. & Lehmann, R. (2000). *TIMSS/III Dritte Internationale Mathematik- und Naturwissenschaftsstudie Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn. Band 1 Mathematische und naturwissenschaftliche Grundbildung am Ende der Pflichtschulzeit*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Bellmann, J. & Müller, T. (Hrsg.). (2011). *Wissen was wirkt. Kritik evidenzbasierter Pädagogik*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Bentler, P. M. (1980). Multivariate Analysis with Latent Variables. Causal Modeling. *Annual review of psychology, 31* (1), 419–456. https://doi.org/10.1146/annurev.ps.31.020180.002223

Bentler, P. M. & Chou, C.-P. (1987). Practical Issues in Structural Modeling. *Sociological Methods & Research, 16* (1), 78–117. https://doi.org/10.1177/0049124187016001004

Bertelsmann Stiftung, Institut für Schulentwicklungsforschung Dortmund & Institut für Erziehungswissenschaft Jena (Hrsg.). (2017). *Chancenspiegel - eine Zwischenbilanz. Zur Chancengerechtigkeit und Leistungsfähigkeit der deutschen Schulsysteme seit 2002* (1. Auflage). Gütersloh: Bertelsmann Stiftung.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examiner's ability. In F. M. Lord & M. R. Novick (Hrsg.), *Statistical theories of mental test scores* (Addison-Wesley series in behavioral science, S. 17–20). Charlotte, N.C.: Information Age Pub. Inc.

Blossfeld, H.-P., Doll, J. & Schneider, T. (2008). Bildungsprozesse im Lebenslauf. Grundzüge der zukünftigen Bildungspanelstudie für die Bundesrepublik Deutschland. *Recht der Jugend und des Bildungswesens, 56* (3), 321–328. Verfügbar unter https://www.bwv-verlag.de/digibib/bwv/apply/viewpdf/opus/229116/contribution/2643/

Blossfeld, H.-P., Maurice, J. von & Schneider, T. (2011). The National Educational Panel Study. Need, main features, and research potential. *Zeitschrift für Erziehungswissenschaft, 14* (S2), 5–17. https://doi.org/10.1007/s11618-011-0178-3

BMBF (Hrsg.). (2008). *Rahmenprogramm zur Förderung der empirischen Bildungsforschung. Framework programme for the promotion of empirical educational research* (Bildung - Ideen zünden!, Bd. 22). Bonn: Autor.

Bollen, K. A. (1989). *Structural equations with latent variables* (Wiley series in probability and mathematical statistics. Applied probability and statistics). New York, NY: Wiley. https://doi.org/10.1002/9781118619179

Bollen, K. A. (2002). Latent variables in psychology and the social sciences. *Annual review of psychology, 53,* 605–634. https://doi.org/10.1146/annurev.psych.53.100901.135239

Bollen, K. A. & Long, J. S. (Hrsg.). (1993). *Testing structural equation models* (Sage focus editions, Bd. 154). Newbury Park: Sage.

Bollen, K. A. & Pearl, J. (2014). Eight Myths About Causality and Structural Equation Models. In S. L. Morgan (Hrsg.), *Handbook of causal analysis for social research* (Handbooks of Sociology and Social Research, S. 301–328). Heidelberg: Springer. https://doi.org/10.1007/978-94-007-6094-3_15

Bond, T. G. & Fox, C. M. (2001). *Applying the Rasch model. Fundamental measurement in the human sciences*. Mawah, N.J.: Lawrence Erlbaum Associates Publishers.

Boomsma, A. & Hoogland, J. J. (2001). The Robustness of LISREL Modeling Revisited. In S. duToit, R. Cudeck & D. Sorbom (Hrsg.), *Structural equation modeling. Present and future ; a festschrift in honor of Karl Jöreskog* (S. 139–168). Lincolnwood, Ill.: Scientific Software International.

Borovcnik, M. (2014). Forschungsprozess und probabilistische Modellbildung – Stochastische Denkweisen. In J. Maaß & H.-S. Siller (Hrsg.), *Neue Materialien für einen realitätsbezogenen Mathematikunterricht 2* (S. 11–30). Wiesbaden: VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-658-05003-0_2

Borsboom, D. (2008). Latent Variable Theory. *Measurement: Interdisciplinary Research & Perspective, 6* (1-2), 25–53. https://doi.org/10.1080/15366360802035497

Borsboom, D., Mellenbergh, G. J. & van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review, 110* (2), 203–219. https://doi.org/10.1037/0033-295X.110.2.203

Borsboom, D., Mellenbergh, G. J. & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111* (4), 1061–1071. https://doi.org/10.1037/0033-295X.111.4.1061

Bos, W. & Postlethwaite, T. N. (2000). Möglichkeiten, Grenzen und Perspektiven internationaler Schulleistungsforschung. In Rolff, H.-G., Bos, W., K. Klemm, H. Pfeiffer & R. Schulz-Zander (Hrsg.), *Jahrbuch der Schulentwicklung. Daten, Beispiele und Perspektiven* (S. 365–386). Weinheim: Juventa Verlag.

Box, G. E. P. & Draper, N. R. (1987). *Empirical model-building and response surfaces* (Wiley series in probability and mathematical statistics. Applied probability and statistics, 7. Aufl.). New York, NY: Wiley.

Brezinka, W. (1975). *Von der Pädagogik zur Erziehungswissenschaft. Eine Einführung in die Metatheorie der Erziehung* (Beltz-Studienbuch, Bd. 22, 3. Aufl.). Weinheim, Basel: Beltz.

Bromme, R. & Prenzel, M. (2014). Zu diesem Sonderheft. *Zeitschrift für Erziehungswissenschaft, 17* (S4), 1–2. https://doi.org/10.1007/s11618-014-0544-z

Bromme, R., Prenzel, M. & Jäger, M. (2014). Empirische Bildungsforschung und evidenzbasierte Bildungspolitik. *Zeitschrift für Erziehungswissenschaft, 17* (S4), 3–54. https://doi.org/10.1007/s11618-014-0514-5

Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology, 37* (1), 62–83. https://doi.org/10.1111/j.2044-8317.1984.tb00789.x

Browne, M. W. & Cudeck, R. (1992). Alternative Ways of Assessing Model Fit. *Sociological Methods & Research, 21* (2), 230–258. https://doi.org/10.1177/0049124192021002005

Brügelmann, H. (2015). *Vermessene Schulen - standardisierte Schüler. Zu Risiken und Nebenwirkungen von PISA, Hattie, VerA & Co*. Weinheim: Beltz.

Buchhaas-Birkholz, D. (2009). Die "empirische Wende" in der Bildungspolitik und in der Bildungsforschung : zum Paradigmenwechsel des BMBF im Bereich der Forschungsförderung. *Erziehungswissenschaft, 20* (39), 27–33.

Bühner, M. (2011). *Einführung in die Test- und Fragebogenkonstruktion* (3. Aufl.). München: Pearson.

Burzan, N. (2016). *Methodenplurale Forschung. Chancen und Probleme von Mixed Methods*. Weinheim: Beltz Juventa.

Busch, V., Schurig, M., Bunte, N. & Beutler-Prahm, B. (2015). „Mir gefällt ja mehr diese Rockmusik.“. Zur Struktur musikalischer Präferenzurteile im Grundschulalter. In W. Auhagen, C. Bullerjahn & R. v. Georgi (Hrsg.), *Offenohrigkeit. Ein Postulat im Fokus* (Jahrbuch der Deutschen Gesellschaft für Musikpsychologie, Bd. 24, S. 133–168). Göttingen: Hogrefe.

Byrne, B. M., Shavelson, R. J. & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures. The issue of partial measurement invariance. *Psychological Bulletin, 105* (3), 456–466. https://doi.org/10.1037/0033-2909.105.3.456

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin, 54* (4), 297–312. https://doi.org/10.1037/h0040950

Campbell, D. T. & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56* (2), 81–105. https://doi.org/10.1037/h0046016

Cheung, G. W. & Rensvold, R. B. (2002). Evaluating Goodness-of-Fit Indexes for Testing Measurement Invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 9* (2), 233–255. https://doi.org/10.1207/S15328007SEM0902_5

Clauss, F. J. (1981). *Wissenschaftslogik und Sozialökonomie*. Berlin: Duncker & Humblot.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (Überarb. Aufl.). New York, NY: Academic Press.

Cook, T. D. & Campbell, D. T. (1979). *Quasi-experimentation. Design & analysis issues for field settings*. Boston, Mass.: Houghton Mifflin.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78* (1), 98–104. https://doi.org/10.1037/0021-9010.78.1.98

Cortina, K. S., Baumert, J., Leschinsky, A., Mayer, K. U. & Trommer, L. (2008). *Das Bildungswesen in der Bundesrepublik Deutschland. Strukturen und Entwicklungen im Überblick ; [der neue Bericht des Max-Planck-Instituts für Bildungsforschung]*. Reinbek bei Hamburg: Rowohlt Taschenbuch Verlag.

Cronbach, L. J. (1982). *Designing evaluations of educational and social programs* (Jossey-Bass series in social and behavioral science and in higher education). San Francisco, Ca.: Jossey-Bass.

Dammer, K.-H. (2015). *Vermessene Bildungsforschung. Wissenschaftsgeschichtliche Hintergründe zu einem neoliberalen Herrschaftsinstrument* (Pädagogik und Politik, Bd. 8, Korr. Nachdr). Baltmannsweiler: Schneider Hohengehren.

Dempster, A. P. (1998). Logicist statistics. I. Models and modeling. *Statistical Science, 13* (3), 248–276. https://doi.org/10.1214/ss/1028905887

Deutschen Institut für Medizinische Dokumentation und Information (Hrsg.). (2014). *Internationale statistische Klassifikation der Krankheiten und verwandter Gesundheitsprobleme (ICD-10). German Modification* (ICD-10-GM, 10. Aufl.). Köln: Dt. Ärzte-Verl.

Deutscher Bildungsrat. (1974). *Empfehlungen der Bildungskommission*. Stuttgart: Klett.

Deutsches Zentrum für Luft- und Raumfahrt e. V. - Projektträger im DLR (Hrsg.). (2013). *Eine Sammlung BMBF-geförderter Projekte. Empirische Bildungsforschung*. Mühlheim a. d. R.

DGfE. (2010). *Kerncurriculum Erziehungswissenschaft. Empfehlungen der Deutschen Gesellschaft für Erziehungswissenschaft (DGfE)* (Erziehungswissenschaft, Bd. 21, 2. Aufl.). Opladen: Budrich.

Diaz-Bone, R. & Weischer, C. (2014). *Methoden-Lexikon für die Sozialwissenschaften*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Diekmann, A. (2012). *Empirische Sozialforschung. Grundlagen, Methoden, Anwendungen* (rororo rowohlts enzyklopädie, Bd. 55678, 6. Aufl.). Reinbek bei Hamburg: Rowohlt Taschenbuch Verlag.

Döring, N. & Bortz, J. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften*. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-41089-5

Dunn, T. J., Baguley, T. & Brunsden, V. (2014). From alpha to omega: a practical solution to the pervasive problem of internal consistency estimation. *British journal of psychology, 105* (3), 399–412. https://doi.org/10.1111/bjop.12046

Dziak, J. J., Coffman, D. L., Lanza, S. T. & Li, R. (2015). Sensitivity and specificity of information criteria. https://doi.org/10.7287/PEERJ.PREPRINTS.1103V2

Edelmann, D., Schmidt, J. & Tippelt, R. (2012). *Einführung in die Bildungsforschung*. Stuttgart: Kohlhammer.

Embretson, S. E. & Reise, S. P. (2000). *Item response theory for psychologists* (Multivariate applications books series, Bd. 4). Mahwah, N.J.: Erlbaum.

Everitt, B. S. (1984). *An Introduction to Latent Variable Models*. Dordrecht: Springer. https://doi.org/10.1007/978-94-009-5564-6

Fahrmeir, L., Kneib, T. & Lang, S. (2007). *Regression. Modelle, Methoden und Anwendungen* (Statistik und ihre Anwendungen). Berlin: Springer. https://doi.org/10.1007/978-3-540-33933-5

Fahrmeir, L., Künstler, R., Pigeot, I. & Tutz, G. (2003). *Statistik. Der Weg zur Datenanalyse* (Springer-Lehrbuch, 4. Aufl.). Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-662-22657-5

Fend, H. (2009). Bildungsforschung von 1965 bis 2008. In B. Wischer & K.-J. Tillmann (Hrsg.), *Erziehungswissenschaft auf dem Prüfstand. Schulbezogene Forschung und Theoriebildung von 1970 bis heute* (S. 15–33). Weinheim: Juventa Verlag.

Finch, W. H. & Bronk, K. C. (2011). Conducting Confirmatory Latent Class Analysis Using M plus. *Structural Equation Modeling: A Multidisciplinary Journal, 18* (1), 132–151. https://doi.org/10.1080/10705511.2011.532732

Finetti, B. d. (1981). *Wahrscheinlichkeitstheorie. Einführende Synthese mit kritischem Anhang* (Scientia Nova, [Nachdr.]. München: De Gruyter Oldenbourg.

Fischer, G. H. (1974). *Einführung in die Theorie psychologischer Tests. Grundlagen und Anwendungen* (Vollst. Neufassung). Bern: Huber.

Friedrichs, J. (1982). *Methoden empirischer Sozialforschung* (10. Aufl.). Wiesbaden: VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-90173-2

Gadenne, V. (1994). Theorien. In T. Herrmann & W. H. Tack (Hrsg.), *Methodologische Grundlagen der Psychologie* (Enzyklopädie der Psychologie. Themenbereich B, Methodologie und Methoden. Serie 1, Forschungsmethoden der Psychologie, Band 1, S. 295–332). Göttingen: Hogrefe.

Gadenne, V. (1984). *Theorie und Erfahrung in der psychologischen Forschung* (Die Einheit der Gesellschaftswissenschaften, Bd. 36). Tübingen: J.C.B. Mohr.

Gagne, P. & Hancock, G. R. (2006). Measurement Model Quality, Sample Size, and Solution Propriety in Confirmatory Factor Models. *Multivariate behavioral research, 41* (1), 65–83. https://doi.org/10.1207/s15327906mbr4101_5

Gangl, M. (2010). Nichtparametrische Schätzung kausaler Effekte mittels Matchingverfahren. In C. Wolf & H. Best (Hrsg.), *Handbuch der sozialwissenschaftlichen Datenanalyse* (S. 931–961). Wiesbaden: VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-92038-2_35

Gigerenzer, G. (2004). Die Evolution des statistischen Denkens. *Unterrichtswissenschaft, 32* (1), 4–22.

Gigerenzer, G. (2013). *Risiko. Wie man die richtigen Entscheidungen trifft* (1. Aufl.). München: Bertelsmann.

Glas, C. A. W. & Dagohoy, A. V. T. (2007). A Person Fit Test for Irt Models for polytomous Items. *Psychometrika, 72* (2), 159–180. https://doi.org/10.1007/s11336-003-1081-5

Glymour, C., Scheines, R. & Spirtes, P. (2014). *Discovering Causal Structure. Artificial Intelligence, Philosophy of Science, and Statistical Modeling*. Burlington, Mass.: Elsevier Science.

Goldhammer, F., Frank, Rölke, H., Scharaf, A. & Upsing, B. (2008) Technology Based Assessment - ein Gemeinschaftsprojekt der Arbeitseinheiten "Informationszentrum Bildung" und "Bildungsqualiät und Evaluation". In DIPF (Hrsg.), *DIPF informiert* (S. 2–6). Verfügbar unter http://www.dipf.de/pdf-dokumente/publikationen/dipf-informiert/dipf-informiert-nr.-12

Goodman, L. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika, 61* (2), 215–231. https://doi.org/10.1093/biomet/61.2.215

Goodman, L. A. (1985). *Analyzing qualitative - categorical data. Log-linear models and latent structure analysis* (Abt Books, [Nachdr.]. Lanham, Md.: Univ. Press of America.

Gräsel, C. (2011). Was ist Empirische Bildungsforschung? In H. Reinders, H. Ditton, C. Gräsel & B. Gniewosz (Hrsg.), *Empirische Bildungsforschung. Strukturen und Methoden* (S. 13–27). Wiesbaden: VS Verlag für Sozialwissenschaften.

Haladyna, T. M. (2006). Roles and Importance of Validity Studies in Test Development. In S. M. Downing & T. M. Haladyna (Hrsg.), *Handbook of test development* (S. 739–758). Mawah, N.J.: Lawrence Erlbaum Associates Publishers.

Hancock, G. R. & Mueller, R. O. (2011). The Reliability Paradox in Assessing Structural Relations Within Covariance Structure Models. *Educational and Psychological Measurement, 71* (2), 306–324. https://doi.org/10.1177/0013164410384856

Hartig, J., Frey, A. & Jude, N. (2008). Validität. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (Springer-Lehrbuch, S. 135–163). Berlin: Springer. https://doi.org/10.1007/978-3-540-71635-8_7

Hayduk, L. A. & Glaser, D. N. (2000a). Doing the Four-Step, Right-2-3, Wrong-2-3. A Brief Reply to Mulaik and Millsap; Bollen; Bentler; and Herting and Costner. *Structural Equation Modeling: A Multidisciplinary Journal, 7* (1), 111–123. https://doi.org/10.1207/S15328007SEM0701_06

Hayduk, L. A. & Glaser, D. N. (2000b). Jiving the Four-Step, Waltzing Around Factor Analysis, and Other Serious Fun. *Structural Equation Modeling: A Multidisciplinary Journal, 7* (1), 1–35. https://doi.org/10.1207/S15328007SEM0701_01

Hayduk, L. A. & Littvay, L. (2012). Should researchers use single indicators, best indicators, or multiple indicators in structural equation models? *British Medical Research Methodology, 12* (159), 1–17. https://doi.org/10.1186/1471-2288-12-159

Herzog, W. (2010). Die Erziehungswissenschaft am Gängelband der Bildungspolitik. *Zeitschrift für pädagogische Histographie* (16), 103–105.

Herzog, W. (2016). Kritik der evidenzbasierten Pädagogik. *Zeitschrift für Erziehungswissenschaft, 19* (S1), 201–213. https://doi.org/10.1007/s11618-016-0711-5

Holland, P. W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association, 81* (396), 945. https://doi.org/10.2307/2289064

Hoogland, J. J. & Boomsma, A. (1998). Robustness Studies in Covariance Structure Modeling. An Overview and a Meta-Analysis. *Sociological Methods & Research, 26* (3), 329–367. https://doi.org/10.1177/0049124198026003003

Hu, L.-t. & Bentler, P. M. (1998). Fit indices in covariance structure modeling. Sensitivity to underparameterized model misspecification. *Psychological Methods, 3* (4), 424–453. https://doi.org/10.1037//1082-989X.3.4.424

Hu, L. & 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. https://doi.org/10.1080/10705519909540118

Hummell, H. J. & Ziegler, R. (1976). *Korrelation und Kausalität* (Flexibles Taschenbuch SOZ). Stuttgart: Enke.

Jaekel, N., Schurig, M., Florian, M. & Ritter, M. (im Erscheinen [2017]. From early starters to late finishers? *Language Learning*.

Jain, A. K. & Dubes, R. C. (1988). *Algorithms for clustering data* (Prentice Hall advanced reference series). Englewood Cliffs, N.J.: Prentice-Hall.

Jöreskog, K. G., Sörbom, D. & Magidson, J. (1979). *Advances in factor analysis and structural equation models*. Cambridge, Mass.: Abt Books.

Jornitz, S. (2009). *Evidenzbasierte Bildungsforschung*. Leverkusen: Budrich Unipress.

Kaplan, D. (2009). *Structural equation modeling. Foundations and extensions* (2. Aufl.). Los Angeles, CA: Sage.

Kasper, D. & Ünlü, A. (2013). On the relevance of assumptions associated with classical factor analytic approaches. *Frontiers in psychology, 4* (109), 1–20. https://doi.org/10.3389/fpsyg.2013.00109

Kaufman, L. & Rousseeuw, P. J. (1990). *Finding Groups in Data*. Hoboken, N.J.: John Wiley & Sons, Inc. https://doi.org/10.1002/SERIES1345

Kelle, U. (2009). *Die Integration qualitativer und quantitativer Methoden in der empirischen Sozialforschung: Theoretische Grundlagen und methodologische Konzepte*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Kenny, D. A., Kaniskan, B. & McCoach, D. B. (2015). The Performance of RMSEA in Models With Small Degrees of Freedom. *Sociological Methods & Research, 44* (3), 486–507. https://doi.org/10.1177/0049124114543236

Kenny, D. A. (1979). *Correlation and causality* (A Wiley-interscience publication). New York, NY: Wiley.

Kenny, D. A., Kashy, D. A. & Bolger, N. (1998). Data analysis in social psychology. In D. T. Gilbert, S. T. Fiske & G. Lindzey (Hrsg.), *The handbook of social psychology* (4. Aufl., S. 233–265). Boston, Mass.: McGraw-Hill.

Klieme, E., Avenarius, H., Blum, W., Döbrich, P., Gruber, H., Prenzel, M. et al. (Hrsg.). (2009). *Zur Entwicklung nationaler Bildungsstandards. Eine Expertise* (Bildungsforschung, Bd. 1, [Nachdr.]. Bonn: BMBF.

Klieme, E. & Baumert, J. (2001). Identifying national cultures of mathematics education. Analysis of cognitive demands and differential item functioning in TIMSS. *European Journal of Psychology of Education, 16* (3), 385–402. https://doi.org/10.1007/BF03173189

Klieme, E. & Hartig, J. (2008). Kompetenzkonzepte in den Sozialwissenschaften und im erziehungswissenschaftlichen Diskurs. In M. Prenzel, I. Gogolin & H.-H. Krüger (Hrsg.)Kompetenzdiagnostik. Zeitschrift für Erziehungswissenschaft. *8,* 11–32 [Themenheft]. Wiesbaden: VS Verlag für Sozialwissenschaften.

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (Methodology in the social sciences, 3. Aufl.). New York, NY: Guilford Press.

Krantz, D. H., Luce, R. D., Suppes, P. & Tversky, A. (2007). *Additive and polynomial representations* (Foundations of measurement, / David H. Krantz … ; Vol. 1). Mineola, N.Y.: Dover Publ.

Kuckartz, U. (2014). *Mixed Methods. Methodologie, Forschungsdesigns und Analyseverfahren*. Wiesbaden: VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-93267-5

Lazarsfeld, P. F. (1959). Latent Structure Analysis. In S. Koch (Hrsg.), *Psychology: A Study of Science* (S. 476–543). New York, NY: McGraw-Hill.

Lazarsfeld, P. F. & Henry, N. W. (1968). *Latent structure analysis*. New York, NY: Houghton Mifflin.

Leutner, D., Klieme, E., Fleischer, J. & Kuper, H. (2013). Editorial. Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen. *Zeitschrift für Erziehungswissenschaft, 16* (S1), 1–4. https://doi.org/10.1007/s11618-013-0378-0

Li, C.-H. (2014). *The performance of MLR, USLMV, and WLSMV estimation in structural regression models with ordinal variables. Dissertation,* Michigan State University. Zugriff am 21.01.2017. Verfügbar unter http://gradworks.umi.com/36/20/3620406.html

Lienert, G. A. & Raatz, U. (1998). *Testaufbau und Testanalyse* (Grundlagen Psychologie, 6. Aufl.). Weinheim: Beltz.

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of applied measurement, 3* (1), 85–106.

Linacre, J. M. (2016). *Fit diagnosis: infit outfit mean-square standardized.* Zugriff am 10.12.2016. Verfügbar unter http://www.winsteps.com/winman/diagnosingmisfit.htm

Little, R. J. & Rubin, D. B. (2000). Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches. *Annual review of public health, 21,* 121–145. https://doi.org/10.1146/annurev.publhealth.21.1.121

Little, T. D., Lindenberger, U. & Nesselroade, J. R. (1999). On selecting indicators for multivariate measurement and modeling with latent variables. When "good" indicators are bad and "bad" indicators are good. *Psychological Methods, 4* (2), 192–211. https://doi.org/10.1037/1082-989X.4.2.192

Lo, Y., Mendell, N. & Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika, 88* (3), 767–778. https://doi.org/10.1093/biomet/88.3.767

Loewenthal, E. (2004). *Plato-Sämtliche Werke in drei Bänden* (Bd. 2, 8. Aufl., 3 Bände). Darmstadt: Wissenschaftliche Buchgesellschaft.

Lord, F. M. & Novick, M. R. (Hrsg.). (1968). *Statistical theories of mental test scores* (Addison-Wesley series in behavioral science). Charlotte, N.C.: Information Age Pub. Inc.

Luce, R. D., Krantz, D. H., Suppes, P. & Tversky, A. (2007). *Representation, axiomatization, and invariance* (Foundations of measurement, / David H. Krantz … ; Vol. 3). Mineola, N.Y.: Dover Publ.

MacCallum, R. C., Roznowski, M. & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: the problem of capitalization on chance. *Psychological Bulletin, 111* (3), 490–504.

MacCallum, R. (1986). Specification searches in covariance structure modeling. *Psychological Bulletin, 100* (1), 107–120. https://doi.org/10.1037/0033-2909.100.1.107

Manski, C. F. (2003). Identification Problems in the Social Sciences and Everyday Life. *Souther Economic Journal, 70* (1), 11–21.

Markus, K. A. & Borsboom, D. (2013). *Frontiers of test validity theory. Measurement, causation, and meaning* (Multivariate applications series). New York, NY: Routledge.

Marsh, H. W., Hau, K. T., Balla, J. R. & Grayson, D. (1998). Is More Ever Too Much? The Number of Indicators per Factor in Confirmatory Factor Analysis. *Multivariate behavioral research, 33* (2), 181–220. https://doi.org/10.1207/s15327906mbr3302_1

Marsh, H. W., Muthén, B., Asparouhov, T., Lüdtke, O., Robitzsch, A., Morin, A. J. S. et al. (2009). Exploratory Structural Equation Modeling, Integrating CFA and EFA. Application to Students' Evaluations of University Teaching. *Structural Equation Modeling: A Multidisciplinary Journal, 16* (3), 439–476. https://doi.org/10.1080/10705510903008220

Maydeu-Olivares, A. & Garcia-Forero, C. (2010). Goodness-of-Fit Testing. In P. Peterson (Hrsg.), *International encyclopedia of education* (3. Aufl., Bd. 7, S. 190–196). Amsterdam u.a.: Elsevier Academic.

Maydeu-Olivares, A. (2013). Goodness-of-Fit Assessment of Item Response Theory Models. *Measurement: Interdisciplinary Research & Perspective, 11* (3), 71–101. https://doi.org/10.1080/15366367.2013.831680

McCutcheon, A. L. (2002). *Latent class analysis* (Sage University papers Quantitative applications in the social sciences, Bd. 64, [Nachdr.]. Newbury Park: Sage.

McDonald, R. P. (1996). Latent Traits and the Possibility of Motion. *Multivariate behavioral research, 31* (4), 593–601. https://doi.org/10.1207/s15327906mbr3104_12

McDonald, R. P. & Mok, M. M. (1995). Goodness of Fit in Item Response Models. *Multivariate behavioral research, 30* (1), 23–40. https://doi.org/10.1207/s15327906mbr3001_2

McDonald, R. P. (1999). *Test theory. A unified treatment*. Mawah, N.J.: Lawrence Erlbaum Associates Publishers.

McElvany, N. & Schwabe, F. (2013). Fairness von Lesetestaufgaben für Kinder aus Familien mit unterschiedlichem sozioökonomischem Status bei Large-Scale-Studien. In N. McElvany & H. G. Holtappels (Hrsg.), *Empirische Bildungsforschung. Theorien, Methoden, Befunde und Perspektiven : Festschrift für Wilfried Bos* (S. 219–233). Münster: Waxmann.

McLachlan, G. & Peel, D. (2000). *Finite Mixture Models*. Hoboken, N.J.: John Wiley & Sons, Inc. https://doi.org/10.1002/0471721182

Meade, A. W. & Lautenschlager, G. (2004). A Comparison of Item Response Theory and Confirmatory Factor Analytic Methodologies for Establishing Measurement Equivalence/Invariance. *Organizational Research Methods, 7* (4), 361–388. https://doi.org/10.1177/1094428104268027

Meade, A. W., Johnson, E. C. & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *The Journal of applied psychology, 93* (3), 568–592. https://doi.org/10.1037/0021-9010.93.3.568

Meijer, R. R. & Sijtsma, K. (2001). Methodology Review. Evaluating Person Fit. *Applied Psychological Measurement, 25* (2), 107–135. https://doi.org/10.1177/01466210122031957

Merkens, H. (2003). Immunisierung gegen Kritik durch Methodisierung der Kritik. In D. Benner, M. Borrelli, F. Heyting & C. Winch (Hrsg.), *Kritik in der Pädagogik. (Zeitschrift für Pädagogik, Beiheft; 46)* (S. 33–53). Weinheim: Beltz.

Messick, S. (1975). The standard problem. Meaning and values in measurement and evaluation. *American Psychologist, 30* (10), 955–966. https://doi.org/10.1037/0003-066X.30.10.955

Messick, S. (1981). Evidence and Ethics in the Evaluation of Tests. *Educational Researcher, 10* (9), 9–20. https://doi.org/10.3102/0013189X010009009

Messick, S. (1995). Validity of psychological assessment. Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50* (9), 741–749. https://doi.org/10.1037/0003–066X.50.9.741

Millsap, R. E. (2007). Structural equation modeling made difficult. *Personality and Individual Differences, 42* (5), 875–881. https://doi.org/10.1016/j.paid.2006.09.021

Mislevy, R. J. & Haertel, G. D. (2006). Implications of Evidence-Centered Design for Educational Testing. *Educational Measurement: Issues and Practice, 25* (4), 6–20. https://doi.org/10.1111/j.1745-3992.2006.00075.x

Moosbrugger, H. & Kelava, A. (Hrsg.). (2008). *Testtheorie und Fragebogenkonstruktion* (Springer-Lehrbuch). Berlin: Springer. https://doi.org/10.1007/978-3-540-71635-8

Moustaki, I. & Knott, M. (2000). Generalized latent trait models. *Psychometrika, 65* (3), 391–411. https://doi.org/10.1007/BF02296153

Mueller, R. O. & Hancock, G. R. (2008). Best practices in structural equation modeling. In J. W. Osborne (Hrsg.), *Best practices in quantitative methods* (S. 488–508). Los Angeles, CA: Sage.

Mulaik, S. (1995). The Metaphoric Origins of Objectivity, Subjectivity, and Consciousness in the Direct Perception of Reality. *Philosophy of Science, 62* (2), 283–303.

Mulaik, S. (2004). Objectivity in Science and Structural Equation Modeling. In D. Kaplan (Hrsg.), *The SAGE Handbook of Quantitative Methodology for the Social Sciences* (S. 425–447). Thousand Oaks, Ca.: Sage. https://doi.org/10.4135/9781412986311.n23

Mulaik, S. A. (2001). The Curve-Fitting Problem. An Objectivist View. *Philosophy of Science, 68* (2), 218–241. https://doi.org/10.1086/392874

Mulaik, S. A. & Millsap, R. E. (2000). Doing the Four-Step Right. *Structural Equation Modeling: A Multidisciplinary Journal, 7* (1), 36–73. https://doi.org/10.1207/S15328007SEM0701_02

Muraki, E. (1992). A Generalized Partial Credit Model. Application of an EM Algorithm. *ETS Research Report Series, 1992* (1), i-30. https://doi.org/10.1002/j.2333-8504.1992.tb01436.x

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika, 49* (1), 115–132. https://doi.org/10.1007/BF02294210

Muthén, B. (2007). Latent Variable Hybrids. In G. R. Hancock & K. M. Samuelsen (Hrsg.), *Advances in Latent Variable Mixture Models* (S. 1–24). Charlotte, N.C.: Information Age Pub. Inc.

Muthén, B. & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology, 38* (2), 171–189. https://doi.org/10.1111/j.2044-8317.1985.tb00832.x

Muthén, B. O. (2002). Beyond SEM. General latent Variable Modelling. *Behaviormetrika, 29* (1), 81–117. https://doi.org/10.2333/bhmk.29.81

Muthén, L. K. & Muthén, B. (1998-2015). *Mplus User's Guide. Seventh Edition*. Los Angeles, CA: Muthén & Muthén.

Muthén, L. K. & Muthén, B. O. (2002). How to Use a Monte Carlo Study to Decide on Sample Size and Determine Power. *Structural Equation Modeling: A Multidisciplinary Journal, 9* (4), 599–620. https://doi.org/10.1207/S15328007SEM0904_8

Newton, P. E. (10.01.2017). *A collection of types of validity: Validity modifier label glossary 29.4.2013* (E-Mail).

Newton, P. E. & Shaw, S. D. (2013). Standards for talking and thinking about validity. *Psychological Methods, 18* (3), 301–319. https://doi.org/10.1037/a0032969

Nunnally, J. C. (1978). *Psychometric theory* (McGraw-Hill series in psychology, 2. Aufl.). New York, NY: McGraw-Hill [u.a.].

Nylund, K. L., Asparouhov, T. & Muthén, B. O. (2007). Deciding on the Number of Classes in Latent Class Analysis and Growth Mixture Modeling. A Monte Carlo Simulation Study. *Structural Equation Modeling: A Multidisciplinary Journal, 14* (4), 535–569. https://doi.org/10.1080/10705510701575396

OECD. (2012). *PISA 2009 Technical Report:* OECD Publishing. https://doi.org/10.1787/9789264167872-en

Oertzen, T. von, Brandmaier, A. M. & Tsang, S. (2014). Structural Equation Modeling With Ωnyx. *Structural Equation Modeling: A Multidisciplinary Journal, 22* (1), 148–161. https://doi.org/10.1080/10705511.2014.935842

Orth, B. (1974). *Einführung in die Theorie des Messens* (Kohlhammer-Standards Psychologie : Studientext : Mathematische Psychologie). Stuttgart: Kohlhammer.

Pearl, J. (2000). *Causality. Models, reasoning, and inference*. Cambridge, Mass.: Cambridge Univ. Press.

Pearl, J. (2013). Linear Models - A Useful "Microscope" for Causal Analysis. *Journal of Causal Inference, 1* (1). https://doi.org/10.1515/jci-2013-0003

Picht, G. (1964). *Die deutsche Bildungskatastrophe. Analyse u. Dokumentation*. Olten: Walter.

Popper, K. R. (1994). *Logik der Forschung* (Bd. 4, 10. Aufl.). Tübingen: Mohr.

Prenzel, M. (2005). Zur Situation der Empirischen Bildungsforschung. In H. Mandl & B. Kopp (Hrsg.), *Impulse für die Bildungsforschung. Stand und Perspektiven. Dokumentation eines Expertengesprächs* (S. 7–21). Berlin: Akademie Verlag.

Rabe-Hesketh, S., Skrondal, A. & Pickles, A. (2004). Generalized multilevel structural equation modeling. *Psychometrika, 69* (2), 167–190. https://doi.org/10.1007/BF02295939

Rammstedt, B. (2010). Reliabilität, Validität, Objektivität. In C. Wolf & H. Best (Hrsg.), *Handbuch der sozialwissenschaftlichen Datenanalyse* (S. 239–258). Wiesbaden: VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-92038-2_11

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests* (Erw. Aufl.). Chicago, Ill.: Univ. of Chicago Press.

Raykov, T. (1997). Scale Reliability, Cronbach's Coefficient Alpha, and Violations of Essential Tau-Equivalence with Fixed Congeneric Components. *Multivariate behavioral research, 32* (4), 329–353. https://doi.org/10.1207/s15327906mbr3204_2

Raykov, T. & Marcoulides, G. A. (2006). *A first course in structural equation modeling* (2. Aufl.). Mawah, N.J.: Lawrence Erlbaum Associates Publishers.

Reinecke, J. (2014). *Strukturgleichungsmodelle in den Sozialwissenschaften* (2. Aufl.). München: De Gruyter Oldenbourg.

Rigdon, E. E. (1995). A Necessary and Sufficient Identification Rule for Structural Models Estimated in Practice. *Multivariate behavioral research, 30* (3), 359–383. https://doi.org/10.1207/s15327906mbr3003_4

Rombach, H. (1970). *Lexikon der Pädagogik*. Freiburg i. Br.: Herder.

Rossiter, J. R. (2008). Content Validity of Measures of Abstract Constructs in Management and Organizational Research. *British Journal of Management, 19* (4), 380–388. https://doi.org/10.1111/j.1467-8551.2008.00587.x

Rost, J. (1990). Rasch Models in Latent Classes. An Integration of Two Approaches to Item Analysis. *Applied Psychological Measurement, 14* (3), 271–282. https://doi.org/10.1177/014662169001400305

Rost, J. (1996). *Lehrbuch Testtheorie, Testkonstruktion* (Aus dem Programm Huber). Bern: Huber.

Roth, H. (1963). Die realistische Wendung in der pädagogischen Forschung. *Die Deutsche Schule* (55), 109–119.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66* (5), 688–701. https://doi.org/10.1037/h0037350

Ryan, R. M. & Sapp, A. (2005). Zum Einfluss testbasierter Reformen: High Stakes Testing (HST). Motivation und Leistung aus Sicht der Selbstbestimmungstheorie. *Unterrichtswissenschaft, 33* (2), 143–159. Verfügbar unter http://nbn-resolving.de/urn:nbn:de:0111-opus-57919

Saint-Mont, U. (2011). *Statistik im Forschungsprozess. Eine Philosophie der Statistik als Baustein einer integrativen Wissenschaftstheorie*. Heidelberg: Physica-Verlag. https://doi.org/10.1007/978-3-7908-2723-1

Saldern, M. von. (1992). Qualitative Forschung – quantitative Forschung: Nekrolog auf einen Gegensatz. *Empirische Pädagogik, 6* (4), 377–399.

Satorra, A. & Bentler, P. M. (1994). Corrections to Test Statistics and Standard Errors in Covariance Structure Analysis. In A. v. Eye & C. C. Clogg (Hrsg.), *Latent variables analysis. Applications for developmental research* (S. 399–419). Thousand Oaks, Ca.: Sage.

Schmidt, B. & Weishaupt, H. (2008). Forschung und wissenschaftlicher Nachwuchs. In K.-J. Tillmann (Hrsg.), *Datenreport Erziehungswissenschaft 2008* (S. 113–138).

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8* (4), 350–353. https://doi.org/10.1037/1040-3590.8.4.350

Schroeder-Heister, P. (2013). Wahrscheinlichkeit. In H. Keuth (Hrsg.), *Karl Popper, Logik der Forschung* (Klassiker auslegen, Bd. 12, 4., bearb. Aufl., S. 187–215). Berlin: Akad.-Verl.

Schulte, K., Nonte, S. & Schwippert, K. (2013). Die Überprüfung von Messinvarianz in international vergleichenden Schulleistungsstudien am Beispiel der Studie PIRLS. *Zeitschrift für Bildungsforschung, 3* (2), 99–118. https://doi.org/10.1007/s35834-013-0062-8

Schurig, M. & Busch, V. (2014). Entwicklung der Musikpräferenz von Grundschulkindern. Individuelle, soziale und musikbezogene Einflüsse. In A. Lehmann-Wermser, V. Busch, K. Schwippert & S. Nonte (Hrsg.), *Mit Mikrofon und Fragebogen in die Grundschule. Jedem Kind ein Instrument (JeKi) – eine empirische Längsschnittstudie zum Instrumentalunterricht* (S. 63–96). Münster: Waxmann.

Schurig, M., Busch, V. & Strauß, J. (2012). Effects of Structural and Personal Variables on Children's Development of Music Preference. In E. Cambouropoulos, C. Tsougras, P. Mavromatis & K. Pastiadis (Hrsg.), *Proceedings of the 12th International Conference on Music Perception and Cognition (ICMPC) and the 8th Triennial Conference of European Society for the Cognitive Sciences of Music (ESCOM)* (S. 896–902). Thessaloniki: School of Music Studies, Aristotle University of Thessaloniki. Zugriff am 17.01.2017. Verfügbar unter http://icmpc-escom2012.web.auth.gr/files/papers/896_Proc.pdf

Schurig, M., Glesemann, B. & Schröder, J. (2016). Dimensionen von Unterrichtsqualität. Die Generalisierbarkeit von Schülerurteilen über Fächer und Zeit. In R. Strietholt, W. Bos, H. G. Holtappels & N. McElvany (Hrsg.), *Jahrbuch der Schulentwicklung Band 19. Daten, Beispiele und Perspektiven* (S. 30–56). Weinheim: Beltz Juventa.

Schurig, M., Wendt, H., Kasper, D. & Bos, W. (2015). Fachspezifische Stärken und Schwächen von Viertklässlerinnen und Viertklässlern in Deutschland im europäischen Vergleich. In H. Wendt, T. C. Stubbe & K. Schwippert (Hrsg.), *10 Jahre international vergleichende Schulleistungsforschung in der Grundschule. Vertiefende Analysen zu IGLU und TIMSS 2001 bis 2011* (S. 35–54). Münster: Waxmann.

Schwippert, K. (2016). Empirische Bildungsforschung: Perspektiven der Erziehungswissenschaft. In D. Fickermann & H.-W. Fuchs (Hrsg.), *Bildungsforschung – disziplinäre Zugänge. Fragestellungen, Methoden und Ergebnisse* (S. 25–37). Münster: Waxmann.

Schwippert, K. & Goy, M. (2008). Leistungsvergleichs- und Schulqualitätsforschung. In W. Helsper & J. Böhme (Hrsg.), *Handbuch der Schulforschung* (2. Aufl., S. 387–421). Wiesbaden: VS Verlag für Sozialwissenschaften.

Sedlmeier, P. (2009). Beyond the Significance Test Ritual. *Zeitschrift für Psychologie / Journal of Psychology, 217* (1), 1–5. https://doi.org/10.1027/0044-3409.217.1.1

Sedlmeier, P. & Renkewitz, F. (2013). *Forschungsmethoden und Statistik für Psychologen und Sozialwissenschaftler* (2. Aufl.). München: Pearson.

Shadish, W. R., Cook, T. D. & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference* [Nachdr.]. Belmont, CA: Wadsworth Cengage Learning.

Shikano, S. (2010). Einführung in die Inferenz durch den nichtparametrischen Bootstrap. In C. Wolf & H. Best (Hrsg.), *Handbuch der sozialwissenschaftlichen Datenanalyse* (S. 191–204). Wiesbaden: VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-92038-2_9

Shrout, P. E. & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86* (2), 420–428. https://doi.org/10.1037/0033-2909.86.2.420

Skrondal, A. & Rabe-Hesketh, S. (2007). Latent Variable Modelling. A Survey*. *Scandinavian Journal of Statistics, 34* (4), 712–745. https://doi.org/10.1111/j.1467-9469.2007.00573.x

Skrondal, A. & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling. Multilevel, longitudinal, and structural equation models* (Chapman & Hall / CRC interdisciplinary statistics series). Boca Raton, Fl.: Chapman & Hall/CRC.

Smith, R. M., Schumacker, R. E. & Bush, M. J. (1998). Using item mean squares to evaluate fit to the Rasch model. *Journal of outcome measurement, 2* (1), 66–78.

Steiger, J. H. (2007). Understanding the limitations of global fit assessment in structural equation modeling. *Personality and Individual Differences, 42* (5), 893–898. https://doi.org/10.1016/j.paid.2006.09.017

Stevens, S. S. (1946). On the Theory of Scales of Measurement. *Science, 103* (2684), 677–680.

Stevens, S. S. (1957). On the psychophysical law. *Psychological Review, 64* (3), 153–181. https://doi.org/10.1037/h0046162

Steyer, R. (1992). *Theorie kausaler Regressionsmodelle*. Stuttgart: Gustav Fischer Verlag.

Steyer, R. (1989). Models of Classical Psychometric Test Theory as Stochastic Measurement Models: Representation, Uniqueness, Meaningfulness, Identifiability, and Testability. *Methodika* (3), 25–60. Zugriff am 15.11.2016. Verfügbar unter https://www.metheval.uni-jena.de/materialien/publikationen/steyer1989_models_of_classical_psychometric_test_theory.pdf

Steyer, R. & Eid, M. (2001). *Messen und Testen. Mit Übungen und Lösungen* (Springer-Lehrbuch, 2. Aufl.). Berlin: Springer Verl. https://doi.org/10.1007/978-3-642-56924-1

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science, 25* (1), 1–21. https://doi.org/10.1214/09-STS313

Suppes, P., Krantz, D. H., Luce, R. D. & Tversky, A. (2007). *Geometrical, thresholg, and prohabilistic representations* (Foundations of measurement, / David H. Krantz … ; Vol. 2). Mineola, N.Y.: Dover Publ.

Szlezák, T. A. (2003). *Die Idee des Guten in Platons Politeia. Beobachtungen zu den mittleren Büchern* (Lecturae Platonis, Bd. 3). St. Augustin: Academia-Verl.

Tabachnick, B. G. & Fidell, L. S. (2007). *Using multivariate statistics* (5. Aufl.). Boston, Mass.: Pearson/Allyn & Bacon.

Tarnai, C. & Rost, J. (1990). *Identifying aberrant response patterns in the Rasch model. The Q index* (Sozialwissenschaftliche Forschungsdokumentationen, Bd. 1). Münster: Inst. f. Sozialwiss. Forschung.

Tashakkori, A. M. & Teddlie, C. B. (2010). *Sage handbook of mixed methods in social & behavioral research* (2. Aufl.). Los Angeles, CA: Sage.

Terhart, E. (2012). „Bildungswissenschaften": Verlegenheitslösung, Sammeldisziplin, Kampfbegriff? *Zeitschrift für Pädagogik, 58* (1), 22–39.

Thissen, D., Steinberg, L. & Gerrard, M. (1986). Beyond group-mean differences. The concept of item bias. *Psychological Bulletin, 99* (1), 118–128. https://doi.org/10.1037/0033-2909.99.1.118

Thistlethwaite, D. L. & Campbell, D. T. (1960). Regression-discontinuity analysis. An alternative to the ex post facto experiment. *Journal of Educational Psychology, 51* (6), 309–317. https://doi.org/10.1037/h0044319

Tippelt, R. (1998). Zum Verhältnis von Allgemeiner Pädagogik und empirischer Bildungsforschung. *Zeitschrift für Erziehungswissenschaft, 1* (2), 239–260.

Tippelt, R. (Hrsg.). (2009). *Steuerung durch Indikatoren. Methodologische und theoretische Reflektionen zur deutschen und internationalen Bildungsberichterstattung* (Vorstandsreihe der Deutschen Gesellschaft für Erziehungswissenschaft). Opladen: Budrich.

Tippelt, R. & Schmidt, B. (Hrsg.). (2010). *Handbuch Bildungsforschung* (3. Aufl.). Wiesbaden: VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-92015-3

Tofighi, D. & Enders, C. K. (2008). Identifying the correct number of classes in growth mixture models. In G. R. Hancock & K. M. Samuelsen (Hrsg.), *Advances in Latent Variable Mixture Models* (S. 317–341). Charlotte, N.C.: Information Age Pub. Inc.

Torres Irribarra, D. & Freund, R. (2014). *Wright Map: IRT item-person map with ConQuest integration.* Zugriff am 05.01.2017. Verfügbar unter http://github.com/david-ti/wrightmap

Tukey, J. W. (1977). *Exploratory data analysis* (Addison-Wesley series in behavioral science). Reading, Mass.: Addison-Wesley Pub. Co.

Van de Schoot, R., Schmidt, P., Beuckelaer, A. de, Lek, K. & Zondervan-Zwijnenburg, M. (2015). Editorial: Measurement Invariance. *Frontiers in psychology, 6,* 1064. https://doi.org/10.3389/fpsyg.2015.01064

Vandenberg, R. J. & Lance, C. E. (2000). A Review and Synthesis of the Measurement Invariance Literature. Suggestions, Practices, and Recommendations for Organizational Research. *Organizational Research Methods, 3* (1), 4–70. https://doi.org/10.1177/109442810031002

Weber, M., Winckelmann, J. & Weber, M. (Hrsg.). (1988). *Gesammelte Aufsätze zur Wissenschaftslehre* (UTB für Wissenschaft, Bd. 1492, 7. Aufl.). Tübingen: Mohr.

Weiber, R. & Mühlhaus, D. (2010). *Strukturgleichungsmodellierung. Eine anwendungsorientierte Einführung in die Kausalanalyse mit Hilfe von AMOS, SmartPLS und SPSS ; [Extras im Web]* (Springer-Lehrbuch). Heidelberg: Springer.

Weinert, F. E. (2002). Vergleichende Leistungsmessung in der Schule - eine umstrittene Selbstverständlichkeit. In F. E. Weinert (Hrsg.), *Leistungsmessungen in Schulen* (Beltz Pädagogik, 2. Aufl., S. 17–32). Weinheim: Beltz.

Wilson, M. (2003). On Choosing a Model for Measuring. *Methods of Psychological Research Online, 8* (3), 1–22. Zugriff am 21.11.2016. Verfügbar unter http://www.dgps.de/fachgruppen/methoden/mpr-online/issue21/mpr122_8.pdf

Wirth, R. J. & Edwards, M. C. (2007). Item factor analysis: current approaches and future directions. *Psychological Methods, 12* (1), 58–79. https://doi.org/10.1037/1082-989X.12.1.58

Wischer, B. & Tillmann, K.-J. (Hrsg.). (2009). *Erziehungswissenschaft auf dem Prüfstand. Schulbezogene Forschung und Theoriebildung von 1970 bis heute*. Weinheim: Juventa Verlag.

Wolf, F. M. (2008). *Meta-analysis. Quantitative methods for research synthesis* (A Sage university paper : Quantitative applications in the social sciences, Bd. 59, 17. Aufl.). Newbury Park: Sage.

Wright, B. D. & Masters, G. N. (1982). *Rating Scale Analysis. Rasch measurement*. Chicago, Ill.: Mesa Press.

Wurpts, I. C. & Geiser, C. (2014). Is adding more indicators to a latent class analysis beneficial or detrimental? Results of a Monte-Carlo study. *Frontiers in psychology, 5,* 920. https://doi.org/10.3389/fpsyg.2014.00920

Yen, W. M. (1984). Effects of Local Item Dependence on the Fit and Equating Performance of the Three-Parameter Logistic Model. *Applied Psychological Measurement, 8* (2), 125–145. https://doi.org/10.1177/014662168400800201

Yousfi, S. & Steyr, R. (2006). Messtheoretische Grundlagen der Psychologischen Diagnostik. In F. Petermann, M. Eid & J. Bengel (Hrsg.), *Handbuch der psychologischen Diagnostik* (Handbuch der Psychologie, Bd. 4, S. 46–56). Göttingen: Hogrefe.

Yuan, K.-H., Chan, W. & Bentler, P. M. (2000). Robust transformation with applications to structural equation modelling. *British Journal of Mathematical and Statistical Psychology, 53* (1), 31–50. https://doi.org/10.1348/000711000159169

Zedler, P. & Döbert, H. (2010). Erziehungswissenschaftliche Bildungsforschung. In R. Tippelt & B. Schmidt (Hrsg.), *Handbuch Bildungsforschung* (3. Aufl., S. 23–45). Wiesbaden: VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-92015-3_2

Zhang, D., Ji, M., Yang, J., Zhang, Y. & Xie, F. (2014). A novel cluster validity index for fuzzy clustering based on bipartite modularity. *Fuzzy Sets and Systems, 253,* 122–137. https://doi.org/10.1016/j.fss.2013.12.013

ANHANG 1

Tabelle 8: Einhundertdreiundfünfzig Kriterien für Validität (vgl. Newton & Shaw, 2013, S. 313)

Abstract	Criteria	External	Judgmental	Retrospective
Administrative	Correlational	External test	Known-groups	Sampling
Aetiological	Criterion	Extratest	Linguistic	Scientific
Artifactual	Criterion-oriented	Face	Local	Scoring
Behavior domain	Criterion-related	Factorial	Logical	Self-defining
Cash	Criterion-relevant	Faith	Longitudinal	Semantic
Circumstantial	Cross-age	Fiat	Lower-order	Single-group
Cluster domain	Cross-cultural	Forecast true	Manifest	Site
Cognitive	Cross-sectional	Formative	Natural	Situational
Common sense	Cultural	Functional	Nomological	Specific
Communication	Curricular	General	Occupational	Statistical
Concept	Decision	Generalized	Operational	Status
Conceptual	Definitional	Generic	Particular	Structural
Concrete	Derived	Higher-order	Performance	Substantive
Concurrent	Descriptive	In situ	Postdictive	Summative
Concurrent criterion	Design	Incremental	Practical	Symptom
Concurrent criterion-related	Diagnostic	Indirect	Predictive	Synthetic
Concurrent true	Differential	Inferential	Predictive criterion	System
Congruent	Direct	Instructional	Predictor	Systemic
Congruent	Discriminant	Internal	Prima Facie	Theory-based
Consensual	Discriminative	Internal test	Procedural	Theoretical
Consequential	Divergent	Interpretative	Prospective	Trait
Construct	Domain	Interpretive	Psychological & logical	Translation
Constructor	Domain-selection	Intervention	Psychometric	Translational
Construct-related	Edumetric	Intrinsic	Quantitative face	Treatment
Content	Elaborative	Intrinsic content	Rational	True
Content sampling	Elemental	Intrinsic correlational	Raw	User
Content-related	Empirical	Intrinsic rational	Relational	Washback
Context	Empirical-judgemental	Item	Relevant	
Contextual	Essential	Job analytic	Representational	
Convergent	Etiological	Job component	Response	

Anm.: Mehrere Validitätsformen sind bei verschiedenen Autoren unterschiedlich definiert worden. Eine annotierte und übersetzte Tabelle der Vorlage von P. E. Newton (persönl. Mitteilung, 10.01.2017) mit den vollständigen Quellangaben kann beim Autor erfragt werden.

ANHANG 2

Übersicht der aufgenommenen Schriften[57]

1. Schurig, M., Busch, V. & Strauß, J. (2012). Effects of Structural and Personal Variables on Children's Development of Music Preference. In: E. Cambouropoulos, C. Tsougras, P. Mavromatis & K. Pastiadis (Hrsg.), *Proceedings of the 12th International Conference on Music Perception and Cognition (ICMPC) and the 8th Triennial Conference of European Society for the Cognitive Sciences of Music (ESCOM)* (S. 896-902). Thessaloniki: School of Music Studies, Aristotle University of Thessaloniki.
2. Schurig, M. & Busch, V. (2014). Entwicklung der Musikpräferenz von Grundschulkindern. Individuelle, soziale und musikbezogene Einflüsse. In: A. Lehmann-Wermser, V. Busch, K. Schwippert & S. Nonte (Hrsg.), *Mit Mikrofon und Fragebogen in die Grundschule. Jedem Kind ein Instrument (JeKi) – eine empirische Längsschnittstudie zum Instrumentalunterricht* (S. 63-96). Münster: Waxmann.
3. Busch, V., Schurig, M., Bunte, N. & Beutler-Prahm, B. (2015). „Mir gefällt ja mehr diese Rockmusik." Zur Struktur musikalischer Präferenzurteile im Grundschulalter. In: W. Auhagen, C. Bullerjahn & R. von Georgi (Hrsg.), *Offenohrigkeit - Ein Postulat im Fokus (Jahrbuch der Deutschen Gesellschaft für Musikpsychologie - Band 24)* (S. 133-168). Göttingen [u.a.]: Hogrefe.
4. Schurig, M., Wendt, H., Kasper, D. & Bos, W. (2015). Fachspezifische Stärken und Schwächen von Viertklässlerinnen und Viertklässlern in Deutschland im europäischen Vergleich. In H. Wendt, T. C. Stubbe & K. Schwippert (Hrsg.), *10 Jahre international vergleichende Schulleistungsforschung in der Grundschule. Vertiefende Analysen zu IGLU und TIMSS 2001 bis 2011* (S. 35–54). Münster: Waxmann.
5. Schurig, M., Glesemann, B. & Schröder, J. (2016). Dimensionen von Unterrichtsqualität. Die Generalisierbarkeit von Schülerurteilen über Fächer und Zeit. In R. Strietholt, W. Bos, H. G. Holtappels & N. McElvany (Hrsg.), *Jahrbuch der Schulentwicklung Band 19. Daten, Beispiele und Perspektiven* (S. 30–56). Weinheim: Beltz Juventa.
6. Jaekel, N., Schurig, M., Florian, M. & Ritter, M. (im Erscheinen [2017]). From early starters to late finishers? A Longitudinal Study of Early Foreign Language Learning in School. *Language Learning*.

[57] Die jeweilige Relevanz der Beiträge wird in kurzen Erläuterungen gegeben.

BEITRAG 1

Erschienen in (Zitierweise):

Schurig, M., Busch, V. & Strauß, J. (2012). Effects of Structural and Personal Variables on Children's Development of Music Preference. In: E. Cambouropoulos, C. Tsougras, P. Mavromatis & K. Pastiadis (Hrsg.), *Proceedings of the 12th International Conference on Music Perception and Cognition (ICMPC) and the 8th Triennial Conference of European Society for the Cognitive Sciences of Music (ESCOM)* (S. 896-902). Thessaloniki: School of Music Studies, Aristotle University of Thessaloniki.

Relevanz:

In dem Beitrag wird eine latente Struktur faktoriell erschlossen, nachdem eine konfirmatorische Prüfung zurückgewiesen werden musste. Der Beitrag wird als Exempel für das Erschließen einer operationalen Definition, eine Nominalen theoretischen Begriffs herangeführt. Eine mehrdimensionale latente Struktur wird operational über die Verwendung Explorativer Faktoranalysen (EFA) abgeleitet und wiederspricht damit Angemessenheit der Abbildung eines theoretischen Begriffs in einer eindimensionalen Struktur. Die beobachtete Struktur wird über Replikationen der Analysen in mehreren Jahrgangsstufen auf der Basis der gleichen Stichprobe abgesichert. Um explorativ auch Informationen über das relative Gewicht möglicher Prädiktoren für vertiefende Analysen zu erlangen, werden Multiple-Causes-Multiple-Indicator (MIMIC) Modelle zu jedem der vorliegenden Messzeitpunkte herangeführt.

Proceedings of the 12th International Conference on Music Perception and Cognition and the 8th Triennial Conference of the European Society for the Cognitive Sciences of Music, July 23-28, 2012, Thessaloniki, Greece
Cambouropoulos E., Tsougras C., Mavromatis P., Pastiadis K. (Editors)

Effects of Structural and Personal Variables on Children's Development of Music Preference

Michael Schurig,[*1] Veronika Busch,[*2] and Julika Strauß[*3]

[*] *Department of Musicology and Music Education, University of Bremen, Germany*
[1]schurig@uni-bremen.de, [2]veronika.busch@uni-bremen.de, [3]julika.strauss@uni-bremen.de

ABSTRACT

Hargreaves' (1982) hypothesis of an age-related decline in children's preference for unfamiliar music genres ("open-earedness") forms the theoretical background of our longitudinal study with four points of measurement between grade one and four. Primary school children answered a sound questionnaire with 8 music examples on a 5-point iconic preference scale. Structural and personal data was collected using standardized questionnaires, and complementary interviews were conducted.

We operationalized open-earedness as a latent construct with "classic" and "ethnic/avant-garde" music preference (Louven, 2011) as distinguishable factors through exploratory factor analyses. The aim is to identify predictor variables (e.g. gender, personality, music experience, migration background, and socio-economic status) using structural equation modelling. This way we tried to assess a measurement model to be used for further investigation of our longitudinal data.

So far, analyses of variance support the expected open-earedness for preference ratings of t_1 (n_1=617), but gender differences already show. Analyses of t_2 (n_2=1142) disclose the beginning decline of open-earedness, with t_3 (n_3=1132) supporting the trend furthermore.

By now, no differences in preference ratings according to migration background, socio-economic status, or music experience were observed in the MIMIC models. Cognitive domains and personality contribute only very low effects. Thus, up to this stage of our analyses, age and gender remain the prime indicators for music preference. Qualitative data also stresses gender differences. Repeated measurement analyses will provide further information on the development of music preference.

I. INTRODUCTION

Musical preference has been extensively investigated over the last decades. Research has covered developmental, social, personal, as well as musical aspects that might influence music preference. Different models have been proposed to group and explain music preference ratings for a variety of music examples from different music genres. Louven (2011) statistically distinguished between preferences of "pop", "classic" and "ethnic/avant-garde" music styles by using principle components analyses. Rentfrow et al. (2011) mention three factors they found in almost all reviewed studies: "classical/jazz", "rock/heavy metal", and "rap/hip hop". In addition to genre-based factors, the authors offer an empirically evolved Five-Factor-Model (MUSIC: "Mellow", "Unpretentious", "Sophisticated", "Intense", and "Contemporary"), as framework for future research, which takes specific musical features and their psychological effects into account (e.g. factor "Mellow": smooth / relaxing; factor "Sophisticated": complex, intelligent, inspiring; Rentfrow et al. (2011). They argue for multiple influences on music preference, like psychological dispositions, social interactions, and exposure to popular media and cultural trend and they point at the curiosity that we do know of the importance of music for people, but that we do not yet know why. Schäfer and Sedlmeier (2009) might offer a possible answer with their investigation of different functions music appears to fulfil.

A time, during which music seems to fulfil many functions and hence appears to be of high importance, is puberty. Hargreaves (1982) described young children as "open-eared" (p. 51), whereas juveniles seemed to have lost this openness for unfamiliar musical styles. His term "open-eared" generated lots of research activities into the development of music preference. Within this research the so-called "open-earedness" is generally understood as acceptance of a large variety of unfamiliar pieces of music. This hypothesis forms the theoretical background of our study on the development of music preference of primary school children.

Previous research generally supports an age-dependency for open-earedness (Hargreaves et al., 2006). LeBlanc (1991; LeBlanc et al., 1996) proposed four stages for age-related differences over the life-span: (1) young children are initially open-eared; (2) they lose their musical openness on the way to puberty; (3) young adults open up again; and (4) older adults show a decline in open-earedness.

While some authors support a decline in open-earedness already during primary school (Gembris & Schellenberg, 2003) or even before (Hargreaves, 1987), Kopiez & Lehmann (2008) provide empirical evidence that the whole primary school time should be seen as a period of open-earedness.

In addition to age-related differences, gender-specific effects were investigated. Most studies point towards differences according to gender and describe girls as more open-eared than boys (Hargreaves, 1995; Gembris & Schellberg 2003, 2007). But Kopiez & Lehmann (2008) do not support these findings and generally found only small effect sizes for gender-specific differences in music preference.

Research has also shown that personality traits influence music preference of juveniles and adults (Delsing et al., 2008; Rawlings & Ciancarelli, 1997). The importance of personality aspects for young children's music preference could thus be assumed, but is difficult to study as their personality is still developing. Research literature provides evidence for a critical time window, during which children are more sensitive towards influences on their developing musical taste (Gembris, 2005; Hargreaves et al., 2006). Young children's open-earedness could in reverse be interpreted as an indicator for this critical time window and might even be seen as an expression of the personality trait "openness to experience" described in the Five-Factor-Model by Costa & McCrae (1992).

In Germany various programs were started to offer children within this critical time window access to learning music instruments and experience musical styles other than mainstream pop (e.g. "JeKi – Jedem Kind ein Instrument").

Proceedings of the 12th International Conference on Music Perception and Cognition and the 8th Triennial Conference of the European Society for the Cognitive Sciences of Music, July 23-28, 2012, Thessaloniki, Greece
Cambouropoulos E., Tsougras C., Mavromatis P., Pastiadis K. (Editors)

Quite often those programs are motivated by expectations regarding transfer effects (especially cognitive ones) and also concerning effects on children's developing music preference (meaning: preserving their initial open-earedness).

Further research indicates that juveniles and adults with migration background prefer music from their country of origin (Sakai, 2011; Cremades et al., 2010; Henninger, 1999; Teo et al., 2008). This music often plays an important role in family life. And it could thus be argued, that "emancipation" from the parents' musical taste is delayed compared to children with no migration background (Greve, 2003; Wurm, 2006; Baumann, 1985).

Additionally, high socio-economic status seems to generally have a positive influence on musical openness (Eijck, 2001; Peterson, 1992).

But so far no satisfying answer has been given as to how these different influencing variables interact and whether they are able to predict open-earedness.

II. AIMS AND QUESTIONS

Our study addresses these general questions concerning the construct of open-earedness. The study is integrated into a larger cooperation project of the Universities of Bremen and Hamburg that investigates the effects of intensified music education of primary school children (for further information on the project visit: www.sigrun2009.de).

A. Aims

The main objective of our study is to analyse the plausibility to interpret open-earedness as a latent construct with distinct predictor variables for further usage.

Therefore we will investigate the influence of independent structural and personal variables on music preference ratings of primary school children. We will try to aggregate several observable variables into a model that might represent open-earedness.

This should provide us with refined ideas of how to construct a measurement model to be used in later latent class change and latent growth curve models. While concluding those tasks there is a wide array of questions that can be answered for our sample alongside.

B. Questions and Hypotheses

The basic question of the presented study is, whether the decline of open-earedness can be predicted.

Thus, our null hypotheses (H_0) would be that open-earedness is neither predicted by age, nor gender, personality, migration background, or socio-economic status. Our alternative hypotheses are:

- H_1: Open-earedness can be distinguished as a single factor via exploratory factor analyses.
- H_2: Older children are less open-eared than younger ones.
- H_3: Boys are less open-eared than girls across all points of measurement.
- H_4: Children with low values at the personality dimension "openness for experience" are less open-eared than children with high values at that dimension.
- H_5: Children with migration background are less open-eared than children without migration background.
- H_6: Children with low socio-economic status are less open-eared than children with high socio-economic status.

III. METHOD

A. Research Design and Participants

We conduct a cohort-design study with four points of measurement ($t_{1\text{-}4}$) between grade one and four of primary school. Pupils and their parents are questioned. By now, three points of data collection are completed (see Table 1).

Table 1: Participants

t_1 (End of 1st Grade)		t_2 (End of 2nd Grade)		t_3 (End of 3rd Grade)	
Children	Parents	Children	Parents	Children	Parents
1143 (617)	914	1223	745	1175	722

n=455 Children took part in all three parts of the sampling, ~52% Girls.

The sample was composed out of groups based on classes from 20 primary schools from North-Rhine-Westphalia and 13 schools from the City of Hamburg. Due to slightly different school programs the sample from Hamburg was not questioned for their musical preference at t_1. For the construction of this basal model all sub-groups are included.

B. Instruments

In addition to the presentation of a sound questionnaire to the children, which provided preference ratings as the dependent variable (see *Musical Examples and Procedure*), children and/or parents answered several standardized questionnaires covering information on age, gender, children's personality ratings, socio-economic status, migration background, cognitive competencies, parental support, and children's participation in learning musical instruments. Complementary interviews were conducted.

1) Definition of Independent Variables.

The independent variables personality, socio-economic status, migration background, and cognitive competencies were defined in the following way:

- *Personality*: Parents answered for their child the Five-Factor-Questionnaire for Children (Fünf-Faktoren-Fragebogen für Kinder, FFFK) by Asendorpf (1998), a tool for the external inquiry of the Big-Five dimensions of personality. The Cronbach's alphas in this study are satisfactory and range from α=.873 (extraversion, t_1) to α=.744 (agreeableness, t_1), and are on the same level as the norm. Though issues of multidimensionality were observed for single items in exploratory factor analyses (EFA) and exploratory structural equation models (ESEM) analyses, it was decided to take this measure into account nonetheless referring to Asendorpf and von Aken (2002). The factors "openness for experience" (1-5 Likert Scale, t_1, AM=4.05 SD=.550) and "extraversion" (t_1, AM=4.06 SD=.657) were taken into account as possible

Proceedings of the 12th International Conference on Music Perception and Cognition and the 8th Triennial Conference of the European Society for the Cognitive Sciences of Music, July 23-28, 2012, Thessaloniki, Greece
Cambouropoulos E., Tsougras C., Mavromatis P., Pastiadis K. (Editors)

regressors. The external rating of the children's personality are stable across all t ($p \leq .001$ in t-tests), therefore a possible change in personality cannot be taken into account.

- *Socio-economic status*: The socio-economic status was derived by Item-Response-Theory-Scaling (IRT) based on a tool for accessing the status for school-children in Hamburg (KESS or LAU 1) including socio-demographic items, like number of owned books in household, yearly income per household, international standards of classification of occupation and education indices (ISCO-88 & ISCED), and belongings (e.g. lawnmower or second car). Some items were added concerning parental behaviour and cultural participation for a stronger inclusion of children's socio-cultural capital. Though already integrated into this index some items were additionally included as separate indicators due to their potential predictive power. The resulting covariance was taken into account.
- *Migration background*: The migration background was indicated by the parents' answers to the questions whether one or both parents were born in Germany.
- *Test for cognitive competencies*: The test for cognitive competencies (KFT 1-3, Heller & Geisler, 1983) was used to measure deductive and numerical thinking capabilities. The results were IRT-scaled. For analyses on musical self-concept and cognitive competencies in our research network see: Nonte & Schwippert (2012).

The descriptive values of the observed variables are summarized in Table 2.

Table 2: Descriptives of the manifest and IRT-scaled covariates (t_1)

Construct	Values	AM	SD
Parental migration background	1 'no parental migration background' to 3 'both parents not born in Germany'	1.40	.718
Socio-economic index	*wle-score*	.077	.387
Books in the household	1 '1-10' to 5 'more than 200'	3.70	1.21
Households income per year	1 'less than 20.000€/year' to 6 'more than 60.000€/year'	3.69	1.90
ISCO-88 Level of the father	1 'ISCO-88 Level 1' to 3'ISCO-88 Level 4'	2.08	.786
ISCO-88 Level of the mother		2.08	.596
ISCED Level of the father	1 'ISCED Level 1' to 4 ISCED Level 5-6'	3.12	.870
ISCED Level of the mother		2.96	.840
KFT 1-3 subscale deductive thinking	*wle-score*	-.030	1.04
KFT 1-3 subscale numerical thinking		-.064	1.40

All predictor variables were taken from t_1. An imputation with data from t_2 will be conducted in one of our future steps concerning longitudinal analyses.

2) *Music Examples and Procedure.*

The sound questionnaire was composed of 8 musical excerpts with durations of 30 seconds each (see Table 3). Four of them were adopted form previous studies to increase comparability of the results. They were chosen to represent the music genres "classic", "contemporary", and "cross-over". Four further examples were included to represent music from different countries, namely Africa, Russia, Turkey, and China. The original sound questionnaire comprises additional music examples for "pop" and "classic" which were especially composed for the study to control for specific musical parameters. But in this paper we will concentrate on investigating the above mentioned music genres only, as they are known from previous research to generate increasing dislike during the course of primary school. Thus, they are the focus point of our investigation of the decline in open-earedeness.

The study was conducted during regular school hours within class. During the test children listened to each musical excerpt via CD-Player and indicated their preference for each example on a 5-point iconic rating scale (smileys), treated as ordered categoricals (ordinal). They are coded as '1'–'Want to hear more often' to '5'–'Don't want to hear'. The *AM* range from 1,75 (*SD*=1,25: African example at t_1) to 3,12 (*SD*=1,49: cross-over example at t_1) with most items being positively skewed and all normally distributed (see paragraph Results: Factor Modelling.).

3) *Statistical Analyses.*

For the analyses of inference techniques of structural equation modeling (SEM) were chosen to take into account multiple latent dimensions, the theory of measurement errors, and especially the easy inclusion of the clustered samples (Reinecke, 2005), correcting the standard error for children per school level. Due to partial non-normal distribution (observed variable migration background) and categorical data in the observed variables correlative / regressive analyses in the SEM environment will be concluded using a robust weighted least squares estimator (WLSMV). The analyses were calculated using SPSS 19 and MPlus 6.11.

IV. RESULTS

The presentation of results will follow our alternative hypotheses and demonstrate the path of our analysing strategy.

A. Factor Modelling

With regard to H_1, we calculated prime EFAs. Results support a distinction between two factors comparable to Louven (2011), who interpreted them as "classic" and "ethnic/avant-garde" music (Table 3). We deemed this two-factor solution as the statistically most fitting across-the-board by comparing all solutions fit indices for all t after corrections which are described in the following passage.

Proceedings of the 12th International Conference on Music Perception and Cognition and the 8th Triennial Conference of the European Society for the Cognitive Sciences of Music, July 23-28, 2012, Thessaloniki, Greece
Cambouropoulos E., Tsougras C., Mavromatis P., Pastiadis K. (Editors)

Table 3: Exemplary rotated communality matrix of an exploratory factor analyses (t_3, all items, n_3=1132)

Musical excerpt	*classic*	*ethnic/ avant-garde*
Sinfonie Nr. 4 op. 90, 1. Satz (Felix Mendelssohn-Bartholdy)	*0.799*	*-0.078*
Air (David Garrett)	*0.681*	*0.007*
"Gavotte I" from Orchestral Suite No. 3, D major (J.S. Bach)	*0.801*	*0.007*
Russia: Smyglyanka (Samovar Russian Folk Ensemble)	*0.412*	*0.315*
3. Sinfonie, 3. Satz „Beschwörungstanz" (H. W. Henze)	*-0.061*	*0.587*
Ümmü (Sümer Ezgü)	*-0.006*	*0.710*
Yu Fu Rong (Chinese Ensemble of Movie Music and Folk Music)	*0.068*	*0.652*
Upepu (Magi Shamba)	*0.007*	*0.606*

Delta Parametersation, Oblique-Geomin Rotation, WLSMV Estimator; RMSEA=.082, TLI=.946.

The communalities of the EFAs at t_1 & t_2 can be interpreted alike. The solution is not satisfactory due to the inconclusive attribution of the Russian folk music. The Russian folk music can be contributed to the "classic" factor at t_1 and to the "ethnic/avant-garde" factor at t_2 a little more clearly, but in either matrix the communality loadings are below 0.550. Therefore it is excluded from further analyses.

After excluding the Russian folk music example the model fits for the two-factor solution ranged from: $RMSEA_{t3}$=.069 TLI_{t3}=.964 (n_3=1132) to $RMSEA_{t1}$=.029 TLI_{t1}=.994 (n_1=599) and can be deemed satisfactory.

Because the extracted factors were to be used in a multiple indicators – multiple causes (MIMIC-) SEM (Joreskog & Goldberger, 1975), those were utilised for creating a confirmatory factor model (CFA) as basis of further analyses.

The factors were reproduced in a CFA for each t. The model of t_3 is shown in Figure 1. The absolute (RMSEA=.043) and the incremental (TLI=.987) model fit indices are absolutly satisfying. It has to be stressed out, that the high correlation (r=.595) between the factors "classic" and "ethnic/avant-garde" imply the possibility of a second order latent. This possibly points to a more generalized construct of open-earedness than it could be operationalised by our data by now. The variance of the residual variance of the avant-garde example (σ^2_{res}=.71) remains high.

The Fit-indices of the CFA for the other t are: $RMSEA_{t1}$=.045, TLI_{t1}=.985 (n_1=599) and $RMSEA_{t2}$=0.042, TLI_{t2}=0.989 (n_2=1125).

A restricted and a one-factor-model were tested and found significantly worse than the two-factor model. The model was tested for factorial invariance and weak metric invariance and found valid, though the fit of gender-specific models of boys was only acceptable, while the girls' models were good.

We assume this model to be reliable and convergent as well as discriminative valid enough for the intended purpose. The indicators' and the latents' distributions can overall be treated as normal (Kolmogorow-Smrinow-tests and Shapiro-Wilk-tests for all t_x p<.01), though kurtosis and skewness values reach factors slightly bigger than 1 in single indicator items. To sum it up, the EFA stressed that open-earedness cannot be distinguished as a single factor via factor analyses ($\neq H_1$) in our data. The two-factorial solution for the preference of unconventional music appears adequate.

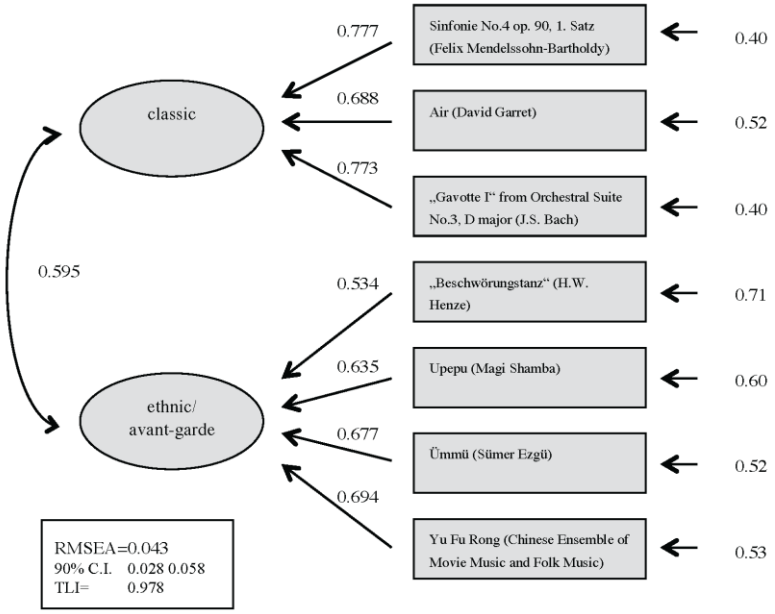


Figure 1: Confirmatory factor analysis, (t_3, n=1132)

Proceedings of the 12th International Conference on Music Perception and Cognition and the 8th Triennial Conference of the European Society for the Cognitive Sciences of Music, July 23-28, 2012, Thessaloniki, Greece
Cambouropoulos E., Tsougras C., Mavromatis P., Pastiadis K. (Editors)

And finally as for H_6, we could not find any relationships between children's music preference and their socio-economic status ($\neq H_6$).

Table 4: MIMIC regressive model (n_{L3}=247)

Regression	classic t_3 *stand β*	ethnic/avant-garde t_3 *stand β*
Factor at t2	.446***	.350***
Gender	.301***	-.053
SES	.046	-.029
Income per Year	.076	.164
Migration background	-.178*	-.012
Books in household	.015	-.016
ISCED of the Father	-.041	-.007
ISCED of the Mother	-.027	.087
ICSO-88 of the Father	-.068	-.069
ISCO-88 of the Mother	-.047	.043
Instrumental tuition	-.015	-.085
KFT 1-3 subscale deductive thinking	-.055	-.059
KFT 1-3 subsclae numerical thinking	-.137	-.127
FFFK Dimension Openness	.166*	-.007
FFFK Dimension Extraversion	-.236*	-.115*
Correlation	*ethnic t_3 with classic t_3* .291	
R Square	.383	.210
Fit Indices		
chi/df/p	419,909/373/0.047	
RMSEA	0.018	
TLI	0.960	

Delta Parameterisation, WLSMV Estimator, Clustered per School (complex), ***=p≤.001, **=p≤.01, *=p≤.05

After reviewing the latent factors r^2 values in this model it has to be stated, that the predictors taken into account so far do not contribute well to the explanation of the children's preference ratings. Additionally the sheer number of possible predictors is making it statistically plausible for an error of the first kind to create a false positive at the 5% level, meaning that 1 out of 20 H_x is a false positive by definition. Therefore, the significant effects should not have been overrated if they supported our hypotheses and shall not be dropped completely before being refined more sophisticated. Though it can be stated that the included variables do not hold major predictive power for our construct of open-earedness, except for gender.

V. CONCLUSION AND PROSPECT

The model is working well with our data. But the low explained variance (r^2) implies that there are other predictors to be taken into account. We found age and gender to be the remaining main predictors for open-earedness ($\neq H_0$). Surprisingly few other factors show predictive power in a generalised model and if so, effect sizes are small.

Especially the result on openness (H_4) appears surprising for the obvious conclusion that "more openness" as a personality trait should bring about "more openness" towards unconventional music. But it has to be taken into account, that the operationalization of openness as a personality trait may also cause more openness or more frankness in expressing negative opinions in situations such as a scientific inquiry, negating known response bias towards the middle and towards positive answers (Schwarz & Sudmann, 1992). It also remains to be analysed, how children's music preference is influenced by other personality scores indicated by the five-factor model of personality (Costa & McCrae, 1992). A serious problem on the measurement of personality through external parental information is the neglect of the children's ups-and-downs in their personality. For it is known that the personality traits are still highly variable in childhood and adolescence. That is why we requested it each *t*. But the traits reflected by the parents were highly invariant over time (t_2-t_1 *and* t_3-t_2, p≤.01 in t-tests). Because of that we treated the children's personality as unchanging, and thus could not live up to more complex facets of personality measurement.

Concerning migration background (H_5) it remains to be analysed whether there are item-level differences on the preference for "ethnic/avant-garde" music that is frequently heard at home and would therefore not be an adequate indicator for open-earedness anymore. Furthermore up to this stage our study did not take into account whether parents with and without migration background differ in their daily musical behaviour and their involvement of their children's musical life.

The differential preference ratings of classic and ethnic music examples between gender groups remain intriguing. Moreover as cognitive capabilities on numerical thinking showed a small significant effect (β=-.254, $p\leq.05$; n_g=140) in this model when calculated for girls alone for the intent of factorial invariance testing. It is to check whether the measurement invariance of group (gender) remains valid for higher levels of statistical invariance especially on latent mean of the factor "classic" music. It is possible that boys already dislike "classic" music at an early age (1st grade), in which case the factor would really measure the same for boys and girls alike. On the other hand it could be possible that even young boys already cling more firmly to the idea of gender-specific music than girls, which would be supported by our preliminary analyses of the interview data. In those interviews girls and boys display strong opinions on gender-specific music alike, but boys were stricter in their rejection of girls' music. In that case the statistical analysis would not come up with an interpretable answer as the measurement itself would be flawed by the boys' bias on their perception of gender-specific music. Explanations for the observed gender differences might be given with regard to theories concerning the development of gender identity (Maccoby, 2000, Ruble et al., 2006). Boys are generally believed to display a stronger fixation on gender stereotypes, which already developed before school age, whereas girls are supposed to be more flexible in this regard. The sensitivity towards atypical gender-specific behaviour increases during primary school. Taken together, children's music preference might not just be explained by socialisation, but could also be seen as an expression of the developing gender identity – and thus (referring to Schäfer & Sedlmeier, 2009) might possibly reflect a specific function for the children.

As mentioned earlier, this study is a part of the refinement of a latent change (LC) / latent growth curve model (LGC). The latent change model will be used to assess group differences furthermore and the latent growth curve modelling will be used to assess the form of change. By now it has to be stated

Proceedings of the 12th International Conference on Music Perception and Cognition and the 8th Triennial Conference of the European Society for the Cognitive Sciences of Music, July 23-28, 2012, Thessaloniki, Greece
Cambouropoulos E., Tsougras C., Mavromatis P., Pastiadis K. (Editors)

that few of the checked covariates have to be taken into account for the further analyses of the latent variables themselves in the LC modelling. Some will be assessed again when analysing the growth curve for some may not contribute direct effects but effects on the development of open-earedness.

Two other steps include the assessment of the preference for different kinds of pop music, that were inquired alongside the less well known kinds of music and the opportunity to approach open-earedness as a possible 2nd order latent variable in a summarizing model.

ACKNOWLEDGMENT

This research is part of the cooperation project SIGrun "Studien zum Instrumentalunterricht an Grundschulen" conducted by the Universities of Bremen and Hamburg, which is funded by the Federal Ministry of Education and Research of Germany within a research program initiated to evaluate the program "JeKi – Jedem Kind ein Instrument".

REFERENCES

Asendorpf, J. D. & van Aken, M. A. G. (2003). Validity of Big Five personality judgments in childhood: A 9 year longitudinal study. *European Journal of Personality, 17*, 1-17.

Baumann, M. P. (Ed.) (1985). *Musik der Türken in Deutschland.* Kassel: Verlag Yvonne Landeck.

Costa, P. T. & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO PI-R) and NEO Five Factor Inventory. Professional Manual*. Odessa, FL: Psychological Assessment Ressources.

Cremades, R., Oswaldo, L., & Lucia, H. (2010). Musical tastes of secondary school students with different cultural backgrounds: A study in the spanish north african city of Melilla. *Musicae Scientiae, 14(1)*, 121–141.

Delsing, M. J. M. H., Bogt, T. F. M. T., Engels, R. C. M. E. & Meeus, W. H. J. (2008). Adolescents' music preferences and personality characteristics. *European Journal of Personality, 22*, 109-130.

Eijck, K. (2001). Social differentiation in musical taste patterns. *Social Forces, 79 (3)*, 1163-1184.

Gembris, H. (2005). Musikalische Präferenzen. In R. Oerter & T. H. Stoffer (Eds.), *Enzyklopädie der Psychologie, Vol. 2, Spezielle Musikpsychologie* (pp. 279-342) Göttingen: Hogreve.

Gembris, H., & Schellberg, G. (2003). Musical preferences of elemantary school children. Paper presented at the *5th Triennial Conference oft he European Society for the Cognitive Sciences of Music* (ESCOM 8.-13.9.2003). Hannover.

Greve, M. (2003). *Die Musik der imaginären Türkei. Musik und Musikleben im Kontext der Migration aus der Türkei in Deutschland.* Stuttgart, Weimar: J. B. Metzler.

Hargreaves, D. J. (1995). Effects of age, gender, and training on musical preferences of British secondary school students. *Journal of Research in Music Education, 43(3),* 242-250.

Hargreaves, D. J., North, A. C., & Tarrant, M. (2006). Musical preference and taste in childhood and adolescence. In G. E. McPherson (Ed.), *The child as musician: A handbook of musical development* (pp. 135-154). New York: Oxford University Press.

Heller, K. & Geisler, H.-J. (1983). KFT 1-3: Kognitiver Fähigkeits-Test. Weinheim: Beltz.

Henninger, J. C. (1999). Ethnically diverse sixth graders' preferences for music of different cultures. *Texas Music Education Research*, 37-43.

Jöreskog, K. G. & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association*, 70(351), 631-639.

Kopiez, R., & Lehmann, M. (2008). The 'open-earedness' hypothesis and the development of age-related aesthetic reactions to music in elementary school children. *British Journal of Music Education, 25(2)*, 121-138.

MPLUS (Version 6.11). [Computer Software]. Los Angeles, CA: Muthén & Muthén.

LeBlanc, A. (1991). Some unanswered questions in music preference research. *Contribution to Music Education, 18*, 66-73.

LeBlanc, A., Sims, W. L., Siivola, C., & Obert, M. (1996). Music style preferences of different age listeners. *Journal of Research in Music Education, 44(1)*, 49-59.

Louven, C. (2011). Mehrjähriges Klassenmusizieren und seine Auswirkungen auf die „Offenohrigkeit" bei Grundschulkindern. Eine Langzeitstudie. *Diskussion Musikpädagogik 50(11)*, 48-59.

Maccoby, E. (2000). *Psychologie der Geschlechter. Sexuelle Identität in den verschiedenen Lebensphasen.* Translated by E. Vorspohl. Stuttgart: Klett-Cotta.

McCrary, J. (1993). Effects of listeners' and performers' race on music preferences. *Journal of Research in Music Education, 41(3)*, 200-211.

Nonte, S. & Schwippert, K. (2012). Musikalische und sportliche Profile an Grundschulen. *Beiträge empirische Musikpädagogik*, 3(1).

Peterson, R. A. (1992). Understanding audience segmentation: From elite and mass to omnivore and univore. *Poetics, 21*, 243-258.

Rawlings, D., & Ciancarelli, V. (1997). Music preference and the five-factor model of the NEO Personality Inventory. *Psychology of Music, 25,* 120-132.

Reinecke, J. (2005), Strukturgleichungsmodelle in den Sozialwissenschaften. München: Oldenbourg.

Rentfrow, P. J., Goldberg, L. R. & Levitin, D. J. (2011). The structure model of musical preferences: A five-factor model. *Journal of Personality and Social Psychology, 100(6)*, 1139-1157.

Ruble, D. N., Martin, C. L. & Berenbaum, S. A. (2006). *Gender development.* In W. Damon & R. M. Lerner (Eds.), Handbook of Child Psychology, Vol. 3, 6th ed. (pp. 858-932). Hoboken: Wiley.

Sakai, W. (2011). Music preferences and family language background: A computer-supported study of children's listening behavior in the context of migration. *Journal of Research in Music Education, 59(2)*, 174-195.

Schäfer, T., & Sedlmeier, P. (2009). From the functions of music to music preference. *Psychology of Music, 37*, 279-300.

Schwarz, N. & Sudman, S. (Eds.) (1992). Context effects in social and psychological research. New York: Springer..

Teo, T., Hargreaves, D. J., & Lee, J. (2008). Musical preference, identification, and familiarity: A multicultural comparison of secondary students from Singapore and the United Kingdom. *Journal of Research in Music Education, 56(1)*, 18–32.

Wurm, M. (2006). *Musik in der Migration. Beobachtungen zur kulturellen Artikulation türkischer Jugendlicher in Deutschland.* Bielefeld: transcript.

BEITRAG 2

Erschienen in (Zitierweise):

Schurig, M. & Busch, V. (2014). Entwicklung der Musikpräferenz von Grundschulkindern. Individuelle, soziale und musikbezogene Einflüsse. In: A. Lehmann-Wermser, V. Busch, K. Schwippert & S. Nonte (Hrsg.), *Mit Mikrofon und Fragebogen in die Grundschule. Jedem Kind ein Instrument (JeKi) – eine empirische Längsschnittstudie zum Instrumentalunterricht* (S. 63-96). Münster: Waxmann.

Relevanz:

Analysiert wird die Entwicklung der Offenheit von Grundschülern für unkonventionelle Musikstile in Abhängigkeit zu verschiedenen Hintergrundmerkmalen. Der Beitrag wird herangeführt, um differenzierende Betrachtungsweisen eines theoretischen Konstrukts auf der Basis eines einzelnen Datensatzes nachvollziehbar zu machen. Aufbauend auf den Ergebnissen der Beiträge 1 und 3 sowie auf Basis einer erweiterten Stichprobe und Datengrundlage wurde eine Latente Profilanalyse durchgeführt, bei der die Indikatoren zweier theoretisch verknüpfter latenter Variablen herangezogen werden, um die Entwicklung eines theoretischen Überkonstrukts über die Zeit beobachtbar zu machen. Dies konnte im Beitrag 1 auf der Basis eines FA Ansatzes nicht geleistet werden. Es kann nachgewiesen werden, dass diese latente Struktur über vier Messzeitpunkte hinweg stabil ist. In der Folge werden die Gruppen interpretiert und Hintergrundmerkmale werden zur inhaltlichen Charakterisierung der Gruppen herangezogen. Zuletzt werden Transitionswahrscheinlichkeiten innerhalb der latenten Profile zwischen den Messzeitpunkten bestimmt, um die Entwicklungen bei der Zusammensetzung der abgeleiteten Klassen, in Abhängigkeit zu den Prädiktoren, zu beobachten. Zudem wird ein Ausblick auf die Ergänzung eines quantiativen Forschungsansatzes durch eine qualitative Folgestudie gegeben.

Michael Schurig und Veronika Busch

IV. Zur Struktur und Beeinflussung musikalischer Präferenzurteile im Grundschulalter. Ergebnisse aus dem Teilprojekt Präferenz

1 Theoretischer Hintergrund

Musikpräferenz wird in dem vorliegenden Beitrag als eine Facette kulturellen Verhaltens verstanden, die vielfach als Ausdruck von sozialer Zugehörigkeit und individueller Identität genutzt wird. Bei dieser Nutzung von Musikpräferenz wird auf frühe musikalische Erfahrungen aufgebaut, die auch im schulischen Musikunterricht ermöglicht werden. Somit ist die Erforschung der Entwicklung kindlicher Musikpräferenz auch für die empirische Bildungsforschung von Interesse.

Unter Musikpräferenz wird generell ein Gefallensurteil verstanden, das sich auf verschiedene Aspekte von Musik, zum Beispiel auf ein konkretes Musikstück, einen Musikstil oder auch einen bestimmten Musiker beziehen kann. In der Fachliteratur lassen sich sehr unterschiedliche und zum Teil konträre Definitionen finden. In der vorliegenden Studie wird Musikpräferenz im Sinne von Behne (1993; vgl. Reinhardt & Rötter, 2012, S. 133) als ein aktuelles Urteil in einer konkreten Situation verstanden und somit von dem längerfristigeren Musikgeschmack eines Menschen abgegrenzt (vgl. Gembris, 2005). Nach Behne (1975; vgl. Bunte, 2013) verweist aktuelles musikbezogenes Präferenzverhalten auf zugrunde liegende Konzepte, die als „Summe von Vorstellungen, Einstellungen, Informationen, Vorurteilen etc., die ein Individuum hinsichtlich eines mehr oder weniger umgrenzten musikalischen Objektes besitzt“ (Behne, 1975, S. 36), beschrieben werden. Diese werden wiederum von den Erfahrungen gespeist, die ein Mensch im Umgang mit Musik sammelt.

Musikbezogene Erfahrungen werden auch im Schulkontext geboten, wobei die JeKi-Programme vermutlich eine besondere Erfahrungsintensität hervorrufen und somit auch die Entwicklung des kindlichen Präferenzverhaltens beeinflussen können. Eine mögliche musikpädagogische Motivation zur Einflussnahme auf die Präferenzentwicklung lässt sich aus der vielfach bestätigten Beobachtung ableiten, dass Kinder im Verlauf der Grundschulzeit ihre anfängliche Offenheit gegenüber einer Vielzahl an musikalischen Stilen verlieren und zum Beginn der Pubertät vor allem die aktuelle populäre Musik der Charts präferieren, was wiederum als Einschränkung in der Teilhabe an ästhetischen Ausdrucksformen begriffen werden kann. Die umfangreichen Forschungsaktivitäten hierzu werden mit dem Begriff Offenohrigkeitsforschung zusammengefasst und gehen auf eine Hypothese von Hargreaves zurück, die besagt: „younger children may be more ‚open-eared’ to forms of music regarded by adults as unconventional“ (Hargreaves, 1982, S. 51). Hierbei werden unter „unconventional“ beispielsweise Formen avantgardistischer, aleatorischer und elektronischer Musik (Hargreaves et al. 2006, S. 144) sowie Formen von „classical music“ und „ethnic music“ verstanden (u.a. Hargreaves et al., 1995; 2006; Kopiez & Lehmann, 2008; Louven, 2011). Der bedeutende Einfluss des Alters auf die Musikpräferenz wird von verschiedenen Studien gestützt (u.a. Hargreaves et al., 2006; LeBlanc et al., 1996), doch sind die Ergebnisse nicht eindeutig: Während einige Autoren bereits im Laufe der Grundschulzeit (Gembris & Schellberg, 2007) oder bereits in der Vorschulzeit (Hargreaves, 1987) von einer Abnahme an Offenohrigkeit ausgehen, sehen beispielsweise Kopiez und Lehmann (2008) die gesamte Grundschulzeit als Periode der Offenohrigkeit an.

In der deutschsprachigen Literatur kommt der Studie von Gembris und Schellberg (2003) eine Vorreiterrolle bei der empirischen Beschreibung des Konstruktes Offenohrigkeit zu. Die Autoren

bewerten das „Verschwinden“ von Offenohrigkeit im Verlauf der Grundschulzeit, also die zunehmende Ablehnung unter anderem von Musik mit ‚klassischen’ Stilmerkmalen, als negativ und sehen in gezielten musikpädagogischen Unterrichtsentwürfen Chancen, diesem entgegen zu wirken (Schellberg, 2006). Diese selten hinterfragte Zielsetzung wird auch in den Richtlinien und Lehrplänen für die Grundschule in Nordrhein-Westfalen (NRW) deutlich. Nach diesen Richtlinien führe der Musikunterricht die Grundschulkinder „zu einem offenen und aktiven Umgang mit Musik“ hin und sei die „[hörende] Auseinandersetzung mit vielfältiger Musik (Popmusik und Jazz, Klassische und Neue Musik sowie Musik anderer Länder und Kulturen)“ von entscheidender Bedeutung, um Aufgeschlossenheit und Neugierde zu erhalten (Ministerium für Schule und Weiterbildung NRW, 2008, S. 87-89). So gehe es im Musikunterricht darum, „für vielfältige Musik offen zu werden“ (Ministerium für Schule und Weiterbildung NRW, 2008, S. 93).

Auch in den JeKi-Programmstandards von NRW wird „stilistische Offenheit gegenüber allen Musikstilen“ [58] explizit als Inhalt formuliert. Nach Behne (1975; 1987) sollte sich dies auch in musikbezogenen Präferenzäußerungen der Kinder ablesen lassen. Diese These zum Zusammenhang von musikalischer Erfahrung und Präferenz lässt sich auch empirisch bestätigen. So haben bereits Hargreaves et al. (1995) in einer Studie mit Jugendlichen festgestellt, dass „level of [musical] training and preference for ‚serious’ style categories“ (Hargreaves et al., 1995, S. 248) positiv korrelieren, wobei die Autoren unter „‚serious’ style categories“ vor allem Musik aus den Bereichen Klassik und Oper, aber auch Jazz und Folk fassen. Louven (2011) hat Grundschulkinder während ihrer ersten vier Schuljahre nach Präferenzurteilen für vorgespielte Musikstücke befragt. Ein Teil der Kinder befand sich in sogenannten Streicherklassen, die während der ersten zwei Schuljahre wöchentlichen zwei Stunden schulischen Instrumentalunterricht im Klassenverbund erhielten und in den letzten beiden Grundschuljahren an Streicher-AGs teilnehmen konnten. Louven konnte feststellen, dass Kinder aus Streicherklassen generell positivere Bewertungen für die Vielfalt der präsentierten Musikstücke abgaben. Insbesondere profitierten von dem Instrumentalunterricht jedoch die Bewertungen der Musikstücke mit ‚klassischen’ Stilmerkmalen, was Louven dadurch erklärt, dass sich Kinder aus Streicherklassen „von der Substanz des musikalischen Materials her mit ‚Klassik’ beschäftigten“ (Louven, 2011, S. 58). Zudem würden sie sich über den vertraut gewordenen Streicherklang mit der „Welt der klassischen Musik“ identifizieren können (Louven, 2011, S. 58). Auch die Studie von Schellberg (2006) verdeutlicht, wie durch die intensive Beschäftigung mit einer Opernarie im Rahmen des schulischen Musikunterrichts die Präferenzurteile für die konkrete Arie deutlich positiver ausfielen, als dies für Musik im Stile des Belcanto zu erwarten sei. Auf einen Zusammenhang von Vertrautheit und Gefallen wird in der musikpsychologischen Präferenzforschung immer wieder verwiesen (vgl. Reinhardt & Rötter, 2012, S. 135 ff.), was die Vermutung bekräftigt, dass musikbezogene Aktivitäten der typischen altersabhängigen Abnahme an Offenohrigkeit im Bereich der „serious“ Stilkategorien entgegenwirken könne (vgl. Hargreaves et al., 1995; Louven, 2011; Schellberg, 2006). Diese Beobachtung ließe sich durchaus als Wert an sich deuten, da sich in der Erweiterung beziehungsweise Aufrechterhaltung eines breit ausgerichteten Präferenzspektrums möglicherweise eine Offenheit widerspiegelt, die den Kindern den Zugang zu verschiedenen Aspekten des musikalisch-kulturellen Lebens erleichtern könnte. Empirisch belegt ist solch ein Zusammenhang von musikalischer Offenheit und genereller kultureller Offenheit bislang jedoch nicht. Andererseits könnte das „Verschwinden“ (Schellberg & Gembris, 2007) von Offenohrigkeit auch als Zeichen einer sich ausbildenden Urteilsfähigkeit und individuellen Musikpräferenz gedeutet werden, was ebenfalls als wertvoll und förderungswürdig erachtet werden kann.

[58] https://www.jedemkind.de/programm/mediathek/pdf/120326_programmstandards_2011_2012.pdf [06.06.2014]

Die erwähnten musikpädagogischen Erwartungen sollten zudem im Kontext der musikbezogenen Transferforschung betrachtet werden, die spätestens seit dem sogenannten „Mozart-Effekt“ (Rauscher et al., 1995) öffentliche Aufmerksamkeit erfahren und eine kontroverse Fachdiskussion generiert hat. Wie Nonte und Schwippert in ihrem Beitrag zum Teilprojekt Transfer in diesem Band (vgl. Kap. III.) darstellen, richten sich pädagogische Erwartungen an musikbezogene Aktivitäten im Wesentlichen auf Transferwirkungen in den kognitiven, aber auch den sozialen Bereich (vgl. BMBF, 2006; 2009). Aus dem Bildungsbericht von 2012 (Autorengruppe Bildungsberichterstattung, 2012) lassen sich zudem Erwartungen auf individualpsychologische Transferwirkungen ablesen:

> Kulturelle/musisch-ästhetische Bildung als integraler Bestandteil individueller und sozialer Identitätsentwicklung ermöglicht die Entwicklung künstlerischer Wahrnehmungs-, Darstellungs-, Gestaltungs- und Ausdruckformen, vor allem über eigene ästhetische Praxis, die in ganz unterschiedlichen sozialen Kontexten ausgeübt wird und so zu spezifischen Gemeinschaftserfahrungen führen kann. (Autorengruppe Bildungsberichterstattung, 2012, S.160)

Bisherige empirische Studien mit Jugendlichen und Erwachsenen lassen die Annahme solch individualpsychologischer Transferwirkungen auch für Kinder berechtigt erscheinen. Wesentliche Aspekte solch einer Argumentation sind zum einen die bereits beschriebene Beeinflussung musikalischer Präferenz durch musikbezogene Erfahrungen sowie zum anderen die Bedeutung von Musikpräferenzäußerungen zur Ausbildung und Darstellung der eigenen Identität. In der Forschung zur Musikpräferenz werden verschiedene Modelle zur Strukturierung musikalischer Präferenzäußerungen für eine Vielzahl musikalischer Beispiele diskutiert. Für die Frage der Identitätsbildung ist das von Rentfrow et al. (2011) vorgeschlagene Five-Factor-Model MUSIC („Mellow”, „Unpretentious”, „Sophisticated”, „Intense” und „Contemporary”) bedeutsam, da hierbei nach den psychischen Wirkungen der jeweiligen Musikstücke differenziert wird. Rentfrow et al. (2011) argumentieren für multiple Einflüsse (wie psychologische Disposition, soziale Interaktion, Umgang mit populären Medien, kultureller Trend) auf musikalische Präferenz und verweisen auf die Eigentümlichkeit, dass wir zwar um die enorme Bedeutung von Musik für Menschen wissen, „[c]uriously, however, very little is known about why music is so important“ (Rentfrow et al., 2011, S. 1155). Aus einer Studie von Schäfer und Sedlmeier (2009) lässt sich ableiten, dass Musik vielfältige bedeutende Funktionen für Individuen übernehmen kann. Der Nutzung von Musik zur Darstellung der eigenen Identität kommt hierbei eine besondere Stellung zu: „the most important reasons why people like their music are its capability to express their identity and their values and its ability to bring people together“ (Schäfer & Sedlmeier, 2009, S. 297).

Als ein Entwicklungsabschnitt, in dem Musik in besonderer Weise Funktionen bei der individuellen Identitätsbildung erfüllen kann, gilt die Pubertät (vgl. Gembris, 2005; Behne, 1986; 1997). Die Ausbildung und Darstellung von Identität durch Musik kann innerhalb der Offenohrigkeitsforschung unter anderem im Zusammenhang mit dem Geschlecht, der Persönlichkeit, dem Migrationshintergrund sowie dem sozialen Status (ökonomisch und kulturell) betrachtet werden. Hinsichtlich des Geschlechts wird zumeist konstatiert, dass Mädchen im Vergleich zu Jungen offenohriger seien (Hargreaves et al., 1995; Gembris & Schellberg, 2007; Busch et al., 2009). In der Studie von Kopiez und Lehmann (2008) zeigten sich hingegen keine bedeutsamen geschlechtsspezifischen Effekte, was sich aber vermutlich aufgrund ihrer Einstufung von ‚klassischer’ Musik als konventionelle Musik relativieren lässt. Erklärungsansätze für Geschlechtsunterschiede im musikalischen Präferenzverhalten sind vermutlich in einer geschlechtsspezifischen musikalischen Sozialisation zu finden (Busch et al., 2009). Hierfür sprechen auch die Befunde von Wilke (2012), nach denen Jungen bereits im Grundschulalter Gangsta Rap zur Auseinandersetzung mit und zur Inszenierung von Männlichkeit nutzen (Wilke, 2012, S. 241 ff.). Hinsichtlich der Persönlichkeitsstruktur (Big Five nach Costa & McCrae, 1992) liegt für Jugendliche und Erwachsene bereits eine Reihe von Studien vor, die auf Zusammenhänge mit musikbezogenen

Präferenzen veweisen (Delsing et al., 2008; Rawlings & Ciancarelli, 1997). So korreliert das Persönlichkeitsmerkmal ‚Offenheit für Erfahrungen' positiv mit einer stilistisch breit angelegten Musikpräferenz, während das Persönlichkeitsmerkmal ‚Extraversion' positive Zusammenhänge zu populären sowie zu in hohem Maße erregenden Musikformen ausweisen (Rawlings & Ciancarelli, 1997). Entsprechend sehen Langmeyer et al. (2012, S. 120) in diesen beiden Persönlichkeitsmerkmalen die besten Prädiktoren für Musikpräferenz. Hinsichtlich der Musikpräferenz von Grundschulkindern kann sich jedoch nicht auf empirische Befunde gestützt werden. Dennoch lassen die berichteten Ergebnisse bezüglich höherer Altersstufen vermuten, dass beispielsweise der Persönlichkeitsfaktor ‚Offenheit für Erfahrungen' auch bereits mit der kindlichen Musikpräferenz korreliert. Möglicherweise lässt sich dieses Persönlichkeitsmerkmal mit dem vielfach beschriebenen kritischen Zeitfenster in der kindlichen Entwicklung, in dem junge Kinder mit erhöhter Sensibilität auf musikbezogene Anregungen hinsichtlich ihrer Präferenzentwicklung reagieren sollen (u.a. Gembris, 2005; Hargreaves et al., 2006), in Verbindung bringen und entsprechend die kindliche Offenohrigkeit als Indikator für diese Zeitfenster deuten. Als ein weiterer Aspekt bei der Identitätsbildung kann der Migrationshintergrund angesehen werden. Diesbezüglich liegen empirische Hinweise vor, wonach Jugendliche und Erwachsene mit Migrationshintergrund die Musik ihres jeweiligen Herkunftslandes präferieren (Sakai, 2011; Cremades et al., 2010; Henninger, 1999; Teo et al., 2008). Zudem wird angenommen, dass die Loslösung vom elterlichen Musikgeschmack aufgrund der häufig bedeutsamen Rolle von Musik im familiären Alltag von Migrationsfamilien verzögert ist (Baumann, 1985; Greve, 2003; Wurm, 2006). Die Ausbildung und Darstellung von Identität muss auch im Zusammenhang mit dem Sozialstatus diskutiert werden und dabei muss auf Bourdieu verwiesen werden. Dieser beschreibt mit dem Habitus-Konzept die soziale Gebundenheit der Gesten und Handlungen eines Menschen, wobei musikalische Vorlieben ausdrücklich als einen bedeutenden Ausdruck eines bestimmtes Habitus und damit einer gesellschaftlichen „Klasse" angesehen werden. Entsprechend ist Bourdieu überzeugt, „daß man von den musikalischen Präferenzen, die jemand hat (oder noch einfacher von den Radiosendern, die er hört), genau so unfehlbar auf die Zugehörigkeit zu einer sozialen Klasse" schließen könne (Bourdieu, 1993 [1980], S. 150). Peterson und Simkus (1992) sowie Chan und Goldthorpe (2007) haben zwar die von Bourdieu postulierte eindeutige Zuordnung von Sozialstatus und bevorzugtem Musikstil nicht bestätigen können, wohl aber einen Zusammenhang von Sozialstatus und stilistischer Breite der präferierten Musik. Demnach steht ein höherer sozialer Status mit der Wertschätzung einer größeren Bandbreite an musikalischen Genres in Verbindung („omnivores"), während ein niedrigerer sozialer Status eher mit der Beschränkung auf einen oder wenige musikalische Stile einhergeht („univores"; Peterson, 1992; Peterson & Simkus, 1992; Chan & Goldthorpe, 2007; Van Eijck, 2001). Eine Beeinflussung von Musikpräferenz durch die Zugehörigkeit zu einer sozialen Schicht scheint somit naheliegend, wenn auch nicht im engen Sinne von Bourdieu (vgl. Lenz, 2013, S. 176 f.). Bourdieu argumentiert zudem (vgl. auch Kleinen, 2011), dass vor allem „[unterschiedliche] Arten des Erwerbs der musikalischen Bildung, unterschiedliche Formen der allerersten Musikerfahrungen" (Bourdieu, 1993, S. 150), die wiederum nach sozialem Status differieren, wesentlich den längerfristigen Musikgeschmack eines Menschen prägen. Eine bedeutende Rolle bei der Aneignung habitueller Verhaltensweisen komme nach Bourdieu (1983) den körperlich gebundenen Erfahrungen zu (vgl. Lenz, 2013, S. 169). Aus der musikpsychologischen Forschung legen etliche Studien ebenfalls eine besondere Körpergebundenheit musikalischer Erfahrung nahe (u.a. Iyer, 2002; Busch, 2005; Lopez Cano, 2003), so dass musikbezogenen Aktivitäten in der frühen Bildung ein nicht zu unterschätzender Einfluss auf die Ausbildung sozialer Verhaltensmuster zugesprochen werden sollte. Hierin liegen somit auch Chancen für die Musikpädagogik: Obgleich der jeweils schichtspezfische gesellschaftliche Habitus seit der Geburt vom Umfeld erfahren und erlernt werde, bestehe nach Bourdieu ein gewisser Handlungsspielraum zur Entfaltung individuellen Verhaltens. Wenn nun wie in Hamburg mit dem JeKi-Programm in besonderer Weise Kinder von bildungsfernen Schichten erreicht werden (vgl. Nonte & Schwippert, 2012), können die Erfahrungen mit musikalischen Aktivitäten, die

üblicherweise nicht zum schichtspezifischen Habitus dieser Kinder gehören, möglicherweise dazu führen, die Grenzen ihrer individuellen Handlungsspielräume zu erweiteren, um alternative oder einfache zusätzliche Verhaltensweisen zu erproben. Dies sollte sich ebenfalls in einer weniger stark ausgeprägten Ablehnung von unkonventioneller Musik im Sinne Hargreaves' (1982) niederschlagen, denn die vermeintlich unkonventionelle Musik wäre den Kindern durch den schulischen Unterricht vertrauter und Teil ihres Erfahrungsschatzes, auf den sie dann für unterschiedliche Bedürfnisse zurückgreifen können.

Musikpräferenz wird in der vorliegenden Studie somit zusammenfassend als ein Teil generellen kulturellen Verhaltens angesehen (vgl. Kap. V. Kulturelle Teilhabe, in diesem Band), der unter anderem die Fähigkeit umfasst, die eigenen musikalischen Vorlieben zu benennen, sowie das Wissen, welche Musik in welchem Moment bevorzugt wird, und wie Musik zur Übernahme bestimmter individueller und sozialer Funktionen genutzt werden kann. So kann Musik beispielsweise gezielt zur Stimmungsaufhellung, zum Emotionsausdruck, zur Ausbildung und Darstellung der eigenen Identität, zur Abgrenzung von sozialen Gruppen sowie zur Strukturierung des Alltags eingesetzt werden. All diese Aktivitäten bedürfen der Fähigkeit, Musik und ihre individuelle, soziale und situationsbezogene Wirkung beurteilen zu können. Es erscheint naheliegend, dass diese Form musikalischer Urteilsfähigkeit in hohem Maße von reichhaltigen Hörerfahrungen und musikalischen Aktivitäten in der frühen Kindheit profitieren und sich erst aus einer Vielfalt des musikalisch Erlebten eine genuin individuelle Musikpräferenz bzw. ein längerfristiger Musikgeschmack herausbilden kann. Somit kann das JeKi-Programm als eine Chance verstanden werden, durch die aktive Auseinandersetzung mit ansonsten möglicherweise nur schwer zugänglichen musikalischen Erlebniswelten das musikalische Stilempfinden und die ästhetische Urteilsfähigkeit zu schulen und die Bandbreite an vertrauten Musikstilen zu vergrößern, so dass die Verhaltensspielräume der heranwachsenden Kinder zur Nutzung von Musik erweitert werden. Das Ziel der vorliegenden Studie ist, einen Beitrag zur empirischen Überprüfung dieser Annahmen beizutragen.

2 Fragestellungen

Die übergeordnete Fragestellung lautet, inwieweit sich die vielfach beschriebene latente (lat. verborgen sein) Offenohrigkeit differenziert beschreiben lässt. Bei solch einer Betrachtung eines latenten Konstruktes wird davon ausgegangen, dass dieses nicht direkt beobachtet werden kann, sondern erst über die Zusammenfassung beobachtbarer Indikatorvariablen fassbar gemacht wird (Bollen, 2002, S. 606). An einem angemessenen Modell können dann Erklärungen anhand weiterer Merkmale vorgenommen werden. Hierfür kommen jene Merkmale in Frage, die im Rahmen des einleitenden Forschungsreviews als mögliche Einflussesvariablen beschrieben wurden. Somit sollen folgende Fragestellungen untersucht werden:

- Ist Offenohrigkeit als singulärer Faktor auf der Basis von Präferenzurteilen für unkonventionelle Musik beschreibbar?
- Gibt es Merkmale von Gruppen (von Kindern), die sich hinsichtlich ihrer Präferenzurteile unterscheiden, und wie lassen sich diese Gruppen ggf. differenzieren?
- Sind ältere Kinder weniger offenohrig als jüngere Kinder?
- Sind Jungen zu allen Messzeitpunkten weniger offenohrig als Mädchen?
- Welche Einflüsse haben schulischer und privater Instrumentalunterricht auf Offenohrigkeit?
- Wie wirkt sich ein Migrationshintergrund der Kinder auf Offenohrigkeit aus?
- Wie wirkt sich der sozio-ökonomischer Status auf die Offenohrigkeit aus?

3 Studiendesign

Das Studiendesign entspricht dem längsschnittlichen Design der SIGrun-Verbundstudie (zur ausführlichen Beschreibung siehe Kap. II.2, in diesem Band). Die spezifischen Bedingungen des hier beschriebenen Teilprojektes werden im Folgenden dargestellt.

3.1 Stichprobe

Die Stichprobe der vorliegenden Kohortenstudie basiert auf Grundschulklassen aus 20 Schulen in NRW und 13 Schulen in Hamburg und lässt sich nach fünf schulspezifischen Erhebungsgruppen differenzieren: JeKi-Schulen in NRW; JeKi-Schulen in Hamburg; JeKi-Schulen in NRW mit zusätzlichem Sportangebot; Schulen mit Instrumentalangebot in Hamburg; Sportschulen in NRW (vgl. Kap. II, in diesem Band). Die quantitativen Erhebungen wurden jeweils zum Ende eines Grundschuljahres durchgeführt ($t_{1\text{-}4}$: 2009 bis 2012), wobei zu jedem Erhebungszeitpunkt mindestens $n = 1\,000$ Kinder in die Erhebungen einbezogen wurden (vgl. Kap. II, in diesem Band). Vollständige Datensätze über alle vier Messzeitpunkte liegen für insgesamt $n = 735$ Kinder vor. Aus dieser Gesamtstichprobe wurden 28 Kinder ausgewählt, die in der Mitte des zweiten und des vierten Schuljahres an qualitativen Interviews in Kleingruppen teilnahmen. Neben den Befragungen der Kinder wurden quantitative als auch qualitative Erhebungen zusätzlich mit deren Eltern und Lehrkräften durchgeführt.
Bei der statistischen Analyse waren einige Herausforderungen aufgrund der zum Teil uneinheitlichen Datenstruktur zu bewältigen. Dies war der notwendigen Ausweitung der Erhebung im Verlauf der Studie geschuldet, die wegen der insgesamt geringen Stichprobenumfänge sowie wegen der nicht hinreichenden Homogenität der Teilstichproben im ursprünglich geplanten Split-Half-Design vorgenommen werden musste. Entsprechend ist eine Verknüpfung der Beurteilungen sämtlicher Musikbeispiele über alle Messzeitpunkte nicht möglich.

3.2 Methodisches Vorgehen und Durchführung

Die hier vorgestellten Befunde umfassen schwerpunktmäßig quantitative Datenanalysen auf Basis standardisierter Fragebögen. Die bereits erwähnten Interviews wurden als komplementäre Ergänzungen in den Längsschnitt integriert (Flick, 2011, S. 83f.), um die Ergebnisse der qualitativen und quantitativen Erhebungen nach Erzberger und Kelle (2003, S. 469) im Sinne eines „complementary model of triangulation“ einander ergänzend betrachten zu können.
Die Durchführung der quantitativen Erhebungen fand zu jedem Messzeitpunkt im Klassenverbund in den jeweiligen Klassenräumen während der regulären Schulzeit statt. Die insgesamt neun leitfadengestützten Interviews wurden ebenfalls während der regulären Schulzeit in Kleingruppen von zwei bis vier Kindern eines Klassenverbundes außerhalb des Klassenraumes durchgeführt. Im Folgenden werden die Messinstrumente der quantitativen und der Leitfaden der qualitativen Erhebungen erläutert.

3.2.1 Klingender Fragebogen als Messinstrument für Musikpräferenz

Bei der Erhebung der abhängigen Variable Musikpräferenz wurde auf einen „Klingenden Fragebogen“ zurückgegriffen, der bereits vielfach in verschiedenen Studien eingesetzt und erprobt wurde (Gembris & Schellberg, 2003, 2007; Kopiez & Lehmann, 2008; ebenfalls verwendet von Louven, 2011). Die Adaption des Klingenden Fragebogens gewährleistet eine Differenzierung für die Fragestellungen der vorliegenden Studie bei gleichzeitiger Vergleichbarkeit der Ergebnisse mit den Befunden der genannten früheren Studien.
Die Auswahl der 16 musikalischen Exzerpte (jeweils 30 Sekunden Dauer, mittleres Tempo um 60 bis 95 bpm; s. Tab. IV.1) orientierte sich an Hargreaves (1982) Beschreibung von Offenohrigkeit und umfasste somit vor allem

Musikstücke, die aus Sicht von Erwachsenen eher unkonventionell für Kinder klingen sollten. Hierbei wurde die nicht unproblematische Frage nach vermeintlicher Unkonventionalität im Sinne von Hargreaves auf eine generelle Vertrautheit mit unterschiedlichen Musikstilen bezogen (Hargreaves, 1982b, 1987; Hargreaves et al., 2006) und zudem angenommen, dass die Hörgewohnheiten der Eltern prägend für die kindliche Vertrautheit mit musikalischen Stilkategorien sind. Hargreaves (1982b, S. 14) vermutet, dass Musik der ‚Klassik' und ‚Avantgarde' nur selten Teil der elterlichen Hörgewohnheiten sind, was im Rahmen dieser Studie durch eine Befragung der elterlichen Hörpräferenzen bestätigt wurde. So wurden im Durchschnitt lediglich die musikalischen Stilkategorien ‚Rock/Pop' und ‚Musical' eindeutig positiv von den Eltern beurteilt, während alle anderen Stilkategorien negative Beurteilungen erhielten. Entsprechend wurde post hoc geschlossen, dass die für den Klingenden Fragebogen ausgewählten Musikstücke aus den Stilkategorien ‚Klassik' und ‚Avantgarde' sowie diejenigen aus nicht-westlichen Musikkulturen keinen bedeutenden Anteil an der (familiären) Hörerfahrung der Kinder haben und somit als eher unvertraut (also vermeintlich unkonventionell) einzustufen sind.

Die beiden Musikstücke mit ‚klassischen' musikalischen Stilmerkmalen (Mendelssohn und Bach) sowie das Musikstück der zeitgenössischen Kunstmusik (Henze) wurden aus früheren Studien übernommen. Das Bach-Exzerpt wurde ab Messzeitpunkt 2 hinzugenommen, da das Mendelssohn-Exzerpt etlichen Kindern bereits zu Messzeitpunkt 1 aus dem Film *Barbie in Die 12 tanzenden Prinzessinnen* (2006) bekannt war. Der Klingende Fragebogen umfasste zudem vier Musikstücke aus unterschiedlichen Musikkulturen (Türkei, Russland, China und Afrika), die für die Mehrheit der Kinder ebenfalls als eher unkonventionell eingestuft wurden. Sie entstammen jedoch aus Herkunftsregionen, für die in den Erhebungsgebieten der vorliegenden Studie (NRW und Hamburg) ein relativ hohes Migrationsaufkommen besteht (Statistisches Bundesamt, 2009, S. 66-68), so dass diese Musikstücke für Kinder mit entsprechendem Migrationshintergrund eher vertraut klingen sollten.

Des Weiteren wurde ein Musikstück (Garrett) aufgenommen, das dem Bereich des musikalischen ‚Cross-Overs' zwischen klassischen und populären Stilmerkmalen zugehört. Diese Kombination von Stilmerkmalen mag insgesamt eher unkonventionell erscheinen, doch möglicherweise dominieren bestimmte musikalische Merkmale eines Stils die kindliche Beurteilung. So könnte einerseits das Vorhandensein eines Schlagzeuges als möglicherweise „strukturelle Fundamentbasis" und „epochales Stilmerkmal" von Popmusik (Jaedtke, 2000, S. 206) zu einer eher konventionellen Beurteilung des Cross-Over-Stückes führen. Andererseits könnte auch der dominante ‚klassische' Klang des Streichinstrumentes Violine eine Kategorisierung als unkonventionell begünstigen (Louven, 2011).

Zudem umfasst der Klingende Fragebogen acht Musikstücke des Berliner Komponisten Achim Gieseler, die dieser als Auftragskompositionen für die vorliegende Studie komponiert hat. Die Vorgabe lautete, die Parameter ‚klassische' versus ‚populäre' Kompositionsweise, ‚klassische' versus ‚populäre' Instrumentationsweise sowie zusätzlich die An- bzw. Abwesenheit eines Schlagzeugklanges (vgl. Jaedtke, 2000) zu verarbeiten. Jede mögliche Kombination dieser Parameter wurde dabei berücksichtigt (s. Tab. IV.2). Die Auftragskompositionen mit einheitlich ‚populären' Stilmerkmalen sollten auf der Basis bereits vorliegender Befunde konventionelle Musik repräsentieren, während die Auftragskompositionen mit einheitlich ‚klassischen' Merkmalen (analog zu Bach und Mendelssohn) a priori als unkonventionell eingestuft wurden. Die Auftragskompositionen wurden hinsichtlich ihrer Kompositions- und Instrumentationsweise so kombiniert, dass sie sowohl konventionelle (‚populäre') als auch unkonventionelle (‚klassische') Stilmerkmale aufwiesen, was die eindeutige Kategorisierung durch die Kinder erschweren sollte. Wie bereits anhand des Cross-Over-Stücks dargelegt wurde, sollten diese Musikstücke somit Rückschlüsse darauf erlauben, welche musikalischen Parameter die kindliche Kategorisierung nach musikalischen Stilmerkmalen dominieren und dadurch möglicherweise die Präferenzäußerungen beeinflussen.

Eine positive Beurteilung der vermeintlich unkonventionellen Musikbeispiele wurde in der vorliegenden Studie als ein hohes Maß an Offenohrigkeit interpretiert.

Tabelle IV.1 Musikbeispiele des Klingenden Fragebogens

Musik-beispiel	Komponist / Interpret / Album	Komposition / Song	Dauer / Ausschnitt	Tempo
Übungs-beispiel	Friedbert Kerschbaumer / Die schönsten Kinderlieder auf der Panflöte	Ein Männlein steht im Walde	30 Sek. / 00:00-00:30	90 bpm
Afrika	Magi Shamba / Colors of Africa	Upepu	30 Sek. / 00:00-00:30	95 bpm
Türkei	Sümer Ezgü / Ege Toros Yörük Türkmen Türküleri (Anatolia Ethnic Music. Turkish Folk Music)	Ümmü	30 Sek. / 00:00-00:30	88 bpm
Russland	Samovar Russian Folk Music Ensemble / Some of our Best	Smyglyanka	28 Sek. / 01:24-01:52	86 bpm
China	Chinese Ensemble of Movie Music and Folk Music / Zhong Guo Dao Jiao Yin Le (Chinese Taoist Music)	Yu Fu Rong	30 Sek. / 01:20-01:50	90 bpm
Garrett	David Garrett / Encore	Air	30 Sek. / 01:52-02:23	60 bpm
Mendels-sohn	Felix Mendelssohn-Bartholdy	4. Sinfonie, 1. Satz	30 Sek. / 00:00-00:30	60 bpm
Henze	Hans Werner Henze	3. Sinfonie, 3. Satz „Beschwörungs tanz“	32 Sek. / 00:41-01:13	ca. 60 bpm
Bach	Johann Sebastian Bach	3. Orchester-Suite, „Gavotte I“	30 Sek. / 00:00-00:30	80 bpm
Acht Auftragskompositionen von Achim Gieseler (s. Tab. IV.2)			30 Sek.	90 bpm

Anmerkung. bpm = beats per minute

Tabelle IV.2 Auftragskompositionen

Musikbeispiel	Kompositionsstil	Instrumentation	Drum Set
Kla-Kla	Klassik	Klassik	Nein
Kla-Kla D			Ja
Kla-Pop		Pop	Nein
Kla-Pop D			Ja
Pop-Kla	Pop	Klassik	Nein
Pop-Kla D			Ja
Pop-Pop		Pop	Nein
Pop-Pop D			Ja

Die Erhebung des Klingenden Fragebogens erfolgte gemeinsam mit den weiteren standardisierten Instrumenten. Den Kindern wurden nacheinander die Musikbeispiele auf einem CD-Player in einheitlicher Lautstärke vorgespielt. Zur Vermeidung von Reiheneffekten wurden die Beispiele in verschiedenen Reihenfolgen dargeboten. Nach jedem Musikbeispiel wurde das Abspielen der CD angehalten, um den Kindern Zeit für ihr Urteil zu geben. Dieses wurde auf einer jeweils zu einem Musikstück dazugehörigen fünfstufigen ikonographischen Likert-Skala erfasst (Smileys unterschiedlichen emotionalen Ausdrucks von 1 „will ich häufiger hören“ bis 5 „will ich nicht hören“, siehe u.a. Gembris & Schellberg, 2007). Das arithmetische Mittel der Präferenz-Ratings reicht von $M = 1.64$ (Pop-Pop D zu t_1: Standardabweichung [SD] = 1.12) bis $M = 3.35$ (Türkei zu t_3: $SD = 1.37$), wobei die Ratings der Musikbeispiele zum Großteil eine rechtsschiefe Verteilung aufweisen (über alle Messzeitpunkte haben nur fünf Variablen eine Schiefe < 0; vier davon zu Messzeitpunkt 4) und nicht-normalverteilt sind. Der Nicht-Normalität wurde über den

Einsatz robuster Schätzer begegnet. Es sei betont, dass es sich bei den erhobenen Daten um Präferenzurteile und keine üblichen Testscores handelt. Während z.B. für einen Mathematiktest zumeist nur die Fähigkeit und die Itemschwierigkeit von Interesse sind, kann hier die differenzierte Wahrnehmung der Musikbeispiele zwischen den Kindern bedeutsam sein (vgl. Maydeu-Olivares & Böckenholt, 2009).

3.2.2 Messinstrumente der potentiellen Prädiktoren

Die Kinder bzw. deren Eltern beantworteten umfassende Fragebögen – unter anderem zum Alter und Geschlecht, zur Teilnahme am JeKi-Programm, zum privatem Instrumentalunterricht, zu sonstigen privat organisierten musikbezogenen Tätigkeiten (z.B. Konzertbesuche), zur kindlichen Persönlichkeit, zum sozialen Status, zum Migrationshintergrund, zu kognitiven Fähigkeiten sowie zum elterlichen Unterstützungsverhalten.
Die zentralen Prädiktoren der hier präsentierten Analysen wurden folgendermaßen operationalisiert: Die JeKi-Teilnahme und das Geschlecht wurden im Schülerfragebogen erhoben. Zusätzlich wurde im Elternfragebogen erfragt, ob das Kind in dem vorangegangenen Jahr privat organisierten Instrumentalunterricht (informell oder non-formal) erhalten hatte.

3.2.3 Leitfaden und Auswertung der Kinderinterviews

Den Interviews mit den Kindern lag ein Leitfaden zugrunde, der unter anderem Fragen zu zwei im Interview vorgespielten Musikstücken des Klingenden Fragebogens (Pop-Pop D und Kla-Kla) stellte sowie Fragen zur Lieblingsmusik und zu möglichen musikbezogenen Geschlechtsstereotypen umfasste. Zu beiden Interviewzeitpunkten wurden sowohl geschlechtshomogene als auch geschlechtsheterogene Gruppen befragt.
Bei der inhaltsanalytischen Auswertung wurde ein kombiniertes Vorgehen aus induktiver Kategoriegewinnung und deduktiver Kategorieanwendung verfolgt (vgl. Mayring, 2007). Hierbei wurde das vorläufige Kategoriensystem des 1. Interviewzeitpunkts (vgl. Beutler-Prahm, 2012) anhand der Daten des 2. Interviewzeitpunkts weiterentwickelt und dann als einheitliches Kategoriensystem für beide verwendet.
Nachfolgend werden jene Ergebnisse der qualitativen Erhebung zusammenfassend berichtet, die für das vertiefende Verständnis der quantitativen Befunde bedeutsam sind. Eine umfassende Analyse des qualitativen Datenmaterials wird aktuell im Rahmen des Promotionsvorhabens von Nicola Bunte vorgenommen (vgl. Bunte, 2013).

3.3 Statistische Analyse

Bei den statistischen Analysen wurden zwei Ansätze verfolgt, um ein möglichst umfassendes Bild der musikalischen Präferenz von Grundschulkindern zu zeichnen und die leitenden Forschungsfragen differenziert zu beantworten. Zum ersten wurden die Musikbeispiele faktorenanalytisch in einem mehrdimensionalen Modell analysiert. An diesem Modell wurden Hypothesentests über latente Regressionen und Mehrgruppenanalysen vorgenommen. Zum zweiten wurden die Kinder auf Basis ihrer Präferenzurteile zu homogenen Gruppen zusammengefasst und Unterschiede zwischen diesen Gruppen wurden inhaltlich interpretiert.

3.3.1 Gruppierung der Musikbeispiele auf Basis der Präferenzurteile

Die Beschreibung der statistischen Analyse zur Gruppierung der Musikbeispiele nach den Präferenzurteilen ist von Schurig und Busch ausführlich dokumentiert (u.a. Schurig et al., 2012; Busch et al., 2014a, 2014b), so dass hier die wesentlichen Aspekte zusammenfassend dargestellt werden können.

Zunächst wurden zur Überprüfung, ob Offenohrigkeit als latentes Konstrukt auf der Grundlage der Präferenzurteile darstellbar ist, Konfirmatorische Faktorenanalysen im Rahmen der Strukturgleichungsmodellierung vorgenommen (Bollen, 2002). Darüber hinaus wurde die für Schuluntersuchungen charakteristische hierarchische Schachtelung des Datenmaterials (vgl. Reinecke, 2005) beachtet, indem durch eine Gewichtungskorrektur die Standardfehler angemessen berücksichtigt wurden und somit auch die Theorie des Messfehlers adäquat einbezogen wurden. Mithilfe von Varianzanalysen mit Messwiederholung wurde die Entwicklung der Faktorscores im Längsschnitt betrachtet.

Darüber hinaus wurden die Urteilsveränderungen über die Zeit auf Ebene der einzelnen Musikstücke untersucht. Da einige Verteilungsvoraussetzungen der Daten (Sphärizität und Normalität) für die Berechnung einfaktorieller Varianzanalyen mit Messwiederholung nicht immer gegeben war, wurde der Friedman-Test als robuste nonparametrische Alternative eingesetzt. Zur post-hoc-Analyse von Messzeitpunkt zu Messzeitpunkt wurden zudem separate Einzeltests mittels des Wilcoxon-Tests berechnet.

3.3.2 Gruppierung der beurteilenden Kinder auf Basis der Präferenzurteile

Ein weiterer Ansatz verfolgt das Ziel, die Kinder auf Basis ihres individuellen Antwortverhaltens zu gruppieren und die empirisch festgestellten Gruppen zu interpretieren. Hierzu wurde das Verfahren der Latenten Klassenanalyse ausgewählt. Die latente Klassenanalyse (Latent Class Analysis) ist ein statistisches Hilfsmittel zum Auffinden empirischer Typologien (Bacher & Vermunt, 2010). Die Latente Klassenanalyse ist ein probabilistisches Verfahren, bei dem alle Kinder auf der Basis von Wahrscheinlichkeiten gewissen Klassen zuordnet werden.

Zunächst wurde auf der Basis gängiger Informationskriterien geprüft, wie viele latente Klassen vorlagen. Die Verrechnung der unabhängigen Variablen (Geschlecht und Teilnahme an privat organisiertem Instrumentalunterricht) erfolgte dabei bei jeder möglichen Klassenlösung in einem Schritt mit der Schätzung der Klassenzugehörigkeiten. Die Auswahl der Klassenzahl geschah auf Basis von heuristischen Vergleichen der logarithmierten Likelihood Funktion (LL) und darauf basierender Indizes (vgl. Rost, 2004).

Diese Klassen wurden nachfolgend auf ihre Validität geprüft und empirisch beschrieben. Die Analysen wurden mit der Statistiksoftware `R` und dem Paket `poLCA` (Linzer & Lewis, 2013) durchgeführt, welches die Modellierung auf der Basis polytomer Items erlaubt. Die Analysen wurden mehrfach mit unterschiedlichen Startwerten wiederholt, um für lokale Maxima der LL-Funktion zu kontrollieren. Es wurden ausschließlich vollständige Datensätze verwendet, die Stichprobenumfänge pro Messzeitpunkt sind in der Tabelle IV.6 vermerkt.

4 Ergebnisse der Studie

Im Folgenden werden die Ergebnisse zu den Konfirmatorischen Faktorenanalysen, zu den Längsschnittanalysen und zu den Regressionsanalysen zusammenfassend beschrieben. Ausführlicher dargestellt werden Analysen zur Varianz auf Itemebene sowie Analysen zu den latenten Klassen.

4.1 Gruppierung der Musikbeispiele auf Basis der Präferenzurteile

Die Konfirmatorischen Faktorenanalysen legt eine Strukturierung der Musikbeispiele in drei Faktoren nahe, die sich musikalisch sinnfällig mit den Begriffen ‚Ethno/Avantgarde', ‚Klassik' und ‚Pop' inhaltlich beschreiben lassen (vgl. Louven, 2011). Diese faktorielle Lösung kann für die ersten drei Messzeitpunkte angenommen werden und ist in Abbildung IV.1 exemplarisch dargestellt (s. Tab. IV.3 für die Fit-Indizes der einzelnen Modelle für Messzeitpunkt 1 bis Messzeitpunkt 4). Drei Musikbeispiele werden aufgrund von Mehrfachladungen ausgeschlossen (Russland, Kla-Kla D und Pop-Kla). Die Variable Geschlecht wird als Instrumentalvariable aufgenommen, nachdem Testungen über das Geschlecht keine Invarianz in der Ladungsstruktur ergeben haben. Damit wird im Strukturmodell für einen Bias durch diese kontrolliert. Mit Ausnahme der Variablen Alter und Geschlecht kristallisiert sich kein weiterer bedeutsamer Prädiktor für alle Messzeitpunkte heraus, d.h. den Variablen schulischer / privater Instrumentalunterricht, Persönlichkeit, sozialer Status und Migrationshintergrund kann keine Vorhersagekraft für die Präferenzäußerungen zugewiesen werden. Ein einfaktorielles Modell, welches Offenohrigkeit direkt hätte beschreiben können, erreicht zu keinem der Messzeitpunkte einen vergleichbar guten Fit.

Für Messzeitpunkt 4 kann das beschriebene Modell nicht akzeptiert werden. Für die ersten drei Messzeitpunkte zeigt sich eine zunehmende Verschlechterung der Präferenzurteile für alle drei Faktoren, für Messzeitpunkt 4 setzt sich dieser Trend nur für die Faktoren ‚Ethno/Avantgarde' und ‚Pop' fort, während er sich für den Faktor ‚Klassik' umzukehren scheint. Zugleich ergeben die geschlechtsspezifischen Verläufe der ersten drei Messzeitpunkte, dass die stets negativeren Beurteilungen des ‚Klassik'-Faktors auf die Präferenzurteile der Jungen zurückzuführen sind.

Bei den Faktoren ‚Ethno/Avantgarde' und ‚Pop' zeigen sich hingegen keine signifikanten geschlechtsspezifischen Unterschiede im Urteilsverhalten.

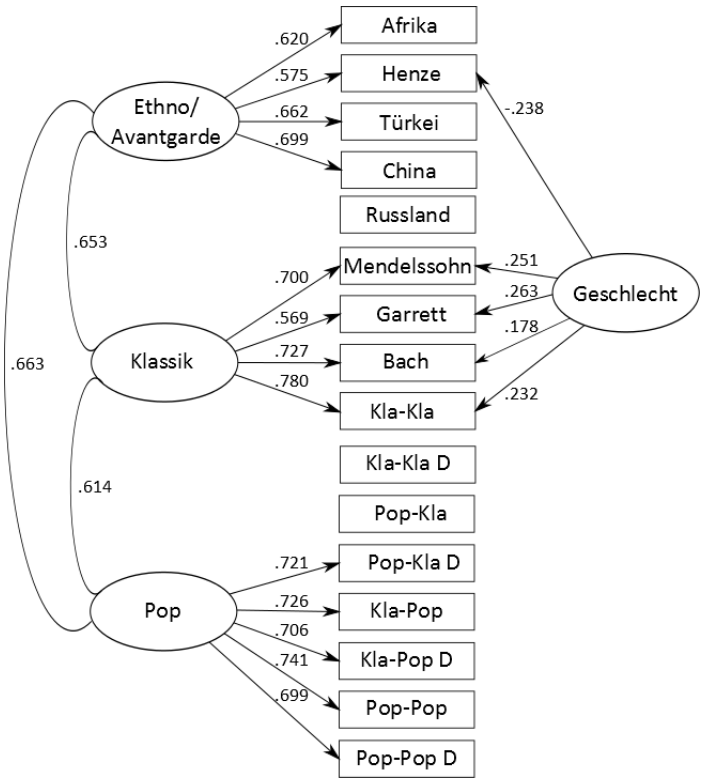


Abbildung IV.1 Faktormodell zum Messzeitpunkt 3

Tabelle IV.3 Fit-Indizes der konfirmatorischen Faktormodelle der vier Messzeitpunkte

MZP	*n*	Chi^2	*df*	*p*	*CFI*	*RMSEA*
1	444	21.8	11	0.026	0.982	0.047
2	890	42.4	16	< 0.001	0.986	0.043
3	1172	92.7	23	< 0.001	0.978	0.051
4	995	395.3	23	< 0.001	0.875	0.110

Anmerkung. Chi^2 = Pearson Chi^2 auf Anpassungsgüte; *df* = Freiheitsgrade; *p* = Signifikanzniveau des Chi^2; *CFI* = Comparative Fit Index; *RMSEA* = Root Means Square Error of Approximation

4.2 Verlaufsstruktur der einzelnen Präferenzurteile

Zum besseren Verständnis der oben beschriebenen veränderten Urteilsstruktur zu Messzeitpunkt 4 werden die folgenden Analysen auf Ebene der einzelnen Musikstücke vorgenommen. Die Einzelanalysen der paarweisen Vergleiche sind in Tabelle IV.4 zusammengetragen. Es wird deutlich, dass sämtliche Musikstücke der (für die Messzeitpunkte 1 bis 3 gültigen) Faktoren ‚Ethno/Avantgarde' und ‚Pop' unabhängig vom Geschlecht der Kinder zu Messzeitpunkt 4 signifikant schlechtere Bewertungen gegenüber Messzeitpunkt 3 erhielten. Allerdings liegen selbst diese negativeren Präferenzurteile immer noch oberhalb des ‚neutralen' Skalenmittelpunktes (3) und werden somit zwar schlechter, aber weiterhin positiv beurteilt. Bei den ‚Ethno/Avantgarde'-Stücken ist hingegen eine deutlichere Verschlechterung der Beurteilung festzustellen, die bis in jenen negativen Skalenbereich (> 3) hineinreicht, der Missfallen signalisiert. In Hinblick auf diese Musikstücke ist also von einem Rückgang an Offenohrigkeit auszugehen.

Tabelle IV.4 Effekte zwischen den Messzeitpunkten pro Musikbeispiel

	Gesamt			Jungen			Mädchen		
	MZP 1 auf 2	MZP 2 auf 3	MZP 3 auf 4	MZP 1 auf 2	MZP 2 auf 3	MZP 3 auf 4	MZP 1 auf 2	MZP 2 auf 3	MZP 3 auf 4
Afrika	-3.27*	-6.61*	-5.31*	-2.55*	-3.18*	-3.67*	-2.05*	-6.20*	-3.84*
Kla-Pop			-5.71*			-4.75*			-3.31*
Mendels-sohn	-2.30*	-4.83*	-7.54*	-1.87	-3.75*	-1.33	-1.37	-3.15*	-8.83*
Pop-Pop			-16.00*			-8.76*			-13.40*
Henze	-0.23	-3.75*	-7.19*	-0.29	-3.12*	-5.367*	-0.03	-2.09*	-4.78*
Garrett	-6.07*	-6.49*	-8.22*	-4.50*	-4.120*	-3.42*	-4.11*	-4.98*	-8.15*
Türkei	-0.59	-6.66*	-15.34*	-0.52	-6.12*	-9.22*	-0.26	-3.35*	-12.29*
Kla-Kla			-8.71*			-2.98*			-8.84*
China	-3.38*	-5.28*	-9.19*	-2.94*	-5.34*	-4.08*	-1.81	-2.06*	-8.72*
Pop-Kla			-8.58*			-7.73*			-4.51*
Russland	-4.92*	-2.94*	-7.15*	-3.41*	-3.13*	-5.62*	-3.58*	-1*	-4.48*
Kla-Pop D			-12.86*			-10.93*			-7.15*
Pop-Pop D			-19.04*			-12.98*			-13.84*
Pop-Kla D			-13.88*			-9.51*			-10.12*
Kla-Kla D			-11.47*			-10.63*			-5.45*
Bach		-7.10*	-10.42*		-6.11*	-4.63*		-3.96*	-9.66*

Anmerkung. Abgetragen sind die *Z*-Werte von Wilcoxon-Rangsummen-Tests. Werte, die auf einem Niveau von $p < 0{,}05$ asymptotisch signifikant sind, sind mit * markiert. MZP = Messzeitpunkt.

Die Musikstücke, die bei der Konfirmatorischen Faktorenanalyse in dem Faktor ‚Klassik' zusammengefasst werden, weisen hingegen im Vergleich von Messzeitpunkt 3 und Messzeitpunkt 4 jeweils signifikant positivere Beurteilungen auf. Lediglich für das Musikstück Mendelssohn zeigt sich ein geschlechtsspezifischer Unterschied, da nur die positivere Beurteilung der Mädchen Signifikanz erreicht. Zusätzlich zu diesen ‚Klassik'-Musikstücken ergibt sich dieselbe positive Veränderung auch bei dem Musikstück Russland, das wegen der Doppelladungen auf den Faktoren ‚Ethno/Avantgarde' und ‚Klassik' in den Strukturgleichungsmodellierungen ausgeschlossen ist. Diese Itemanalysen verdeutlichen somit in markanter Weise, dass die Kinder zu Messzeitpunkt 4 sämtliche Musikstücke mit klassischen Stilmerkmalen – und hierzu zählt durchaus auch das Musikstück Russland – positiver beurteilen und somit in keiner Weise dem üblichen „Verschwinden" von Offenohrigkeit entsprechen. Die für diese Stücke bisher charakteristischen geschlechtsspezifische Differenzierung der Präferenzurteile zeigt sich zu Messzeitpunkt 4 nicht mehr. Abbildung IV.2 bietet eine zusammenfassende Darstellung dieser Befunde auf Basis der drei Faktoren ‚Klassik', ‚Pop' und ‚Ethno/Avantgarde'.

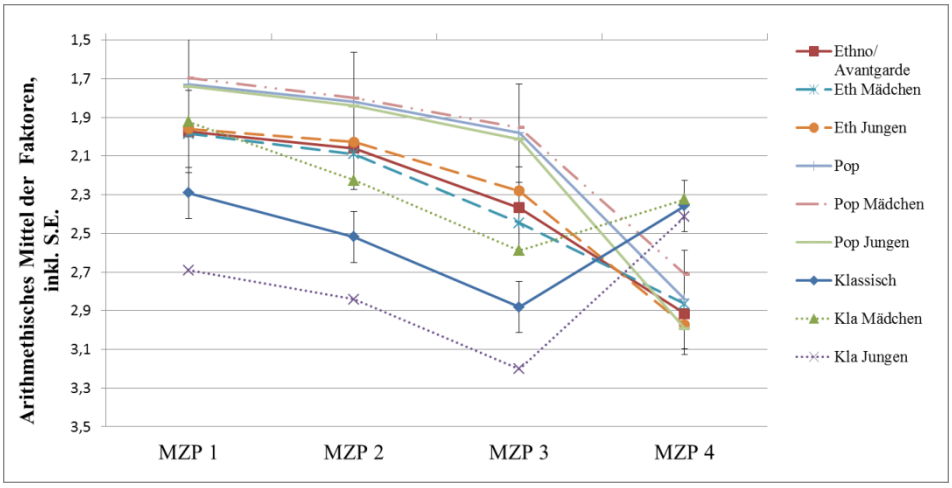


Abbildung IV.2 Verlauf der mittleren Präferenzurteile pro Faktor

Anmerkung. Abgetragen sind die arithmetischen Mittel der drei Faktoren ‚Ethno/Avantgarde', ‚Pop' und ‚Klassik' inklusive Standardfehler pro Messzeitpunkt (gesamt und jeweils differenziert nach Geschlecht).

4.3 Gruppierung der urteilenden Kinder auf Basis der Präferenzurteile

Für die Klassifizierung der Kinder anhand ihrer Präferenzurteile werden nur jene vermeintlich unkonventionellen Musikstücke einbezogen, die zu jedem der vier Messzeitpunkte erhoben wurden. Dies umfasst die Musikstücke aus Afrika, der Türkei, China und Russland sowie die Musikbeispiele von Mendelssohn, Henze und Garrett.

Da aus bisherigen Analysen bekannt ist, dass das Geschlecht einen Einfluss hatte, wird dieses als Kovariate einbezogen. Darüber hinaus wird die Teilnahme am Instrumentalunterricht einbezogen, dessen Einfluss zuvor theoretisch hergeleitet wurde. Unterschieden wird hierbei nach der Teilnahme an privat organisiertem Instrumentalunterricht in der Freizeit (entweder informell, z.B. durch die Eltern, oder non-formal professionell, z.B. durch eine Musikschule) und die Teilnahme an Jeki als ein non-formales professionelles Angebot im schulischen Rahmen. Es kann beobachtet werden, dass weder informelle Lernmöglichkeiten (z.B. durch die Eltern) noch JeKi einen Einfluss auf die Gruppenbildung ausübten. Zur Verringerung der Modellkomplexität werden daher diese beiden Merkmale aus der Klassenanalyse ausgeschlossen. Als zweiter Prädiktor neben den Präferenzurteilen verbleibt somit die Teilnahme an privat organisiertem, non-formal professionellem Instrumentalunterricht in der Freizeit (z.B. in einer Musikschule).

Für die Analysen wird angenommen, dass sich die latenten Klassen durch Kinder definieren lassen, die mit einer jeweils hohen Wahrscheinlichkeit entweder positive, negative oder neutrale Präferenzurteile fällen, also durch Kinder, die die unkonventionellen Musikstücke gern wieder hören wollen, die diese nicht wieder hören wollen oder die diesbezüglich indifferent sind. Es ist zu erwarten, dass Jungen dabei häufiger der negativ urteilenden Gruppe zugeordnet werden. Obwohl das Geschlecht für die Beurteilung der Präferenz von Musikstücken unterschiedlicher Kulturregionen und der ‚Pop'-Beispiele keinen starken Einfluss hat, ist der Effekt auf Musikstücke mit ‚klassischen' Stilmerkmalen hoch. Die Verteilung der Prädiktoren pro Messzeitpunkt ist in der Tabelle IV.5 aufgezeigt.

In der Tabelle IV.6 sind die Maximierte Log-Likelihood Funktion und das Akaike Information Criterion sowie das Bayes Information Criterion pro Messzeitpunkt und Klassenzahl dargestellt. Die Lösung mit einem Cluster und ohne Kovariaten wird als Referenz angeführt.

Tabelle IV.5 Häufigkeiten in den Prädiktorvariablen pro Messzeitpunkt

	Geschlecht		Privat organisierter Instrumentalunterricht	
	Jungen (Code = 1)	Mädchen (Code = 2)	Ja (Code = 1)	Nein (Code = 0)
MZP 1	322	256	233	242
MZP 2	378	299	450	227
MZP 3	341	266	403	204
MZP 4	255	214	271	198

Tabelle IV.6 Modellkennzahlen für unterschiedliche Klassenzahlen und Messzeitpunkte

MZP	Klassen	*n*	*npar*	*LL*	*AIC*	*BIC*
1	1	578	28	-5175	10407	10529
	2	578	59	-4838	9795	10052
	3	578	90	-4732	9644*	10036*
	4	578	121	-5033	10309	10836
	5	578	152	-5091	10486	11149
2	1	677	28	-6657	13370	13496
	2	677	59	-6182	12483	12749
	3	677	90	-6000	12180	12587*
	4	677	121	-5937	12117*	12663
	5	677	152	-6278	12861	13547
3	1	607	28	-6451	12959	13082
	2	607	59	-6141	12400	12661
	3	607	90	-5997	12175	12571*
	4	607	121	-5991	12224	12757
	5	607	152	-5883	12071*	12741
4	1	469	28	-5028	10113	10229
	2	469	59	-47223	9563	9808
	3	469	90	-4594	9368	9742*
	4	469	121	-4523	9288*	9791
	5	469	152	-4476	9257	9887

Anmerkung. *npar* = Zahl der im Modell zu schätzenden Parameter; *LL* = maximierte Log-Likelihood Funktion; *AIC* = Akaike Information Criterion; *BIC* = Bayes Information Criterion. Die jeweilig niedrigsten Informationskriterien sind mit * hervorgehoben.

Für den Messzeitpunkt 1 ergibt sich ein klares Bild auf Basis der Informationskriterien. Die Kriterien sprechen für die Annahme der 3-Klassenlösung.

Für den Messzeitpunkt 2 spricht das AIC für eine 4- und das BIC für eine 3-Klassenlösung. Aber die -2LL verbessert sich zwischen der 3-Klassen- und der 4-Klassenlösung nur noch um 1%. Zum Messzeitpunkt 3 spricht das AIC sogar für eine 5-Klassenlösung, der BIC aber erneut für eine 3-Klassenlösung. Die Verbesserung der -2LL ist zwischen der 3- und der 4-Klassenlösung nur 0,1% groß und zwischen 4- und 5-Klassenlösung ebenfalls nur 1%. Zum Messzeitpunkt 4 ist das AIC des 4-Klassenmodells am geringsten und das BIC des 3-Klassenmodells. Die Verbesserung in der -2LL beträgt erneut nur 1%. Zusammenfassend lässt sich festhalten, dass eine 3-Klassenlösung für alle Messzeitpunkte annehmbar ist und der besseren Vergleichbarkeit halber einheitlich gewählt wird.

Die prozentualen Anteile der Kinder pro Klasse sind in Tabelle IV.7 aufgezeigt. Die Ordnung der Klassen wird der Ordnung des ersten Durchlaufes angepasst, so dass diese vergleichbar sind.

Tabelle IV.7 Prozentuale Anteile der Kinder pro latenter Klasse und Messzeitpunkt

MZP	Klasse 1	Klasse 2	Klasse 3
1	17,5%	30,5%	52,1%
2	16,5%	36,5%	47,0%
3	32,1%	30,0%	37,9%
4	28,8%	41,2%	30,0%

Die geringste Zellhäufigkeit liegt bei 110 Kindern (16,5%) in der Klasse 1 des Messzeitpunkts 2. Der durchschnittliche Modus der $p(k_x)$ pro Fall, also der Mittelwert der höchsten Klassenzugehörigkeitswahrscheinlichkeiten jedes Kindes, kann bei einer probabilistischen Klassifizierung als ein Reliablilitätsmaß interpretiert werden (Rost, 2004, S. 161). Diese Maßzahlen liegen sämtliche in einem guten Bereich und deuten somit auf eine gute Separation der Klassen durch das Modell hin (0.895 für t_1, 0.925 für t_2, 0.896 für t_3 und 0.908 für t_4). Die Klassenlösung wird durch das Entfernen und die Hinzugabe einzelner Parameter (z.B. einzelner Präferenzen) auf Stabilität geprüft und für ausreichend befunden.

Die Charakterisierung der Klassen erfolgt auf der Basis der Antwortwahrscheinlichkeiten pro Item. Es zeigt sich, dass die Kinder der Gruppe 1 zu allen Messzeitpunkten dazu tendieren, eher negative Präferenzäußerungen für die dargebotenen Musikbeispiele abzugeben. Bei den Kindern der Gruppe 2 fallen diese Präferenzäußerungen eher neutral aus, während die Kinder der Gruppe 3 positive Äußerungen charakterisieren. In Anlehnung an das Konstrukt der Offenohrigkeit werden die Gruppen demnach als ‚Verschlossene Gruppe', ‚Indifferente Gruppe' und ‚Offene Gruppe' interpretiert. Über diese erste Charakterisierung hinaus lässt sich jedoch festhalten, dass die ‚Verschlossene Gruppe' nicht ausschließlich negativ urteilt. Der Skalenpunkt 1 („will ich häufiger hören") wird sogar häufiger gewählt als in der ‚Indifferenten Gruppe'. Dafür entfallen die Skalenpunkte 2 bis 4 bis zum Messzeitpunkt 4 nahezu. Es wäre also falsch, von einer rein negativ urteilenden Gruppe zu sprechen. Die Kinder antworten aber absoluter als die Kinder anderer Gruppen. Für die ‚Offene Gruppe' hingegen gilt, dass nahezu keine negativen Äußerungen abgegeben werden.

In der Abbildung IV.3 sind die Prozentanteile pro Messzeitpunkt aufgezeigt. Hier lässt sich eine Verschiebung der Präferenzurteile beobachten. Während zum Messzeitpunkt 1 noch über 50% der Kinder der ‚Offenen Gruppe' zugeteilt werden, sind es zum Messzeitpunkt 4 nur noch circa 30%. Die Gruppe wird sukzessive kleiner. Die ‚Verschlossene Gruppe' erfährt zwischen Messzeitpunkt 2 und Messzeitpunkt 3 einen Anstieg von circa 15%. Zwischen den Messzeitpunkten 1 und 2 verändert sie sich aber nur um etwa 1%. Zwischen den Messzeitpunkten 3 und 4 nimmt ihr Umfang um etwa 3% ab. Die ‚Indifferente Gruppe' wächst zwischen den Messzeitpunkten 1 und 4 um etwa 11% an.

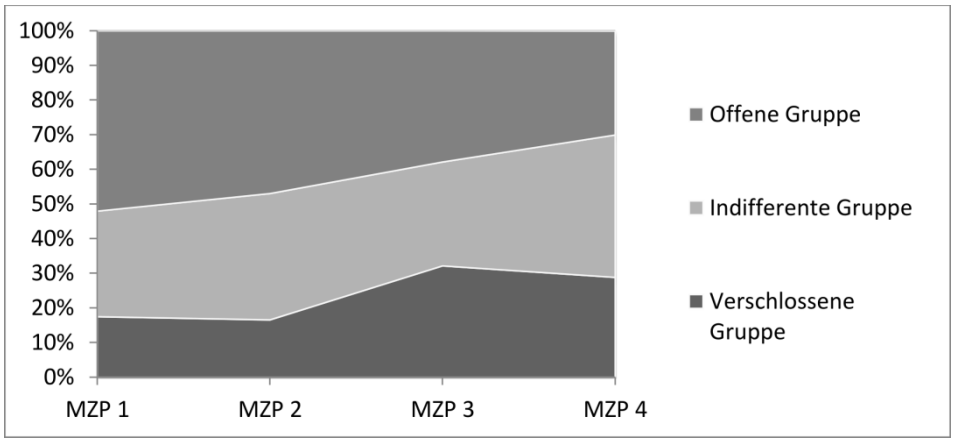


Abbildung IV.3 Prozentuale Häufigkeiten der drei Gruppen pro Messzeitpunkt

Nachfolgend werden die Effekte der unabhängigen Variablen zwischen den Klassen beschrieben (s. Tab. IV.8). Es handelt sich dabei um eine logistische Regression, d.h. es wird eine Klasse als Referenzgruppe definiert. In dem vorliegenden Fall wird hierfür die Klasse 1 gewählt, die jene Gruppe an Kindern umfasst, die zu allen Messzeitpunkten überproportional häufig negative Urteile abgeben, so dass eine klare Richtung der Effekte erkennbar wird.

Tabelle IV.8 Regressionsgewichte der Analyse latenter Klassen

MZP	Prädiktorvariable	Klasse 2 gegenüber Klasse 1			Klasse 3 gegenüber Klasse 1		
		Beta Koeff.	*SE*	*p*	*Beta Koeff.*	*SE*	*p*
1	(Intercept)	1.57	0.63	0.013	2.48*	0.54*	< 0.001
	Geschlecht	-0.69	0.37	0.058	-1.04*	0.32*	< 0.001
	Privat organisierter Instrumentalunterricht	-0.01	0.05	0.784	0.02	0.04	0.634
2	(Intercept)	3.53*	0.68*	<0.001	2.99*	0.689	< 0.001
	Geschlecht	-1.74*	0.37*	<0.001	-1.49	0.37	< 0.001
	Privat organisierter Instrumentalunterricht	0.41	0.31	0.178	0.31	0.30	0.297
3	(Intercept)	2.45*	0.54*	<0.001	1.21*	0.48*	0.011
	Geschlecht	-1.78*	0.34*	<0.001	-0.80*	0.27*	0.003
	Privat organisierter Instrumentalunterricht	-0.05	0.28	0.851	0.23	0.25	0.364
4	(Intercept)	1.28	0.55	0.020	1.00	0.56	0.073
	Geschlecht	-0.98*	0.32*	0.003	-0.91*	0.32*	0.005
	Privat organisierter Instrumentalunterricht	0.95*	0.30*	0.002	0.78*	0.31*	0.012

Anmerkung. Die Regressionen wurden in einem Schritt mit der Klassifikation vorgenommen. *Beta-Koeffizient* = Regressionsgewicht; *SE* = Standardfehler des Regressionsgewichtes; *p* = Signifikanzniveau des Regressionsgewichtes. Die Ergebnisse von Regressionen, die auf einem Niveau von $p < 0.05$ signifikant wurden, sind mit * hervorgehoben.

Für eine Zusammenfassung der Ergebnisse wird der Prozentanteil der Kinder mit einer spezifischen Ausprägung auf einer der Prädiktorvariablen pro Gruppe in den Abbildungen IV.4 und IV.5 dargestellt.

Von Interesse sind dabei Unterschiede in den Häufigkeiten innerhalb der Klassen. Es sind also die Differenzen in den prozentualen Häufigkeiten abgetragen.

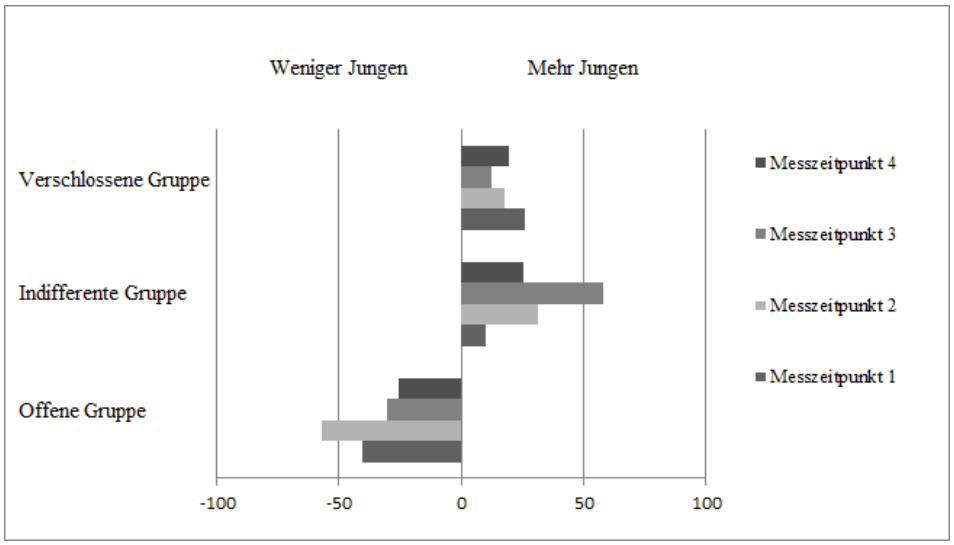


Abbildung IV.4 Differenzen der prozentualen Häufigkeiten des Geschlechtes pro Klasse

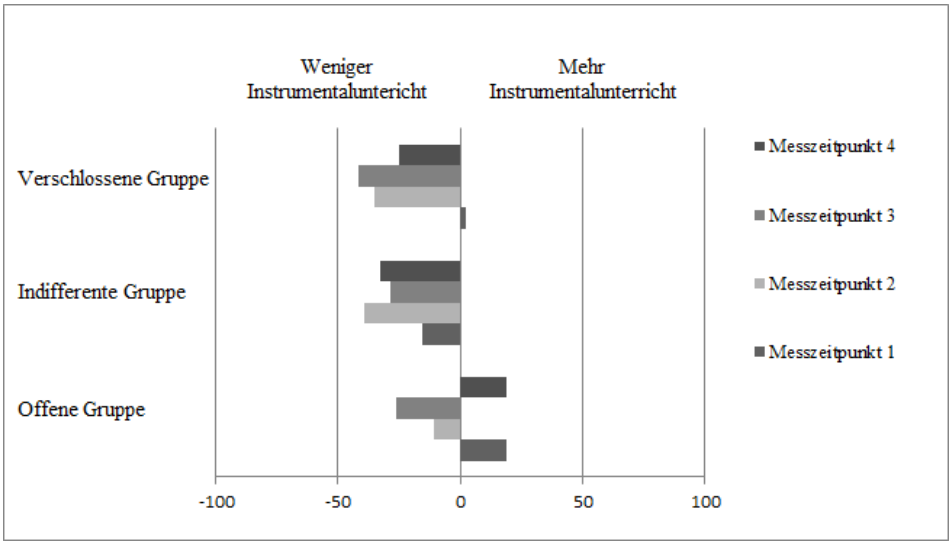


Abbildung IV.5 Differenzen der prozentualen Häufigkeiten des Erhaltens von privatem Instrumentalunterricht pro Klasse

In der Abbildung IV.4 ist zu erkennen, dass die Wahrscheinlichkeit für Jungen, zu der ‚Verschlossenen Gruppe' zu gehören, immer signifikant höher ist als für Mädchen ($p < 0.001$). Nur für den Messzeitpunkt 1 und nur für die ‚Indifferente Gruppe' gibt es keinen signifikanten Unterschied zwischen den Geschlechtern.

Bezogen auf die Teilnahme an privat organisiertem Instrumentalunterricht im Vorjahr lässt sich anhand von Abbildung IV.5 beobachten, dass es basierend auf Verteilungsvergleichen nur einen hochsignifikanten Unterschied zwischen den Gruppen zum Messzeitpunkt 4 gibt ($p < 0.001$). Trotzdem zeichnen sich Differenzen innerhalb der Gruppen klar ab und es kann erkannt werden, dass die Wahrscheinlichkeit zu

der ‚Offenen Gruppe' zu gehören zu den Messzeitpunkten 1 ($p < 0.048$) und 4 ($p < 0.001$) höher ist, wenn die Schülerinnen und Schüler privaten Instrumentalunterricht erhalten.

Abschließend wird die Transitivität zwischen den Messzeitpunkten beschrieben, also der Anteil jener Kinder, die zwischen den Messzeitpunkten gleichen respektive unterschiedlichen Gruppen zugeordnet werden. Fehlende Werte, also Werte von Kindern, die nicht an zwei aufeinander folgenden Messzeitpunkten teilnehmen, werden ausgeschlossen.

Die Kontingenzkoeffizienten, also der Zusammenhang der Klassen über die Messzeitpunkte, betragen 0.31 für den Abgleich von Messzeitpunkt 1 zu Messzeitpunkt 2, 0.36 von Messzeitpunkt 2 zu Messzeitpunkt 3 und 0.26 von Messzeitpunkt 3 zu Messzeitpunkt 4. Es werden also mittlere Übereinstimmungen beobachtet, die zwischen den Messzeitpunkten 3 und 4 am geringsten sind.

Übertragen auf Wahrscheinlichkeiten stellen sich die Transitivitäten wie in Tabelle IV.9 gezeigt dar. Dabei wird abgetragen, wie hoch die Wahrscheinlichkeiten sind, von einer Klasse zu einem Messzeitpunkt in eine Klasse in dem darauf folgenden Messzeitpunkt zu gelangen. Zusätzlich werden zur besseren Interpretation die Zellhäufigkeiten abgetragen. Wie erwartet sind auf den spezifischen Diagonalen relativ hohe Wahrscheinlichkeiten zu erkennen, im Durchschnitt ungefähr 47%. Gleichzeitig sind die Wahrscheinlichkeiten, eine Klasse zu überspringen, also von der ‚Verschlossen Gruppe' in die ‚Offene Gruppe' oder vice versa zu wechseln, verhältnismäßig gering. Im Durchschnitt liegen diese bei ungefähr 21%. Die summierte Wahrscheinlichkeit eines Klassen-Wechsels steigt im Vergleich von Messzeitpunkt 3 zu Messzeitpunkt 4 etwas an (26% und 25% gegenüber 28%). Aufgrund der teilweise sehr geringen Zellhäufigkeiten sollten diese Ergebnisse aber nicht überinterpretiert werden. Über alle Vergleiche hinweg kann also ein Großteil der Kinder einer einzelnen Klasse zugeordnet werden, wobei die Klassen aber keinesfalls vollkommen undurchlässig sind.

Tabelle IV.9 Transitionswahrscheinlichkeiten zwischen Gruppen und Messzeitpunkten

		Verschlossen		Indifferent		Offen	
		p	*n*	*p*	*n*	*p*	*n*
MZP 1 zu MZP 2 (*n*=303)	Verschlossen	0.377	20	0.415	22	0.208	11
	Indifferent	0.170	17	0.620	62	0.210	21
	Offen	0.167	25	0.373	56	0.460	69
MZP 2 zu MZP 3 (*n*=416)	Verschlossen	0.500	35	0.143	10	0.357	25
	Indifferent	0.300	60	0.440	88	0.260	52
	Offen	0.178	26	0.226	33	0.596	87
MZP 3 zu MZP 4 (*n*=310)	Verschlossen	0.389	37	0.453	43	0.158	15
	Indifferent	0.265	26	0.500	49	0.235	23
	Offen	0.197	23	0.385	45	0.419	49

Anmerkung. *p* = Wahrscheinlichkeiten, *n* = Zellhäufigkeiten. Zeilensummen ungleich 1 entstanden durch Rundungsfehler.

4.4 Erklärungsansätze aus den Analysen der Kinderinterviews

Die bisherigen Analysen der Kinderinterviews fokussierten auf die Frage, ob sich Hinweise auf das Vorhandensein von musikalischen Konzepten finden lassen und inwieweit sich die Aussagen der Kinder zur musikalischen Präferenz auf diesen Konzepten gründen lassen. Im Folgenden werden die Ergebnisse zusammenfassend in ihrem Bezug zu den Ergebnissen der präsentierten quantitativen Erhebung dargestellt (für eine ausführlichere Darstellung siehe Busch et al., 2013; Busch et al., 2014a, 2014b; Bunte, 2013).
Im Rahmen der qualitativen Analyse wird ein Kategoriensystem entwickelt, das Behnes (1975) „musikalische Konzepte" als Bezugspunkt nimmt, wobei zunächst vor allem Aussagen zu inter-individuell bedeutsamen musikalischen Konzepten interessieren. Somit wird die Analyse von der Frage geleitet, ob sich aus den

Interviews inter-individuell einheitliche Vorstellungen, Einstellungen, Annahmen oder Vorurteile „hinsichtlich eines mehr oder weniger umgrenzten musikalischen Objektes“ (Behne, 1975, S. 36; beispielsweise einer Stilkategorie) aufspüren lassen, die in Bezug zur Musikpräferenz relevant erscheinen. Aus den Aussagen der Kinder lassen sich induktiv – auch unabhängig von den konkreten Interviewfragen – vier solcher musikalischer Konzepte ableiten, die folgendermaßen umschrieben werden: ‚Mädchenmusik’, ‚Jungenmusik’, ‚Rockmusik’ und ‚Chartsmusik’.
Wie bereits an anderer Stelle publiziert (vgl. Busch et al., 2014b) bieten die Interviewergebnisse einige Erklärungsansätze für die in der Strukturgleichungsmodellierung beobachteten negativeren Präferenzäußerungen der Jungen bezüglich des Faktors ‚Klassik’ zu Messzeitpunkt 1 bis Messzeitpunkt 3. So äußern vor allem Jungen zum 1. Interviewzeitpunkt im Rahmen des Konzeptes ‚Rockmusik’ persönliche Präferenzen für das Instrument Schlagzeug, das in den Musikbeispielen des Faktors ‚Klassik’ nicht vorkommt und somit eine Ablehnung nahelegt. Zudem steht das Konzept ‚Rockmusik’ im Zusammenhang mit den zum 1. Interviewzeitpunkt stark ausgeprägten geschlechtsspezifischen Konzepten ‚Jungenmusik’ und ‚Mädchenmusik’. Bei den Jungen waren zum 1. Interviewzeitpunkt die Konzepte ‚Rockmusik’ und ‚Jungenmusik’ quasi deckungsgleich, während ‚Mädchenmusik’ von Jungen mit Charakteristika beschrieben wurde, die auf Assoziationen mit ‚klassischer’ Musik schließen ließen (Cello, ruhiger etc.). Möglicherweise grenzten sich die Jungen auch durch die Ablehnung des Faktors ‚Klassik’ zusätzlich von ‚Mädchenmusik’ ab. Zum 2. Interviewzeitpunkt hatte sich die Begrenzung von ‚Jungenmusik’ auf ‚Rockmusik’ jedoch zugunsten ausdifferenzierter und individueller Konzepte aufgelöst. Zudem zeigte sich eine Sensibilisierung für die normative Bewertung von ‚Jungenmusik’ und ‚Mädchenmusik’, die die persönliche Zuordnung zu den beiden geschlechtsspezifischen musikalischen Konzepten nicht mehr zuzulassen scheint.
In Bezug auf den Faktor ‚Klassik’ ergeben die statistischen Analysen auf der Ebene der einzelnen Musikstücke, dass diese unabhängig vom Geschlecht zum Messzeitpunkt 4 positiver beurteilt werden. Aus den Analysen des 2. Interviewzeitpunkts lässt sich hingegen bei den Jungen der geschlechtshomogenen Gruppen ein stark ausgeprägter Peergruppen-Effekt herauslesen. Dieser Effekt verweist zum Messzeitpunkt 4 möglicherweise auf markante Unterschiede zwischen Urteilen, die im Rahmen des Klingenden Fragebogen erhoben werden, und Urteilen, die im Diskurs mit anderen Jungen gefällt werden. Während der Einfluss der Peers am Ende der Grundschulzeit bei der Fragebogen-Erhebung nicht mehr so stark zum Ausdruck kommt und vermutlich eine differenziertere Beurteilung anhand musikalischer Stilmerkmale möglich ist, zeigt sich im Gespräch ein bedeutender Einfluss der Peergruppe.
Hinsichtlich der negativeren Bewertung des Faktors ‚Pop’ zum Messzeitpunkt 4 zeigen die Interviewanalysen, dass diese mit dem Aufkommen des Konzeptes ‚Chartsmusik’ zusammenfallen. Die im Faktor ‚Pop’ zusammengefassten Musikbeispiele sind speziell für die vorliegende Studien komponiert worden, so dass sie die für „Chartsmusik“ bedeutsame Aktualität und Bekanntheit gar nicht aufweisen können, was die negativere Bewertungen erklären dürfte (vgl. Busch et al., 2014b).

5 Diskussion und Fazit

Die Ergebnisse der Studie belegen einerseits eine bereits zum Schulbeginn vorhandene Fähigkeit zur Klassifizierung gehörter Musik nach verschiedenen Stilkategorien. Andererseits lassen die Ergebnisse einen generellen altersabhängigen Rückgang an Offenohrigkeit fraglich erscheinen. Vielmehr erscheint naheliegend, dass neben der bedeutenden Einflussvariable Alter eine differenzierte Betrachtung der Entwicklung kindlicher Musikpräferenz nach musikalischen Stilkategorien, nach dem Geschlecht der urteilenden Kinder sowie nach dem Erhalt von Instrumentalunterricht vorzunehmen.

Eine Limitation der vorliegenden Studie besteht darin, dass die Präferenzurteile der Kinder aufgrund der erläuterten Herausforderungen (vgl. Abschnitt 3.1, in diesem Beitrag) nicht für alle Musikstücke über die gesamte Projektlaufzeit vorgenommen werden konnte. Somit konnte die faktorielle Struktur der Messzeitpunkte nur nebeneinander abgetragen werden und in die Klassenanalysen nur eine Auswahl der

präsentierten Musikstücke eingehen. Zudem erscheint diskussionswürdig, inwieweit die Präferenzäußerungen der Kinder als Ausdruck ihrer tatsächlichen Musikpräferenz interpretiert werden können. Dies betrifft auch die (forschungspraktische) Entscheidung für Gruppenerhebungen mittels Smiley-Skalen, bei der Peergruppen-Effekte beobachtet werden konnten. Zukünftige Studien sollten dies berücksichtigen und beispielsweise als alternativen Ansatz zur Erfassung von Musikpräferenz individuell von den Kindern festzulegende Hördauern pro Musikbeispiel in Betracht ziehen (vgl. die Software *OpenEar* vorgestellt von Louven, 2011).

Im Bewusstsein dieser Limitationen lässt sich dennoch festhalten, dass sich in sämtlichen der berichteten quantitativen und qualitativen Analysen die Notwendigkeit einer Differenzierung nach Jungen und Mädchen zeigte. Dieser Geschlechtseffekt wurde bereits zu Beginn der Grundschulzeit deutlich, schwächte sich erst am Ende der Grundschulzeit ab und konnte im Wesentlichen auf die negativere Beurteilung von Musik mit ‚klassischen' Stilmerkmalen durch die Jungen zurückgeführt werden. Da Kopiez und Lehmann (2008) in ihrer Studie gerade die ‚klassischen' Musikstücke von der Analyse ausschlossen, ist es somit nicht verwunderlich, dass sie keinen Geschlechtseffekt fanden. Der beobachtete Geschlechtseffekt verwies auf einen jungenspezifischen Peergruppen-Effekt (Schurig et al., 2012). Dieser wies zwar korrelative Zusammenhänge zu der Persönlichkeitsdimension ‚Offenheit für Erfahrung' auf, aber die Ausprägung der Persönlichkeitsdimension hatte über die Konfundierung zu diesem Effekt hinaus keinen prädiktiven Wert für die musikalische Präferenz. So lässt sich der geschlechtsspezifische Effekt weniger als Ausdruck von Persönlichkeit, sondern vielmehr als Funktionalisierung von Musikpräferenzäußerungen zur Ausbildung und zum Ausdruck von (Geschlechts-)Identität deuten. Dies wird in dem Sinne verstanden, wie es North und Hargreaves (1999, S. 90) bereits für Jugendliche und Erwachsene beschrieben haben: „music functions as a ‚badge' in adolescents' social cognitions". Auch Schäfer und Sedlmeier (2009) konnten aus Musikpräferenzäußerungen verschiedene Dimensionen (u.a. evaluative und behaviorale) mit jeweils spezifischen Funktionen für das Individuum ableiten, wobei dem Identitätsausdruck eine hervorgehobene Stellung von den Befragten zugewiesen wurde. Nach den Befunden der vorliegenden Studie setzt die Nutzung von musikbezogenen Präferenzäußerungen zum Ausdruck spezifischer individueller und sozialer Funktionen bereits während der Grundschulzeit ein und somit früher als bislang vermutet (vgl. Behne, 1997; Baacke, 1993). Dass Jungen von dieser Art der Funktionalisierung in besonderer Weise Gebrauch machen, erfährt durch die Beobachtungen von Wilke (2012) Bestätigung, nach denen Jungen mit Migrationshintergrund in der vierten Grundschulklasse Präferenzäußerungen für Gangsta Rap zur Inszenierung von Männlichkeit nutzen. Erkenntnisse zur generellen Entwicklung von Geschlechtstypisierungen bestätigen ebenfalls diese geschlechtsspezifische Differenzierung, da bei Jungen die Fixierung auf die Stereotypen ‚männlich' und ‚weiblich' grundsätzlich ausgeprägter ist als bei Mädchen (Maccoby, 2000; Ruble et al., 2006; vgl. Beutler-Prahm, 2012). Es erscheint naheliegend, dass geschlechtsspezifische Unterschiede in der musikalischen Präferenzentwicklung bei Grundschulkindern nicht nur als Ergebnis unterschiedlicher Sozialisation, sondern auch als Grundlage für weitere Entwicklungs- und Sozialisationsprozesse und damit als sinnfälliger Schritt in der Entwicklung von Geschlechtsidentität zu deuten sind. Hieraus ergibt sich eine mögliche Erklärung für die von Rentfrow et al. (2011) formulierte Frage, warum Musik so bedeutsam für Menschen sei. Denn wenn Musik bereits im jungen Kindesalter die Möglichkeit zur Darstellung und Ausbildung ihrer psychosozialen (Geschlechts-)Identität bieten und somit grundlegende Funktionen der Entwicklung übernehmen kann, erscheint nachvollziehbar, dass Musik ihre Bedeutsamkeit bis ins Erwachsenenalter beibehält. Für zukünftige Studien zur Musikpädagogik werfen diese Zusammenhänge die Frage auf, wie Instrumentalunterricht konzipiert sein sollte, damit „Kulturelle/musisch-ästhetische Bildung als integraler Bestandteil individueller und sozialer Identitätsentwicklung" (Autorengruppe Bildungsberichterstattung, 2012, S. 160) zu neuen Handlungsoptionen und vielschichtigen Erfahrungen führen kann. Von besonderem Interesse wäre zudem, ob die berichteten Geschlechtseffekte sich vor allen in Gruppenerhebungen so

ausgeprägt darstellen und außerhalb von Gruppensituationen abschwächen, was sich auf Grundlage der Ergebnisse der qualitativen Erhebung vermuten ließe. Des Weiteren wäre zu überprüfen, warum sich keine persönlichkeitsspezifischen Effekte auffinden ließen. Hierbei stellt sich u.a. die Frage nach einem geeigneten Erhebungsinstrument für kindliche Persönlichkeit. Zudem sollte in zukünftigen Studien der Übergang von der dritten zur vierten Grundschulklassen detailliert in den Blick genommen werden, um die damit zusammenfallende Auflösung der faktoriellen Strukturierung nach musikalischen Stilkategorien sowie den Rückgang an geschlechtsspezifischen Präferenzunterschieden präziser fassen zu können.

Die Befunde von Wilke (2012) scheinen auf den ersten Blick Bourdieus Annahme zu stützen, dass Musikpräferenz zur Darstellung der Zugehörigkeit zu einer bestimmten gesellschaftlichen Klasse genutzt wird (u.a. Bourdieu, 1993), obgleich Migrationshintergrund keinesfalls nur in bestimmten sozialen Schichten der Gesellschaft vorkommt. In der vorliegenden Studie konnten hingegen weder in Bezug auf Migrationshintergrund noch in Bezug auf Sozialstatus Einflüsse auf die Entwicklung der kindlichen Musikpräferenz beobachtet werden. Entsprechende Effekte werden möglicherweise erst zu einem späteren Entwicklungsstadium (Eintritt in die Pubertät) deutlich, wozu das Verbundprojekt WilmA, die Folgestudie zu SIGrun, sicher aufschlussreiche Hinweise liefern wird. Gesellschaftlicher Status zeigt sich jedoch auch im Erhalt von privatem Instrumentalunterricht. So haben u.a. Nonte und Schwippert (2014) dargelegt, dass Kinder mit erhöhtem Risikopotenzial – wozu unter anderem gesellschaftliche Distinktionsfaktoren wie höchster familiärer Bildungsabschluss oder das jährliche Haushaltsnettoeinkommen beitragen – weniger häufig privaten Instrumentalunterricht erhalten, während sich beim Schulprogramm JeKi in dieser Hinsicht keine Unterschiede zeigen. In der vorliegenden Studie sind die Befunde hinsichtlich des Einflusses von Instrumentalunterricht auf die musikbezogene Präferenzentwicklung uneinheitlich, so dass frühere Publikationen (Schurig et al., 2012; Busch et al., 2013; 2014a; 2014b) präzisiert werden müssen. Während bei der faktoriellen Strukturierung der Musikstücke nach den Präferenzurteilen keine Beeinflussung festgestellt werden konnte, wurde in der Gruppierung der beurteilenden Kindern zu latenten Klassen auf Basis ihrer Präferenzurteile am Ende der Grundschulzeit ein Einfluss von privat organisiertem Instrumentalunterricht sichtbar, nicht aber von schulischem Instrumentalunterricht (wie JeKi). In einer ergänzenden Analyse (Schurig et al., 2012) konnte beobachtet werden, dass Kinder mit privat organisiertem Instrumentalunterricht häufig weniger lange an JeKi teilnahmen. In einer Prädiktionsanalyse für den Erhalt von Instrumentalunterricht (Nagelkerke geschätzer $R^2 = 0.452$) mit den weiteren Variablen Bildungsaspiration der Eltern (*Odds-Ratio* 10.4; $p < 0.001$) und aktives Musizieren der Eltern (*Odds-Ratio* 2.8; $p < 0.001$) ergab sich für die vorliegende Stichprobe ein *Odds-Ratio* von 0.03 ($p < 0.006$) dafür, den JeKi Unterricht bis zur vierten Klasse zu besuchen. Dies impliziert, dass sich die Wahrscheinlichkeit, ein Instrument außerhalb der Schule zu erlernen, stark verringerte, wenn eine JeKi-Teilnahme bis zur vierten Jahrgangsstufe erfolgte. Dieses oberflächlich ungünstig erscheinende Ergebnis lässt sich aber positiv interpretieren. So ist zu vermuten, dass die Kinder aus JeKi ausscheiden, die ein Instrument in einem privaten Kontext erlernen, aber jene in JeKi verbleiben, die ansonsten keinen privat organisierten Instrumentalunterricht erhalten würden. Somit wäre das Programmziel erfüllt und zugleich auf die Notwendigkeit verwiesen, die Frage nach Wirkungen von JeKi auch auf der Ebene von Kultureller Teilhabe zu analysieren (vgl. Kap. V. Kulturelle Teilhabe, in diesem Band).

Die latente Klassenanalyse führte zu allen vier Messzeitpunkten zu einer drei Klassen-Lösung, wobei sich die latenten Klassen nach der Art ihrer Präferenzen unterschieden. Diese drei Gruppen wurden folgermaßen umschrieben: Offene Hörer, die unkonventionelle Musikstücke tendenziell positiv beurteilten; Indifferente Hörer, deren Urteile eher im neutralen Bereich angesiedelt waren; und Verschlossenere Hörer, die vor allem deutliche Abneigungen, aber auch spezifische Zuneigungen erkennen ließen. Im Verlauf der Grundschulzeit wurde die Gruppe der Offenen Hörer kleiner. Zudem

zeigte sich, dass Kinder mit privat organisiertem Instrumentalunterricht generell weniger indifferente Urteile fällten und somit eine höhere Bereitschaft hatten, sich eindeutig in ihrer Präferenz positiv oder negativ zu positionieren. Dies lässt sich möglicherweise durch die häufigeren Gelegenheiten zu musikbezogenen Erfahrungen sowie den Zuwachs an musikalischer Expertise und an musikalischem Stilempfinden erklären, was vermutlich auch mit einem Anstieg im musikbezogenen Selbstkonzept einhergeht und infolge eine deutlichere Präferenz-Positionierung befördern könnte. Während also bei der konkreten Beurteilung einzelner Musikbeispiele kein Einfluss von Instrumentalunterricht (weder privat noch schulisch) beobachtet wurde, zeigte sich hinsichtlich des generellen Urteilsverhaltens ein Einfluss des privaten Instrumentalunterrichts. Warum sich dieser Effekt nicht auch bei schulischem Instrumentalunterricht (wie JeKi) beobachten ließ, steht zur Diskussion – zumal JeKi einen positiven Einfluss auf die Entwicklung des musikalischen Selbstkonzeptes ausübt (Nonte, 2013). Möglicherweise sind bei der Ausbildung eines differenzierten Urteilsvermögens also doch Aspekte wie musikalische Expertise und musibezogenes Stilempfinden bedeutsamer, die aber im JeKi-Programm eventuell weniger stark im Vordergrund stehen als die sozialen Aspekte des gemeinsamen Musizierens. Bei anderen Schulprogrammen zum Instrumentalunterricht haben sich in Bezug auf Musik mit ‚klassischen' Stilmerkmalen ähnliche Effekte gezeigt, wie sie hier für den privat organisierten Instrumentalunterricht beschrieben wurden (Louven, 2011). Möglicherweise wird diese Entwicklung zusätzlich durch eine größere Vertrautheit mit ‚klassischer' Musik und einem ausgeprägteren musikalischen Stilempfinden befördert, so dass eine differenziertere und von sozialen Einflüssen unabhängigere Beurteilung vorgenommen werden kann.

Interessant ist jedoch, dass sich in der vorliegenden Studie die drei Gruppen an Kindern nicht nur bei den ‚klassischen' Musikstücken herauskristallisieren, sondern auch bei den Stücken des Faktors ‚Ethno/Avantgarde'. Diese Musikstücke wurden am Ende der Grundschulzeit als einzige Musikstücke wirklich negativ beurteilt, also unterhalb des Skalenmittelwertes von 3, so dass das „Verschwinden" an Offenohrigkeit nur für eben diese Musikstücke deutlich wurde. Dies könnte darin begründet sein, dass die Kinder diese Musikstücke durch ihr generell gestiegenes musikalisches Stilempfinden zunehmend als ‚andersartig' erkannten und somit im Sinne des eigenen Identitätsausdruckes eher nicht als zu sich zugehörig erlebten und daher ablehnten. Die Präferenzentwicklung für die ‚Klassik'-Stücke ließ jedoch vermuten, dass eine Verstärkung der musikbezogenen Erfahrungen mit Musik anderer Kulturen ebenfalls dazu beitragen könnte, diese Musik zwar weiterhin als ‚besonders' zu erkennen, sie aber in ihrer eigenen Art wertzuschätzen und somit differenzierter zu beurteilen.

Die Herausforderung für die Musikpädagogik wird somit darin gesehen, einerseits Musikpräferenz als einen Teil kulturellen Verhaltens wahrzunehmen, der bedeutende individual- und sozialpsychologische Funktionen in der kindlichen Entwicklung übernehmen kann. Andererseits aber sollte im Rahmen von musikalischen Angeboten zugleich die Chance genutzt werden, die Ausbildung einer differenzierten und von psycho-sozialen Aspekten weniger abhängigen musikbezogenen Urteilsfähigkeit zu befördern. Musikpräferenz könnte somit zwischen individuell sinnfälliger Funktionalisierung und kritischer Urteilsfähigkeit oszillieren, dadurch – durchaus im Sinne von Bourdieu – Handlungsspielräume erweitern und Offenheit für neue musikalische Erfahrungen aufrecht erhalten oder generieren.

Literaturverzeichnis

Autorengruppe Bildungsberichterstattung (2012). Bildung in Deutschland 2012. Ein indikatorengestützter Bericht mit einer Analyse zur kulturellen Bildung im Lebensverlauf. Bielefeld: Bertelsmann.

Baacke, D. (1993). Jugendkulturen und Musik. In H. Bruhn, R. Oerter & H. Rösing (Hrsg.), Musikpsychologie. Ein Handbuch (S. 228-237). Reinbek: Rowohlt.

Bacher, J. & Vermunt, J. K. (2010). Analyse latenter Klassen. In C. Wolf & H. Best (Hrsg.), Handbuch der Sozialwissenschaftlichen Datenanalyse. (S. 543-574). Wiesbaden: Verlag für Sozialwissenschaften.

Baumann, M. P. (Hrsg.) (1985). Musik der Türken in Deutschland. Kassel: Verlag Yvonne Landeck.

Behne, K.-E. (1975). Musikalische Konzepte. Zur Schicht- und Altersspezifität musikalischer Präferenzen. In E. Kraus (Hrsg.), Forschung in der Musikerziehung (S. 35-61). Mainz: Schott.

Behne, K.-E. (1986). Hörertypologien. Zur Psychologie jugendlichen Musikgeschmacks. Regensburg: Bosse.

Behne, K.-E. (1987). Urteile und Vorurteile: Die Alltagsmusiktheorien Jugendlicher Hörer. In H. de la Motte-Haber (Hrsg.), Psychologische Grundlagen des Musiklernens, Handbuch der Musikpädagogik, Bd. 4 (S. 221-272). Kassel: Bärenreiter.

Behne, K.-E. (1993). Musikpräferenzen und Musikgeschmack. In H. Bruhn, R. Oerter & H. Rösing (Hrsg.), Musikpsychologie. Ein Handbuch (S. 339-353). Reinbek: Rohwolt.

Behne, K.-E. (1997). The development of „Musikerleben“ in adolescence. How and why young people listen to music. In I. Deliège & J. A. Sloboda (Hrsg.), Perception and Cognition of Music (S. 143-159). Hove, UK: Psychology Press.

Beutler-Prahm, B. (2012). Geschlechtsspezifische Aspekte in der musikalischen Präferenz bei Grundschulkindern. Unveröffentlichte Bachelor-Arbeit, Universität Bremen.

Bourdieu, P. (1983). Ökonomisches Kapital, kulturelles Kapital, soziales Kapital. In R. Kreckel (Hrsg.), Soziale Ungleichheit (S. 183-198). Göttingen: Schwartz.

Bourdieu, P. (1993 [1980]). Soziologische Fragen. Frankfurt / M.: Suhrkamp.

Bundesministerium für Bildung und Forschung (BMBF) (Hrsg.) (2006). Macht Mozart schlau? Die Förderung kognitiver Kompetenzen durch Musik. Bildungsforschung: Bd. 18. Bonn, Berlin.

Bundesministerium für Bildung und Forschung (Hrsg.) (2009). Pauken mit Trompeten. Lassen sich Lernstrategien, Lernmotivation und soziale Kompetenzen durch Musikunterricht fördern? Bildungsforschung, Bd. 32. Berlin.

Bollen, K. A. (2002). Latent Variables in Psychology and the Social Sciences. Annual Review of Psychology, 53, 605-34.

Bunte, N. (2013). Die Entwicklung musikalischer Konzepte im Grundschulalter und ihre Bedeutung für kindliche Musikpräferenzen. Exposé zum Promotionsvorhaben, unveröffentlichtes Manuskript, Universität Bremen.

Busch, V. (2005). Tempoperformance und Expressivität. Eine Studie zwischen Musikpsychologie und Musiktherapie. Frankfurt a.M., Peter Lang Verlag.

Busch, V., Lehmann-Wermser, A. & Liermann, C. (2009). The Influence of Music Genre, Style of Singing, and Gender of Singing Voice on Music Preference of Elementary School Children. In J. Louhivuori et al. (Hrsg.). Proceedings of the 7th Triennial Conference of European Society for the Congitive Sciences of Music (ESCOM) in Jyväskylä, 2009 (S. 33-37), Jyvväskylä, Finland.

Busch, V., Schurig, M., Bunte, N. (2013). Mädchen- oder Jungenmusik? JeKi und die Entwicklung musikalischer Präferenzen im Grundschulalter. In der Broschüre der Koordinierungsstelle des BMBF-Forschungsschwerpunkts zu Jedem Kind ein Instrument (Hrsg.), Empirische Bildungsforschung zu Jedem Kind ein Instrument (S. 52-54). Bielefeld: Universität Bielefeld.

Busch, V., Schurig, M., Bunte, N. & Beutler-Prahm, B. (2014a). Teilprojekt "Präferenz" – Entwicklung musikbezogener Präferenz von Grundschulkindern. Im Abschlussband des BMBF-Forschungsschwerpunktes zu JeKi. Manuskript im Druck.

Busch, V., Schurig, M., Bunte, N. & Beutler-Prahm, B. (2014b). „Mir gefällt ja mehr diese Rockmusik.“ Zur Struktur der Präferenzurteile im Grundschulalter. Jahrbuch Musikpsychologie, Bd. 24 (Themenband zur Offenohrigkeitsforschung). Manuskript im Druck.

Costa, P. T. & McCrae, R. R. (1992). Revised NEO Personality Inventory (NEO PI-R) and NEO Five Factor Inventory. Professional Manual. Odessa, FL: Psychological Assessment Ressources.

Cremades, R., Oswaldo, L., & Lucia, H. (2010). Musical tastes of secondary school students with different cultural backgrounds: A study in the spanish north african city of Melilla. Musicae Scientiae, 14 (1), 121-141.

Delsing, M. J. M. H., Bogt, T. F. M. T., Engels, R. C. M. E. & Meeus, W. H. J. (2008). Adolescents' music preferences and personality characteristics. European Journal of Personality, 22, 109-130.

Eijck, K. van (2001). Social differentiation in musical taste patterns. Social Forces, 79 (3), 1163-1184.

Erzberger, C. &. Kelle U. (2003). Making inferences in mixed methods: The rules of Integration. In A. Tashakkori und C. Teddlie (Hrsg.), Handbook of mixed methods in social & behavioral research. Thousand Oaks, Calif. [u.a.]: SAGE.

Gembris, H. (2005). Musikalische Präferenzen. In R. Oerter & T. H. Stoffer (Hrsg.), Enzyklopädie der Psychologie, Vol. 2, Spezielle Musikpsychologie (S. 279-342). Göttingen: Hogrefe.

Gembris, H., & Schellberg, G. (2003). Musical preferences of elemantary school children. Paper presented at the 5th Triennial Conference of the European Society for the Cognitive Sciences of Music. Hannover, 2003.

Gembris, H. & Schellberg, G. (2007). Die Offenohrigkeit und ihr Verschwinden bei Kindern im Grundschulalter. Musikpsychologie, 19, 71-92.

Greve, M. (2003). Die Musik der imaginären Türkei. Musik und Musikleben im Kontext der Migration aus der Türkei in Deutschland. Stuttgart, Weimar: J. B. Metzler.

Hargreaves, D. J. (1982). The development of aesthetic reactions to music. Psychology of Music (Special issue), 51-54.

Hargreaves, D. J. (1987). Development of liking for familiar and unfamiliar melodies. Bulletin of the Council for Research in Music Education, 91, 65-69.

Hargreaves, D. J., Comber, C. & Colley, A. (1995). Effects of age, gender, and training on musical preferences of British secondary school students. Journal of Research in Music Education, 43 (3), 242-250.

Hargreaves, D. J., North, A. C., & Tarrant, M. (2006). Musical preference and taste in childhood and adolescence. In G. E. McPherson (Hrsg.), The child as musician: A handbook of musical development (S. 135-154). New York: Oxford University Press.

Henninger, J. C. (1999). Ethnically diverse sixth graders' preferences for music of different cultures. Texas Music Education Research, 37-43.

Hu, L. & Bentler, P. M. (1998). Fit Indices in Covariance Structure Modeling: Sensitivity to Underparameterized Model Misspecification. Psychological Methods, 3, 424-453.

Iyer, V. (2002). Embodied mind, situated cognition, and expressive microtiming in African-American music. Music Perception 19(3): 387-414.

Jaedtke, W. (2000). Popmusik als Epochenstil. Versuch einer musikhistorischen und musiktheoretischen Aufarbeitung. In H. Rösing & T. Phleps (Hrsg.). Populäre Musik im kulturwissenschaftlichen Diskurs [=Beiträge zur Popularmusikforschung 25/26] (S. 201-216), Karben: Coda.

Kleinen, G. (2011). Musikalische Sozialisation. In H. Bruhn, R. Kopiez & A. C. Lehmann (Hrsg.), Musikpsychologie. Das neue Handbuch (S. 37-66). Reinbek: Rowohlt.

Kopiez, R., & Lehmann, M. (2008). The 'open-earedness' hypothesis and the development of age-related aesthetic reactions to music in elementary school children. British Journal of Music Education, 25 (2), 121-138.

Kulin, S. & Schwippert, K. (2012). Kooperationsbeziehungen im JeKi-Kontext: Beweggründe zur Kooperation und Merkmale gemeinsamer Reflexion methodischer und didaktischer Fragen. In J. Knigge & A. Niessen (Hrsg.), Musikpädagogisches Handeln. Begriffe, Erscheinungsformen, politische Dimensionen [= Musikpädagogische Forschung 33] (S. 152-171). Essen: Die Blaue Eule.

LeBlanc, A. (1991). Some unanswered questions in music preference research. Contribution to Music Education, 18, 66-73.

LeBlanc, A., Sims, W. L., Siivola, C., & Obert, M. (1996). Music style preferences of different age listeners. Journal of Research in Music Education, 44 (1), 49-59.

Lehmann-Wermser, A. & Jessel-Campos, C. (2013). Aneignung von Kultur. Wege zu kultureller Teilhabe und zur Musik. In A. Hepp & A. Lehmann-Wermser (Hrsg.), Transformation des Kulturellen. Prozesse des gegenwärtigen Kulturwandels (S. 129-144). Wiesbaden: vs-Verlag.

Lenz, F. (2013). Soziologische Perspektiven auf musikalische Sozialisation. In In Heyer, R., Wachs, S. & Palentien, C. (Hrsg.) (2013). Handbuch Jugend – Musik – Sozialisation (S. 157-185). Wiesbaden: Springer.

Leopold, E. (2013). Urteilshomogenität und Klassengemeinschaft – Ein Beitrag zur Offenohrigkeitshypothese. Musikpsychologie, 22, 74-90.

Linzer, D. & Lewis, J. (2013). poLCA: Polytomous variable Latent Class Analysis (R Paket), http://poLCA.r-forge.r-project.org.

Lopez Cano (2003). Setting the body in music: Gesture, schemata and stylistic-cognitive types. Paper given at the International Conference Music and Gesture, University East Anglia, Norwich.

Louven, C. (2011). Mehrjähriges Klassenmusizieren und seine Auswirkungen auf die „Offenohrigkeit" bei Grundschulkindern. Eine Langzeitstudie. Diskussion Musikpädagogik, 50 (11), 48-59.

Louven, C. & Ritter, A. (2011). Hargreaves' „Offenohrigkeit" – Ein neues, softwarebasiertes Forschungsdesign. Beitrag zur AMPF-Tagung 2011 in Stuttgart (S. 275-299).

Maccoby, E. (2000). Psychologie der Geschlechter. Sexuelle Identität in den verschiedenen Lebensphasen. Translated by E. Vorspohl. Stuttgart: Klett-Cotta.

Maydeu-Olivares A. & Böckenholt, U. (2009). Modeling Preference Data. In: R. Millsap & A. Maydeu-Olivares (Hrsg.), The SAGE Handbook of Quantitative Methods in Psychology (S. 264–282) London: SAGE Publications.

Mayring, P. (2007). Qualitative Inhaltsanalyse. Grundlagen und Techniken. Weinheim: Beltz.

MacDonald, R., Hargreaves, D. J. & Miell, D. (2002). Musical Identities. Oxford: Oxford University Press.

Ministerium für Schule und Weiterbildung des Landes Nordrhein-Westfalen (Hrsg.) (2008). Richtlinien und Lehrpläne für die Grundschule in Nordrhein-Westfalen, Heft 2012, 1. Auflage.

MPLUS (Version 6.11). [Computer Software]. Los Angeles, CA: Muthén & Muthén.

Müller, R. (1999). Musikalische Selbstsozialisation. In J. Fromme, S. Kommer, J. Mansel & K.-P. Treumann (Hrsg.), Selbstsozialisation, Kinderkultur und Mediennutzung (S. 113-125). Oladen: Leske + Budrich.

Nonte, S. (2013). Herausforderungen und Probleme bei der Entwicklung eines Instruments zur Selbsteinschätzung musikalischer Fähigkeiten im Grundschulalter. *Beiträge empirischer Musikpädagogik, 4* (2), 1-30.

Nonte, S. & Schwippert, K. (2012). Musikalische und sportliche Profile an Grundschulen – Auswirkungen auf Klassenklima und Selbstkonzept. *Beiträge empirische Musikpädagogik, 3* (1), 1-25.

Nonte, S. & Schwippert, K. (2014). Teilprojekt "Transfer" – Effekte von JeKi-Programmen auf die Entwicklung sozialer und motivationaler Aspekte von Kindern mit kumulierten Risikofaktoren. Manuskript in Druck.

North, A. C. & Hargreaves, D. J. (1999). Music and adolescent identities. Music Education Research, 1, 75-92.

Peterson, R. A. (1992). Understanding audience segmentation: From elite and mass to omnivore and univore. Poetics, 21, 243-258.

Peterson, R. A., & Simkus, A. (1992). How musical tastes mark occupational status groups. In M. Lamont & M. Fournier (Hrsg.), Cultivating differences. Symbolic boundaries and the making of inequality. Chicago: Chicago Press.

Rauscher, F. H., G. L. Shaw & K. N. Ky (1995), Listening to Mozart enhances spatial-temporal reasoning: towards a neurophysiological basis, in: Neuroscience Letters 1285, S. 44-47.

Rawlings, D., & Ciancarelli, V. (1997). Music preference and the five-factor model of the NEO Personality Inventory. Psychology of Music, 25, 120-132.

Reinecke, J. (2005). Strukturgleichungsmodelle in den Sozialwissenschaften, München: Oldenbourg.

Reinhardt, J. & Rötter, G. (2013). Musikpsychologischer Zugang zur Jugend-Musik-Sozialisation. In Heyer, R., Wachs, S. & Palentien, C. (Hrsg.) (2013). Handbuch Jugend – Musik – Sozialisation (S. 127-155). Wiesbaden: Springer.

Rentfrow, P. J. & Goslings, S. D. (2003). The do re mi's of everyday life: The structure and personality correlates of music preference. Journal of Personality and Social Psychology, 84, 1236-1256.

Rentfrow, P. J., Goldberg, L. R. & Levitin, D. J. (2011). The structure model of musical preferences: A five-factor model. Journal of Personality and Social Psychology, 100 (6), 1139-1157.

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

Ruble, D. N., Martin, C. L. & Berenbaum, S. A. (2006). Gender development. In W. Damon & R. M. Lerner (Hrsg.), Handbook of Child Psychology, Vol. 3, 6th ed. (S. 858-932). Hoboken: Wiley.

Sakai, W. (2011). Music preferences and family language background: A computer-supported study of children's listening behavior in the context of migration. Journal of Research in Music Education, 59 (2), 174-195.

Schäfer, T., & Sedlmeier, P. (2009). From the functions of music to music preference. Psychology of Music, 37, 279-300.

Schellberg, G. (2006). Zum Einfluss von Unterricht auf Musikpräferenzen von Grundschulkindern für Opernarien. In N. Knolle (Ed.), Lehr- und Lernforschung (S. 71-84). Essen: Die blaue Eule.

Schellenberg, E.G. (2006). Long-term positive associations between music lessons and IQ. Journal of Educational Psychology, 98, 457-468.

Schurig, Michael (2012). Response Bias und Messinvarianz in einem Urteil zu musikalischer Präferenz. Hinter der Messinvarianz. Vortrag gehalten auf der 77. Tagung der Arbeitsgruppe für Empirische Pädagogische Forschung in Bielefeld, 2012.

Schurig, M., Busch, V. & Strauß, J. (2012). Effects of Structural and Personal Variables on Children's Development of Music Preference. In E. Cambouropoulos, C. Tsougras, P. Mavromatis & K. Pastiadis (Hrsg.), Proceedings of the 12th International Conference on Music Perception and Cognition (ICMPC) and the 8th Triennial Conference of European Society for the Cognitive Sciences of Music (ESCOM) in Thessaloniki, 2012 (S. 896-902).

Statistisches Bundesamt (2009). Bevölkerung und Erwerbstätigkeit. Ausländische Bevölkerung. Ergebnisse des Ausländerzentralregisters, Fachserie 1, Reihe 2. Wiesbaden [verfügbar unter: https://www.destatis.de/DE/Publikationen/Thematisch/Bevoelkerung/MigrationIntegration/AuslaendBevoelkerung 2010200087004.pdf?__blob=publicationFile].

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

Teo, T., Hargreaves, D. J., & Lee, J. (2008). Musical preference, identification, and familiarity: A multicultural comparison of secondary students from Singapore and the United Kingdom. Journal of Research in Music Education, 56 (1), 18-32.

Vermunt, J. K. (2010). Latent Class Models. En: E. Baker, P. Peterson & B. McGaw (Hrsg.), *International Encyclopedia of Education*, Band 7 (S. 238-244). Oxford: Elsevier, 3. Auflage.

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

Wilke, K. (2012). Bushido oder bunt sind schon die Wälder?! Musikpräferenz von Kindern in der Grundschule. Münster: Lit.

Wurm, M. (2006). Musik in der Migration. Beobachtungen zur kulturellen Artikulation türkischer Jugendlicher in Deutschland. Bielefeld: transcript.

Zweigenhaft, R. (2008). A Do Re Mi Encore. A Closer Look at the Personality Correlates of Music Preference. Journal of Individual Differences, 29 (1), 45-55.

BEITRAG 3

Erschienen in (Zitierweise):

Busch, V., Schurig, M., Bunte, N. & Beutler-Prahm, B. (2015). „Mir gefällt ja mehr diese Rockmusik." Zur Struktur musikalischer Präferenzurteile im Grundschulalter. In: W. Auhagen, C. Bullerjahn & R. von Georgi (Hrsg.), *Offenohrigkeit - Ein Postulat im Fokus (Jahrbuch der Deutschen Gesellschaft für Musikpsychologie - Band 24)* (S. 133-168). Göttingen [u.a.]: Hogrefe.

Relevanz:

Der Beitrag beschätigt sich mit der lägsschnittlichen Validierung eines Modells und der Prädiktion von Dimensionen der Musikpräferenz von Grundschülern. Dieser Beitrag ergänzt ein Exempel für den Erklärungswert von methodologisch gemischten Ansätzen. Aufbauend auf den Ergebnissen des Beitrags 1 sowie auf Basis einer erweiterten Stichprobe und Datengrundlage wurde eine mehrdimensionale konfirmatorische Struktur auf die nun vollständigen vier Messzeitpunkte übertragen und geprüft. Es wird beobachtet, dass sich die beobachtete faktorielle Struktur zum vierten Messzeitpunkt auflöst. Die Struktur dieses Messzeitpunkts wird gesondert analysiert. Für die ersten drei Messzeitpunkte, in denen die abgeleitete Struktur als hinreichend belastbar bewertet wurde, werden verschiedene Prädiktoren verarbeitet, um deren Relevanz für die einzelnen Faktoren innerhalb der Messzeitpunkte zu bestimmen. Da die Befunde inhaltlich nicht aus dem quantitativen Datenmaterial erklärt werden konnten, wurde eine begleitende qualitative Studie ergänzt, so dass erklärende Aussagen abgeleitet werden können.

„Mir gefällt ja mehr diese Rockmusik.“

Zur Struktur musikalischer Präferenzurteile im Grundschulalter

Veronika Busch[1], Michael Schurig[2], Nicola Bunte[1] und Bettina Beutler-Prahm[1]
[1]Universität Bremen, [2]Technische Universität München

Keywords:

Musikalische Präferenz, Offenohrigkeit, Faktoranalyse, Mixed-Method, Geschlecht, Grundschulzeit

Diese Forschungsarbeit ist Teil des Verbundprojektes „SIGrun – Studie zum Instrumentalunterricht an Grundschulen“, das von den Universitäten Hamburg und Bremen durchgeführt und vom Bundesministerium für Bildung und Forschung im Rahmen des Forschungsprogramms zu „JeKi – Jedem Kind ein Instrument“ finanziert wurde.

Zusammenfassung

Hargreaves (1982) knappe Hypothese einer altersbedingten Abnahme kindlicher Offenheit gegenüber unbekannten und unkonventionellen Musikstücken (sogenannte Offenohrigkeit, „open-earedness“) bildet den Hintergrund der vorliegenden explorativen Längsschnittstudie mit vier Messzeitpunkten im Verlauf der ersten vier Grundschuljahre. Die beteiligten Schülerinnen und Schüler beanworteten einen Klingenden Fragebogen mit 16 Musikbeispielen auf einer fünfstufigen ikonographischen Rating-Skala. Strukturelle und personelle Daten wurden mittels standardisierter Fragebögen erhoben, leitfadengestützte Interviews ergänzen den Datensatz. Auf der Basis von Faktorenanalysen wird Offenohrigkeit über die latenten Faktoren „Klassik“, „Pop“ und „Ethno/Avantgarde“ (vgl. Louven, 2011) operationalisiert. Das Ziel der Studie ist die Ableitung von Messmodellen für längsschnittliche Strukturgleichungsanalysen, an denen Prädiktorvariablen (z.B. Geschlecht, Alter, Instrumentalunterricht, Persönlichkeitsdimensionen, Migrationshintergrund, sozio-ökonomischer Status) identifiziert und deren Effekte getestet werden können. Die Ergebnisse legen nahe, dass Kinder bereits im ersten Schuljahr über ein musikspezifisches Kategoriensystem für Präferenzurteile verfügen und im Verlauf der ersten drei Schuljahre eine stetige Abnahme an Offenohrigkeit aufweisen. Im vierten Schuljahr löst sich hingegen die Faktorstruktur vermutlich aufgrund zunehmend individueller Musikpräferenzen auf. Neben diesem Alterseffekt zeigt sich ein ausgeprägter geschlechtsspezifischer Effekt, während den anderen unabhängigen Variablen keine bedeutende Vorhersagekraft für die Faktoren zukommt. Die Analyse der qualitativen Interviewdaten unterstützt die Vermutung, dass musikbezogene Präferenzäußerungen von den Kindern als ein Mittel zur Darstellung ihrer psychosozialen (Geschlechts-) Identität genutzt werden.

Abstract

Hargreaves' (1982) brief hypothesis of an age-related decline in children's preference for unfamiliar and unconventional pieces of music („open-earedness") forms the theoretical background of our exploratory longitudinal study with four points of measurement between grade one and four. Primary school children answered a sound questionnaire with 16 music examples on a 5-point iconic rating scale. Structural and personal data was collected using standardized questionnaires, and complementary guided interviews were conducted. Open-earedness is operationalized as a construct with „classic", „pop", and „ ethnic/avant-garde" music preference (cf. Louven, 2011) as distinguishable latent factors through factor analyses. This way measurement models for the investigation of longitudinal data will be assessed using structural equation modelling. The aim is to derive a measurement model that can be used to identify and test predictor variables (e.g. age, sex, music tuition, personality, music experience, migration background, and socio-economic status). The analyses indicate that already one year pupils possess a music specific categorial system for their preference ratings. During the first three years of primary school a decline of open-earedness was observed, while during the fourth year the factor structure dissolves – probably due to increasing individual preferences. In addition to this age effect, a strong effect for sex was found, while all the other independent variables showed no relevant predictive power for the factors. The analysis of the interviews supports the assumption that music preference ratings are used by the children as a means to indicate their psycho-social (gender-) identity.

1 Einleitung

Musikpräferenz war in den vergangenen Jahrzehnten Gegenstand vielfältiger Untersuchungen. Mittlerweile liegen verschiedene Modelle zur Gruppierung musikalischer Präferenzäußerungen für eine Vielzahl musikalischer Beispiele unterschiedlicher musikalischer Genres vor. So gelangt Louven (2011) beispielsweise analog zu Gembris und Schellberg (2007) zu einer Differenzierung der Präferenzen von Grundschulkindern nach den stilistisch interpretierbaren Kategorien „Pop“, „Klassik“ und „Ethno/Avantgarde“. Rentfrow et al. (2011) stellen hingegen das „genre-free“ Five-Factor-Model MUSIC („Mellow“, „Unpretentious“, „Sophisticated“, „Intense“, and „Contemporary“) als Rahmen für zukünftige Forschung zur Diskussion. Ein interessanter Aspekt dieses Modells ist, dass es sowohl musikalische Eigenschaften als auch deren psychische Effekte umfasst. Rentfrow et al. (2011) argumentieren für multiple Einflüsse (wie psychologische Disposition, soziale Interaktion, Umgang mit populären Medien, kultureller Trend) auf musikalische Präferenz, und sie verweisen auf die Eigentümlichkeit, dass wir zwar um die enorme Bedeutung von Musik für Menschen wissen, „[c]uriously, however, very little is known about why music is so important“ (Rentfrow et al., 2011, S. 1155). Schäfer und Sedlmeier (2009) bieten auf diese Frage eine mögliche Antwort, indem sie durch ihre Studien vielfältige Funktionen von Musik untersucht haben und ein theoretisches Modell zur Entwicklung der musikalischen Präferenz im Verhältnis zu den individuellen Funktionen von Musik vorschlagen. Als ein Lebensabschnitt, in dem Musik bedeutende Funktionen bei der individuellen Identitätsbildung zu übernehmen scheint und musikalisches Erleben durch ein hohes Maß an Emotionalisierung gekennzeichnet ist, gilt die Pubertät (vgl. Gembris, 2005; Behne, 1986; 1997). Bei jüngeren Kindern vermutete Hargreaves (1982) eine größere Offenheit gegenüber solchen Musikstücken, die Erwachsene als unkonventionell ansehen. Hargreaves umschreibt diese ursprüngliche Offenheit mit dem Begriff „open-eared“ (Hargreaves, 1982, S. 51). Diese „open-earedness-hypothesis“ (vgl. Kopiez und Lehmann, 2008, S.122) bildet auch den theoretischen Hintergrund für die vorliegende Längsschnittstudie zur Entwicklung von Musikpräferenz bei Grundschulkindern.

Im Folgenden wird der Forschungsstand zu jenen potentiellen Einflussfaktoren zusammengefasst, die für die vorliegende Studie von Bedeutung sind. Bisherige Forschung

unterstützt eine Altersabhängigkeit von Offenohrigkeit (Hargreaves et al., 2006). LeBlanc (1991; LeBlanc et al., 1996) differenziert vier Stadien altersbezogener Unterschiede im Verlauf einer Lebensspanne, u.a. eine anfängliche Offenohrigkeit junger Kinder sowie der Verlust an Offenohrigkeit während der Pubertät. Während einige Autoren bereits im Laufe der Grundschulzeit (Gembris & Schellberg, 2003) oder sogar in der Vorschulzeit (Hargreaves, 1987) von einer Abnahme an Offenohrigkeit ausgehen, sehen Kopiez und Lehmann (2008) die gesamte Grundschulzeit als Periode der Offenohrigkeit an.

Kontrovers wird auch die Diskussion um Einflüsse des Geschlechts der Rezipienten geführt. Eine Vielzahl an Studien liefert Hinweise auf geschlechtsspezifische Unterschiede, wobei Mädchen gegenüber Jungen zumeist als offenohriger beschrieben werden (Hargreaves et al., 1995; Gembris & Schellberg, 2003, 2007; Busch et al., 2009). Kopiez und Lehmann (2008) fanden hingegen nur sehr geringe geschlechtsspezifische Effektgrößen.

Für Jugendliche und Erwachsene liegen zudem Studien vor, die einen Einfluss von Persönlichkeitsfaktoren auf die Musikpräferenz nahelegen (Delsing et al., 2008; Rawlings & Ciancarelli, 1997). Inwieweit auch bei der kindlichen Musikpräferenz Persönlichkeitsfaktoren von Bedeutung sind, ist bislang nicht erforscht, was vermutlich auch in der enormen Herausforderung der Untersuchung von in der Entwicklung befindlicher Persönlichkeit begründet ist. Häufig wird ein kritisches Zeitfenster beschrieben, in dem junge Kinder sensibler auf Einflüsse auf ihre sich ausbildenden musikalischen Vorlieben reagieren (Gembris, 2005; Hargreaves et al., 2006). So ließe sich die Offenohrigkeit junger Kinder quasi als ein Indikator für dieses kritische Zeitfenster interpretieren und sogar als Ausdruck des Persönlickeitsmerkmales „openness for experience“ (Costa & McCrae, 1992) deuten. In Deuschland wurden verschiedene Programme initiiert, um Kindern während dieses kritischen Zeitfensters den Zugang zum Erlernen eines Musikinstrumentes sowie die Erfahrung mit musikalischen Stilen jenseits der aktuellen Rock-/Pop-Musik zu eröffnen (u.a. „JeKi – Jedem Kind ein Instrument“).

In Bezug auf den Einfluss von musikpraktischen Erfahrungen, wie sie im Rahmen von privatem oder schulischem Instrumental- bzw. Gesangsunterricht erworben werden können, liegen bereits Hinweise auf Zusammenhänge mit der Entwicklung von Musikpräferenz vor. So konnten u.a. Hargreaves et al. (1995) bei Jugendlichen einen positiven Zusammenhang zwischen „level of [musical] training and preference for ‚serious’ style categories“

nachweisen (Hargreaves et al., 1995, S. 248). Dies wird durch Louven (2011) bekräftigt, dessen Studie nahelegt, dass schulischer Instrumentalunterricht der typischen altersabhängigen Abnahme an Offenohrigkeit im Bereich dieser „serious“ Stilkategorien entgegenwirken könne.
Des Weiteren lassen sich Studien zum Einfluss des Migrationshintergrundes auf die Musikpräferenz dahingehend zusammenfassen, dass Jugendliche und Erwachsene die Musik ihres jeweiligen Herkunftslandes bevorzugen (Sakai, 2011; Cremades et al., 2010; Henninger, 1999; Teo et al., 2008) und die Loslösung vom elterlichen Musikgeschmack bei Kindern mit Migrationshintergrund verzögert ist (Greve, 2003; Wurm, 2006; Baumann, 1985).
Als ein weiterer Einflussfaktor wurde der sozio-ökonomische Status beschrieben, wobei Peterson (1992) bei Mitgliedern einer höheren Statusgruppe gegenüber Mitgliedern einer niedrigeren Statusgruppe generell die Wertschätzung einer größeren Bandbreite an musikalischen Genres beobachtet („omnivores“ vs. „univores“). Van Eijck (2001) spezifiziert diesen Befund dahingehend, dass höhere Statusgruppen ihre Musikpräferenz lediglich innerhalb einer größeren Vielfalt an musikalischen Genres dingfest machen.
Bislang zeichnet sich aus der Forschung zur Offenohrigkeit jedoch keine befriedigende Antwort ab, inwieweit diese vielfältigen Einflussfaktoren interagieren und in der Lage sind, als Prädiktoren für Offenohrigkeit zu fungieren.

2 Generelles Studiendesign

Die vorliegende Studie befasst sich mit der generellen Frage nach der Beschaffenheit des Konstruktes Offenohrigkeit. Sie stellt eines von vier empirischen Teilprojekten des Verbundprojektes „SIGrun – Studien zum Instrumentalunterricht an Grundschulen“ der Universitäten Bremen und Hamburg dar (siehe „www.sigrun2009.de“), das vom Bundesministerium für Forschung und Bildung (BMBF) zur Evaluation der Programme „JeKi – Jedem Kind ein Instrument“ in den Bundesländern Nordrhein-Westfalen und Hamburg gefördert wurde (2009 bis 2012). Die übergeordnete Fragestellung des SIGrun-Projektes lautet, inwieweit schulischer Instrumentalunterricht Bereiche wie Selbstkonzept und Klassenklima (Nonte & Schwippert, 2012), Kulturelle Teilhabe (Lehmann-Wermser & Jessel-Campos, 2013) und Musikpräferenz (Schurig et al., 2012) beeinflusst und welche Bedeutung

kooperativen Prozessen der Akteure (wie Schulleitung, JeKi-Lehrkraft) hierbei zukommt (Kulin & Schwippert, 2012).

2.1 Explorative Fragestellungen

Lässt sich die vielfach beschriebene Abnahme an Offenohrigkeit bestätigen, differenziert beschreiben und vorhersagen? Hierfür werden folgende explorative Fragestellungen untersucht, die sich aus dem bisherigen Stand der Forschung ableiten lassen:

1. Ist Offenohrigkeit als singulärer Faktor über Präferenzurteile für unkonventionelle Musik beschreibbar?
2. Sind ältere Kinder weniger offenohrig als jüngere Kinder?
3. Sind Jungen zu allen Messzeitpunkten weniger offenohrig als Mädchen?
4. Ist schulischer / privater Instrumentalunterricht ein Prädiktor für ausgeprägtere Offenohrigkeit?
5. Ist eine hohe Ausprägung der Persönlichkeitsdimension „Offenheit für Erfahrungen" ein Prädiktor für ausgeprägtere Offenohrigkeit?
6. Ist der Migrationshintergrund ein Prädiktor für geringe Offenohrigkeit?
7. Ist geringer sozio-ökonomischer Status ein Prädiktor für geringe Offenohrigkeit?

2.2 Methodisches Vorgehen

Den explorativen Fragestellungen wird sich mittels quantitativer Datenerhebung und Analyseverfahren genähert (standardisierte Fragebögen, Klingender Fragebogen). In das Forschungsdesign sind zusätzlich qualitative Methoden (qualitatitve Interviews) integriert. Im Sinne eines „complementary model of triangulation" (Erzberger & Kelle, 2003, S. 469) sollen die Ergebnisse der qualitativen und quantitativen Methoden einander ergänzend in Beziehung gesetzt werden (siehe Abschnitt 5).

3 Quantitative Erhebung: Offenohrigkeit als latentes Konstrukt

Es stellt sich die Frage, inwiefern eine Interpretation von Offenohrigkeit im Sinne eines latenten Konstrukts mit distinkten Prädiktorvariablen für weitere Analysen sinnfällig ist. Um dieser Frage nachzugehen, unternehmen wir den Versuch, ein latentes Faktormodell zur Musikpräferenz zu spezifizieren, dies auf alle vier Messzeitpunkte (MZP) zu übertragen und damit die Vorhersagekraft verschiedener Prädiktoren musikalischer Präferenz zu explorieren.

3.1 Stichprobendesign

Es wurde eine Kohorten-Studie mit jeweils einem MZP zum Ende jedes der vier Grundschuljahre (t_{1-4}: 2009 bis 2012) durchgeführt. Zusätzlich zu den Schulkindern wurden auch deren Eltern sowie deren Lehrkräfte (hier nicht berücksichtigt) befragt (siehe Tabelle 1). Das Sample setzt sich aus Gruppen zusammen, die auf Schulklassen aus 20 Grundschulen in Nordrhein-Westfalen und 13 Grundschulen in Hamburg basieren. Zur Konstruktion des basalen Modells werden sämtliche Untergruppen des Samples einbezogen (komplette Datensätze MZP 1 bis 4: n=735 Kinder). Die Erhebung wurde aus statistischen Erwägungen im Verlauf der Studie auf eine größere Stichprobe ausgeweitet, woraus sich neue Herausforderungen für die längsschnittlichen Analysen des Datenmaterials ergaben, wie im Abschnitt 3.4 zur statistischen Analyse näher erläutert wird.

Tabelle 1

Stichprobengröße

Erhebungs-methode	Befragte	2008/09 1. Schuljahr	2009/10 2. Schuljahr	2010/11 3. Schuljahr	2011/12 4. Schuljahr
Fragebögen	Schülerinnen und Schüler	1.143	1.244	1.180	1.066
	Eltern	914	761	735	508
Interviews	Schülerinnen und Schüler		31		28
	Eltern		25		10

Anmerkung: Interviews wurden nur im zweiten und vierten Schuljahr durchgeführt.

3.2 Messinstrument der abhängigen Variable „Musikpräferenz“

Als Messinstrument zur Erhebung der abhängigen Variablen zur Musikpräferenz diente ein Klingender Fragebogen, der in Anlehnung an entsprechende Messinstrumente anderer Studien konstruiert wurde (u.a. Gembris & Schellberg, 2007).

3.2.1 Musikbeispiele

Der Klingende Fragebogen setzt sich aus 16 (+1) musikalischen Exzerpten mit jeweils einer Dauer von etwa dreißig Sekunden und einem mittleren Tempobereich (zwischen 60 und 95 bpm, siehe Tabelle 2) zusammen. Drei dieser Stimuli wurden zur Erhöhung der Vergleichbarkeit mit Ergebnissen früherer Studien von diesen übernommen (Gembris & Schellberg, 2003, 2007; Kopiez & Lehmann, 2008; ebenfalls verwendet von Louven, 2011). Diese Stücke sollten die musikalischen Stilkategorien „Klassik“ (Bach und Mendelssohn) und „Avantgarde“ (Henze) repräsentieren. Das Bach-Exzerpt wurde erst ab dem zweiten MZP hinzugenommen, da das Mendelssohn-Beispiel einer Vielzahl an Kindern bereits zum ersten MZP aus dem Film „Barbie in Die 12 tanzenden Prinzessinnen“ (2006) bekannt war.

Zusätzlich wurde ein Musikbeispiel als „Cross-Over“-Stück (Garrett) zwischen den Stilkategorien „Klassik“ und „Pop“ einbezogen. Vier weitere Stimuli wurden in Hinblick auf eine möglichst große stilistische Vielfalt aus vier unterschiedlichen Musikkulturen (Türkei, Russland, China und Afrika) ausgewählt. Für diese vier Herkunftsregionen besteht ein relativ hohes Migrationsaufkommenin unseren Erhebungsgebieten (Statistisches Bundesamt, 2009, S. 66-68).

Tabelle 2

Musikstücke des klingenden Fragebogens

Musik-beispiel	**Komponist / Interpret / Album**	**Komposition / Song**	**Dauer / Ausschnitt**	**Tempo**
Übungs-beispiel	Friedbert Kerschbaumer / Die schönsten Kinderlieder auf der Panflöte	Ein Männlein steht im Walde	30 Sek. / 00:00-00:30	90 bpm
Afrika	Magi Shamba / Colors of Africa	Upepu	30 Sek. / 00:00-00:30	95 bpm
Türkei	Sümer Ezgü / Ege Toros Yörük Türkmen Türküleri (Anatolia Ethnic Music. Turkish Folk Music)	Ümmü	30 Sek. / 00:00-00:30	88 bpm
Russland	Samovar Russian Folk Music Ensemble / Some of our Best	Smyglyanka	28 Sek. / 01:24-01:52	86 bpm
China	Chinese Ensemble of Movie Music and Folk Music / Zhong Guo Dao Jiao Yin Le (Chinese Taoist Music)	Yu Fu Rong	30 Sek. / 01:20-01:50	90 bpm
Garrett	David Garrett / Encore	Air	30 Sek. / 01:52-02:23	60 bpm
Mendels-sohn	Felix Mendelssohn-Bartholdy	4. Sinfonie, 1. Satz	30 Sek. / 00:00-00:30	60 bpm
Henze	Hans Werner Henze	3. Sinfonie, 3. Satz „Beschwörungstanz“	32 Sek. / 00:41-01:13	ca 60 bpm
Bach	Johann Sebastian Bach	3. Orchester-Suite, "Gavotte I“	30 Sek. / 00:00-00:30	80 bpm
Acht Auftragskompositionen von Achim Gieseler (siehe Tabelle 3)			30 Sek.	90 bpm

Des Weiteren umfasst der Klingende Fragebogen acht Musikstücke, die eigens für die Studie von dem Berliner Komponisten Achim Gieseler komponiert wurden (Musikstücke können bei den Autoren erfragt werden). Diese Kompositionen wurden unter der Maßgabe erstellt, die Parameter klassische versus populäre Kompositionsweise, klassische versus populäre Instrumentationsweise sowie zusätzlich An- bzw. Abwesenheit eines Schlagzeuges (vgl. Jaedtke, 2000) zu kontrollieren. Jede mögliche Kombination dieser Parameter wurde dabei berücksichtigt (siehe Tabelle 3).

Tabelle 3

Auftragskompositionen

Musikbeispiel	**Kompositionsstil**	**Instrumentation**	**Drum Set**
Kla-Kla	Klassik	Klassik	Nein
Kla-Kla D			Ja
Kla-Pop		Pop	Nein
Kla-Pop D			Ja
Pop-Kla	Pop	Klassik	Nein
Pop-Kla D			Ja
Pop-Pop		Pop	Nein
Pop-Pop D			Ja

Anmerkung: Die hier verwendeten Kürzel werden in der Folge als Label verwendet.

3.2.2 Operationalisierung von Offenohrigkeit

Theoretisch wurde erwartet, dass die Auftragskompositionen mit einheitlich populären Stilmerkmalen dem musikalischen Genre „Pop“ zugeordnet werden und somit „konventionelle“ Musik repräsentieren. Als eher „unkonventionell“ wurdenhingegen die Auftragskompositionen mit einheitlich „klassischen“ Merkmalen sowie die Musikbeispiele der Stilkategorie „Klassik“ betrachtet. Mit dieser Einteilung wird deutlich, dass sich die Operationalisierung a priori an Hargreaves’ (1982) Annahmen orientiert, ohne jedoch die Problematik solch einer Zuordnung zu negieren (vgl. Kopiez & Lehmann, 2008; Louven,

2011). Für das Musikbeispiel der Stilkategorie „Avantgarde“ sowie die Beispiele aus den verschiedenen Musikkulturen wird angenommen, dass diese als eher „unkonventionell“ gelten. Eine positive Beurteilung der vermeintlich „unkonventionellen“ Musikbeispiele wird in der vorliegenden Studie als Offenohrigkeit interpretiert.
Bei dem „Cross-Over“-Beispiel sowie den stilistisch uneinheitlichen Auftragskompositionen fällt die Voraussage einer Zuordnung hingegen schwer. Die Kombination von Stilmerkmalen erscheint zwar eher „unkonventionell“, möglicherweise dominieren aber die Merkmale eines Stiles die kindliche Beurteilung. Die Präferenzäußerungen zu diesen a priori schwer zu kategorisierenden Musikbeispielen sollten im Rückschluss Anhaltspunkte dafür liefern, welche musikalischen Parameter die Gruppierung nach Stilkategorien und damit die Präferenzäußerungen beeinflussen. Hinsichtlich des Parameters Instrumentation erscheint insbesondere das Vorhandensein bzw. Nicht-Vorhandensein eines Schlagzeugs (Drum Set) relevant, da nach Jaedtke (2000, S. 206) der Schlagzeugrhythmus „strukturelle Fundamentbasis“ und „epochales Stilmerkmal“ von Popmusik sei. Allerdings könnte beim „Cross-Over“-Beispiel auch der dominante Klang des Streichinstrumentes Violine eine Kategorisierung als „Klassik“ begünstigen (Louven, 2011).

3.2.3 Durchführung der Erhebung des Klingenden Fragebogens

Während eines vollen Unterrichtstages wurden den Kindern im Klassenverbund nacheinander die Musikbeispiele mittels CD-Player mit standardisierter Lautstärke vorgespielt. Die Musikbeispiele wurden zur Vermeidung von Reihen-Effekten in verschiedenen Reihenfolgen dargeboten. Nach jedem Musikbeispiel wurde das Abspielen der CD angehalten, um den Kindern Zeit für ihr Urteil zu geben. Dieses wurde auf einer jeweils zu einem Musikstück dazugehörigen fünfstufigen ikonographischen Likert-Skala erhoben (Gesichter unterschiedlichen Ausdrucks von 1 „Will ich häufiger hören“ bis 5 „Will ich nicht hören“, siehe u.a. Gembris & Schellberg, 2007). Das arithmetische Mittel der Präferenz-Ratings reichte von AM=1,64 (Pop-Pop D zu t1: Standardabweichung [SD]=1,12) bis AM=3,35 (Türkei zu t3: SD=1,37), wobei die Ratings der Musikbeispiele zum Großteil eine positive Schiefe aufweisen (über alle MZP haben nur 5 Variablen eine Schiefe < 0; 4 davon im vierten MZP) und nicht-normalverteilt sind.

3.3 Messinstrumente der unabhängigen Variablen

Zur Erfassung der unabhängigen Variablen beantworteten die Grundschulkinder und / oder deren Eltern standardisierte Messinstrumente, die unter anderem Informationen zu Alter und Geschlecht, zur Teilnahme am JeKi-Programm sowie zum Erhalt von privatem Instrumentalunterricht, zur kindlichen Persönlichkeit, zum sozio-ökonomischen Status und Migrationshintergrund, zu kognitiven Fähigkeiten, zum elterlichen Unterstützungsverhalten sowie zu außerschulischen musikbezogenen Verhaltensweisen erfassen. Die Operationalisierung der bislang verwendeten unabhängigen Variablen Persönlichkeit, sozio-ökonomischer Status, Migrationshintergrund, schulischer (JeKi-Teilnahme) sowie privater Instrumentalunterricht wird im Folgenden erläutert (siehe Tabelle 4 für deskriptive Werte der beobachteten Variablen beispielhaft dargestellt anhand von MZP 3).

3.3.1 Kindespersönlichkeit

Eltern beantworteten für ihre Kinder den Fünf-Faktoren-Fragebogen für Kinder (FFFK) von Asendorpf (1998), mit dem fünf Dimensionen kindlicher Persönlichkeit (Extraversion, Emotionale Stabilität, Gewissenhaftigkeit, Verträglichkeit, Offenheit für Kultur) auf der Basis von bipolaren Adjektivpaaren erhoben werden. Der Faktor Offenheit für Kultur weicht in seiner Bedeutung vom gängigeren Faktor Offenheit der Big-Five ab (z.B. Costa & McCrae, 1992), wurde aber hieraus abgeleitet. Er wird von Asendorpf & van Aken (2003) als „culture/intellect/openness with antecedents and outcomes of school achievement" beschrieben. Cronbachs Alphas liegen in unserer Studie im Bereich der Normwerte und können als befriedigend angesehen werden (Werte von Extraversion zu t_4: α=,875 bis Verträglichkeit zu t_1: α=,744). Obgleich für einzelne Items Multidimensionalität beobachtet wurde (Wagschal et al., 2010), integrieren wir mit Bezug auf Asendorpf und von Aken (2003) dieses Messinstrument dennoch in unserer Analyse. Als mögliche Regressoren wurden die Faktoren „Kultur/Intellekt/Offenheit" und „Extraversion" berücksichtigt. Da die elterliche Beurteilung der kindlichen Persönlichkeit über die vier Messzeitpunkte jedoch stabil war, die

Eltern also keine Veränderungen in der Persönlichkeit ihrer Kinder angaben, können mögliche Persönlichkeitsentwicklungen nicht mit einbezogen werden.

3.3.2 Weitere potentielle Prädiktoren

Der sozio-ökonomische Status wird von uns über den höchsten von einem Elternteil erreichten Bildungsabschluss auf einer vierstufigen Skala gebildet (z.B. Ehmke & Siegle, 2005). Die JeKi-Teilnahme wurde mittels Fragebogen-Angaben der Kindern erhoben. Der Migrationshintergrund beschreibt, ob mindestens ein Elternteil sich selbst bzw. den anderen Elternteil als Zuwanderer beschreibt. Mit der Variable Instrumentalunterricht wurde erfragt, ob das Kind in dem vorangegangenen Jahr außerschulischen Instrumentalunterricht erhalten hat. Diese Variable wurde pro MZP neu errechnet, bei den anderen potentiellen Prädiktoren wurden die Angaben für die hier präsentierten Analysen aus MZP 1 als Basis verwendet und bei fehlenden Werten aus den folgenden Messzeitpunkten aufgefüllt.

Tabelle 4

Deskriptive Beschreibung der Prädiktoren für den dritten Messzeitpunkt

Variable	**Skala**	**N**	**Zutreffend**			
JeKi-Teilnahme	0=Nein, 1=Ja	817	69,0%			
Migrationshinter-grund der Eltern	0=Nein, 1=Ja	1.167	30,9 %			
Instrumental-unterricht Klasse 3	0=Nein, 1=Ja	709	64,6 %			
Variable	**Skala**	**N**	**Min.**	**Max.**	**Mittel-wert**	**Standard-abweich-ung**
Höchster Bildungsabschluss der Eltern	1=kein Abschluss, 2=Haupt- oder Realschulabschluss, 3=Allg. (Fach-) Hochschulreife, 4=(Fach-) Hochschul-abschluss	1.052	1	4	3,27	0,82
Persönlichkeits-dimension Kultur/Intellekt/ Offenheit	Mittelwert 1-5, 5=Hohe Zustimmung	520	1,75	5	3,98	0,61
Persönlichkeits-dimension Extraversion	Mittelwert 1-5, 5=Hohe Zustimmung	521	1,75	5	4,06	0,66

3.4 Statistische Analyse

Es kamen faktorenanalytische Verfahren (Faktoranalyse: FA) im Strukturgleichungsansatz zum Einsatz. Diese Verfahren der multivariaten Statistik erlauben das Schließen von empirischen Beobachtungen mehrerer verschiedener manifester Variablen auf darunter vermutete latente (theoretische/nicht-beobachtbare) Variablen oder latente Faktoren (s.a.

Bollen, 2002). Durch den Einsatz von Verfahren der Strukturgleichungsmodellierung (SEM) können geschachtelte Samples (Reinecke, 2005) sowie multiple latente Faktoren und die Theorie des Messfehlers berücksichtigt werden. Allgemeines Ziel dieses Vorgehens ist es, Variablen auf einer höheren Abstraktionsebene zusammenzufassen, um den Einfluss einzelner Indikatoren auf das unterstellte gemeinsame, zugrunde liegende Merkmal zu bestimmen und zeitgleich Hypothesentests vorzunehmen. In diesem Rahmen werden Varianz/Kovarianzanalysen sowie Korrelationen und lineare Regressionen auf manifester und latenter Ebene verwendet. Die Analysen der einzelnen MZP werden nebeneinander präsentiert, da die uneinheitliche Datenstruktur keine vollständige Verknüpfung über alle MZP zulässt.

Das ursprüngliche Studiendesign sah vor, den Klingenden Fragebogen nur in Nordrhein-Westfalen einzusetzen und die acht Auftragskompositionen zudem im split-half-Design darzubieten. Es zeigte sich jedoch nach MZP 1, dass die Stichproben nicht die erwarteten Umfänge aufwiesen. Daraufhin wurde die Erhebung ab MZP 2 auf Hamburg ausgeweitet. Erste Analysen ergaben zudem, dass die durch das split-half-Design entstandenden Teilstichproben für die angestrebten Verfahren nicht hinreichend homogen waren. Daher wurde ab MZP 3 der gesamte Fragebogen in Nordrhein-Westfalen und in Hamburg eingesetzt. Für die statistischen Analysen bedeutet dies, dass die Stichprobenumfänge zwischen den ersten beiden und den letzten beiden MZP voneinander abweichen. Des Weiteren bedeutet dies, dass der Faktor „Pop“ zum MZP 1 und 2 nicht modelliert werden kann.

Die für Schuluntersuchungen typische hierarchische Schachtelung der Daten wurde durch Korrektur des Standardfehlers über einen implementierten Sandwich-Schätzer berücksichtigt. Der auf die Klassenzugehörigkeit zurückzuführende Anteil der Varianz (Intra-Klassen-Korrelation, ICC) lag bei 6%-10%, womit dieses Vorgehen angemessen erscheint (vgl. vertiefend bei Leopold, 2012). Im Rahmen der SEM-Verfahren wurde der robuste weighted least squares Schätzer *(*WLS*;* in MPlus WLSMV*)* eingesetzt, um nicht-normale Verteilungen in den Indikatoren zu berücksichtigen. Die Analysen wurden mit MPlus 6.11 gerechnet. MPlus verwendet die Full Information Maximum Likelihood Funktion (FIML) im Umgang mit fehlenden Werten.

Im ersten Schritt der Analyse wird der konfirmatorischer Fit der faktoriellen Struktur im Schema wie in Abbildung 1 dargestellt getestet.

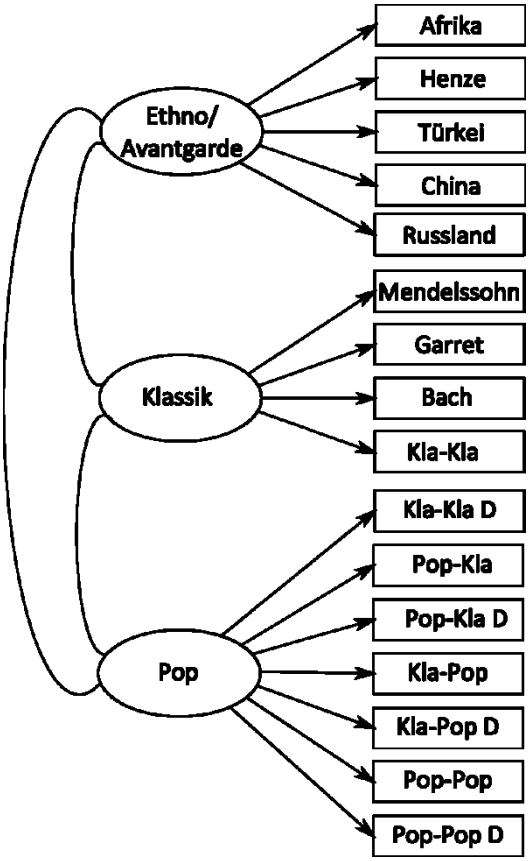


Abbildung 1.: Unterstellte faktorielle Struktur der Messzeitpunkte 3 und 4.

Entsprechend dem gängigen Vorgehen in der konfirmatorischen Faktoranalyse wird das Modell in der Folge angepasst und Invarianztestungen unterworfen, bei denen geprüft wird, inwieweit sich Modelle in Bezug auf ihre Parameter (z.B. Ladungsstruktur) unterscheiden. Dieses methodische Vorgehen der Modellanpassung ist in vielen Bereichen der empirischen Bildungs- und Sozialforschung üblich (vgl. Bollen, 2002, S.1-2). Es muss betont werden, dass die Modell-Anpassung keinesfalls beliebig verläuft und eine Reduktion des berücksichtigten Datenmaterials nur solange verfolgt wird, wie diese theoretisch sinnfällig gerechtfertigt werden kann.

Der Chi^2-Anpassungstest testet die prognostizierte Kovarianzmatrix des Modells gegen die Kovarianzmatrix in den Daten, wobei die Nullhypothese lautet: Das Modell passt exakt. Es ist

jedoch bekannt, dass der Test übersensibel bei großen Stichproben (n<400) und nicht-parametrischen Indikatoren reagiert (Hu & Bentler, 1998). Daher schlagen Bentler und Chou (1987) als Daumenregel eine Abhängigkeit des Stichprobenumfangs zur Modellkomplexität, also den Freiheitsgraden (df), vor (5 zu 1). Der inkrementelle Comparative Fit Index (CFI) sollte sich 1 annähern und gilt ab etwa ,95 als ausreichend. Der approximative Index Root Means Square Error of Approximation (RMSEA) sollte höchstens Werte bei ,08 aufweisen. Sowohl CFI als auch RMSEA basieren auf dem Chi^2-Wert.

Im Rahmen dieser Faktor-Modellierung wurde zudem innerhalb jedes MZP ein Differenz-Test auf Invarianz über Gruppen (Chi^2-Modellvergleichstest) vorgenommen, um die Vergleichbarkeit der Sub-Samples zu prüfen (vgl. Steenkamp & Baumgartner, 1998).
Im zweiten Schritt der Analyse werden die längsschnittlichen Daten aufgrund der oben beschriebenen Uneinheitlichkeit der Datenstruktur über Varianzanalysen mit Messwiederholung beschrieben. Dafür werden die Faktoren als Mittelwerte der Indikatoren pro MZP operationalisiert, um eine einfache Form der Mittelwertsvergleiche vornehmen zu können.
Im dritten Schritt der Analyse werden latente Regressionen gerechnet. Als Basis hierfür dienen die zuvor erstellten Messmodelle, die nun mit möglichen Prädiktorvariablen ergänzt werden. Hierbei wurde auf mögliche Suppressionseffekte über schrittweise Hinzugabe der Prädiktoren kontrolliert.

3.5 Ergebnisse

Bei Betrachtung der musikalischen Beschaffenheit der Musikstücke sind diverse mehrfaktorielle Lösungen denkbar. Vorgestellt wird zunächst die sparsamste einheitliche Lösung, die den rechnerisch größten Modellfit bei größter Nähe zum theoretischen Modell pro MZP aufweist. In den darauffolgenden Analyseschritten wird dann auf dieser Faktor-Modellierung aufgebaut.

3.5.1 Faktor-Modellierung

Im ersten Schritt wurden die vollen CFAs modelliert. Erwartungsgemäß erreichte keine Lösung sofort einen angemessenen Fit, sodass die im Folgenden beschriebene Anpassung des Modells notwendig war. Von den insgesamt 16 Musikbeispielen konnten drei Beispiele (Russland, Kla-Kla D und Pop-Kla) keinem Faktor eindeutig zugeordnet werden und wurden daher für Folgeanalysen ausgeschlossen (Doppelladungen mit Effektstärken jeweils bei $\beta \approx ,4$). Bei dem Vergleich der Teilstichproben stellte sich heraus, dass zwischen Jungen und Mädchen in MZP 1 bis 3 nur partielle schwache Invarianz vorliegt (z.B. ergaben sich zum MZP 3: $Chi^2=5,979$, $df=3$, $p=,113$ für metrische Invarianz; $Chi^2=108,132$, $df=7$, $p<,001$ für skalare Invarianz und $Chi^2=4,388$, $df=4$, $p=,356$ für partiell skalare Invarianz). Bereits in der Ansicht der Mittelwertstruktur (siehe Tabelle 5) eröffnet sich, was sich in den komplexeren Verfahren zeigte: Das Geschlecht hat vor allem bei den MZP 1 bis 3 einen signifikanten Einfluss auf spezifische Variablen. Um dies angemessen modellieren zu können, wurde eine Instrumentalvariable „Geschlecht“ eingefügt (siehe Abbildung 2 dargestellt für MZP 3), die einen strukturellen Vergleich zwischen den Modellen ermöglicht (z.B. Weiber & Mühlhaus, 2010). Die Fit-Indizes der einzelnen Modelle lassen eine Annahme der Modelle für die MZP 1 bis 3 zu (siehe Tabelle 6).

Tabelle 5

Deskriptiver Vergleich über das Geschlecht

MZP	Geschlecht	Werte	Henze	Mendelssohn	Garret	Bach	Kla-Kla
1	Mädchen	Mean	1,47	1,78	2,05	n.z.	2,06
		SD	1,93	1,30	1,36	n.z.	1,27
	Jungen	Mean	1,46	2,78	2,69	n.z.	2,75
		SD	1,63	1,63	1,50	n.z.	1,59
2	Mädchen	Mean	2,29	2,03	2,40	2,28	2,17
		SD	1,50	1,39	1,45	1,39	1,37
	Jungen	Mean	1,92	2,79	3,03	2,87	2,78
		SD	1,36	1,74	1,62	1,59	1,51
3	Mädchen	Mean	2,64	2,30	2,80	2,71	2,50
		SD	1,45	1,37	1,45	1,44	1,41
	Jungen	Mean	2,11	3,02	3,52	3,15	3,15
		SD	1,34	1,51	1,45	1,58	1,46
4	Mädchen	Mean	3,04	2,20	2,51	2,32	2,31
		SD	1,32	1,20	1,25	1,25	1,19
	Jungen	Mean	2,45	2,17	2,78	2,20	2,37
		SD	1,43	1,21	1,23	1,26	1,48

Anmerkung(en). Die jeweilige Stichprobengröße variert aufgrund des Samplings. Generell liegt sie zwischen n=273 (MZP 1 Jungen) und n=459 (MZP 3 Mädchen). Bei dem Musikbeispiel Kla-Kla liegt die Untergrenze zum MZP 1 bei n=152.

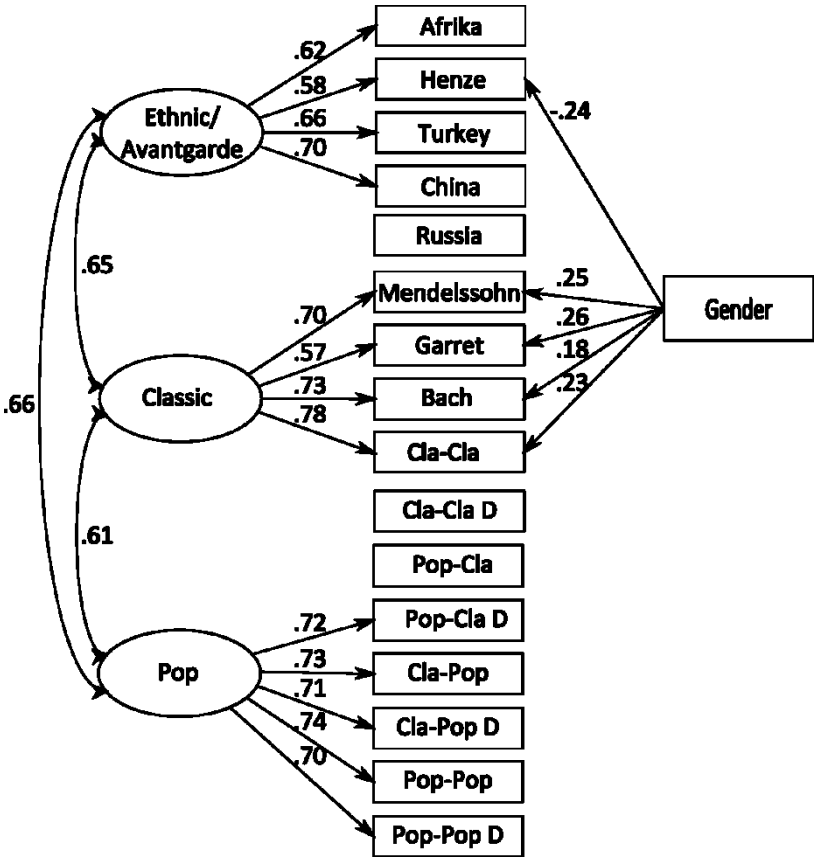


Abbildung 2. Getrimmtes Modell.

Tabelle 6

Modell-Fit-Indizes

MZP	n	Chi^2	df	p	CFI	RMSEA
1	444	21,794	11	,0261	,982	,047
2	890	42,381	16	<,001	,986	,043
3	1172	92,694	23	<,001	,978	,051
4	995	395,345	23	<,001	,875	,110

Anmerkungen: Chi^2: Pearson Chi^2 auf Anpassungsgüte; df: Freiheitsgrade; p: Signifikanzniveau der Chi^2; CFI: Comparative Fit Index; RMSEA: Root Means Square Error of Approximation

Für den MZP 4 lässt die Faktorstruktur jedoch eine Annahme des Modells nicht zu. Dies bedeutet, dass sich die Präferenzurteile der Kinder zum MZP 4 nicht auf dieselbe Weise faktoriell gruppieren lassen wie zu den MZP 1 bis 3. Das bisher passende Modell kann also

zum vierten Messzeitpunkt nicht erneut zur strukturellen Analyse der Präferenzäußerungen herangezogen werden. Daher wird die Struktur deskriptiv auf Basis einer explorativen Faktoranalyse vorgestellt. In Tabelle 7 ist die rotierte Faktorladungsmatrix dargestellt. Die Anzahl der Faktoren wurde über Modellvergleiche und das Kaiser-Kriterium (Faktoren mit Eigenvalue >1) abgeleitet, d.h. das sparsamste, nach Fit-Kriterien gerade passende Modell wurde gewählt. Die Korrelationen zwischen den rotierten Faktoren 1 und 2 liegen bei r=,469, zwischen den rotierten Faktoren 1 und 3 bei r=,406 und zwischen den Faktoren 2 und 3 bei r=,213.

Tabelle 7

Faktorladungsmatrix MZP 4

Musikbeispiel	**Faktor MZP 1 – 3**	**Faktor 1**	**Faktor 2**	**Faktor 3**
Afrika	Ethno/Avantgarde	,676		
Türkei	Ethno/Avantgarde			,475
Henze	Ethno/Avantgarde		,507	
China	Ethno/Avantgarde	,415		,538
Russland	n.z.	,544		
Mendelssohn	Klassik	,737		
Garret	Klassik	,606		
Bach	Klassik (MZP 2-3)	,744		
Kla-Kla	Klassik (MZP 3)	,758		
Kla-Kla D	n.z.		,512	,407
Kla-Pop	Pop (MZP 3)		,584	
Kla-Pop D	Pop (MZP 3)		,672	
Pop-Kla	n.z.		,582	
Pop-Kla D	Pop (MZP 3)	,724		
Pop-Pop	Pop (MZP 3)			,640
Pop-Pop D	Pop (MZP 3)			,709

Anmerkungen: Oblique-Geomin-Rotierte Faktorladungsmatrix, Ladungen <,4 wurden zur besseren Übersichtlichkeit unterdrückt

Zu erwähnen bleiben die Effektstärken in den passend erscheinenden Modellen. Die geringsten Faktorladungen von Indikatoren (standardisiertes Effektmaß: β_{stand}) betreffen die Variablen „Afrika“ und „Henze“ zum MZP 1 sowie „Henze“ zum MZP 3 (jeweils <,6). Der Durchschnitt der Ladungsstärken liegt bei ~,690. Die latenten Korrelationen der Faktoren „Klassik“ und „Ethno/Avantgarde“ liegen zum MZP 1 bei ,758 und zum MZP 2 bei ,794. Zum MZP 3 liegt die latente Korrelation zwischen „Klassik“ und „Ethno/Avantgarde“ bei ,653, zwischen „Klassik“ und „Pop“ bei ,614 und zwischen „Pop“ und „Ethno/Avantgarde“ bei ,663. Restriktivere einfaktorielle Modelle zeigen zu allen MZP einen signifikant schlechteren Fit als die mehrfaktoriellen Modelle. Hinsichtlich der ersten explorativen Fragestellung muss somit angenommen werden, dass Offenohrigkeit nicht als singulärer Faktor über Präferenzurteile für unkonventionelle Musik beschreibbar ist.

3.5.2 Deskriptiver Längsschnitt

Die Ergebnisse der Varianzanalysen mit Messwiederholung zur Beschreibung des Längsschnitts sind in Abbildung 3 dargestellt. Abgetragen sind hierbei die arithmetischen Mittelwerte der Faktoren (operationalisiert als Mittelwerte der Indikatoren pro MZP) inklusive der Standardfehler und nach geschlechtsspezifischen Gruppen aufgeteilt. In dieser Abbildung ist auch MZP 4 abgebildet. Es muss jedoch wie oben erläutert beachtet werden, dass sich die der Analyse zugrundeliegende faktorielle Gliederung der Präferenzurteile in Bezug auf MZP 4 nicht mehr aus dem empirischen Material ergibt und somit eine Interpretation der Werte für MZP 4 nicht zulässig ist.

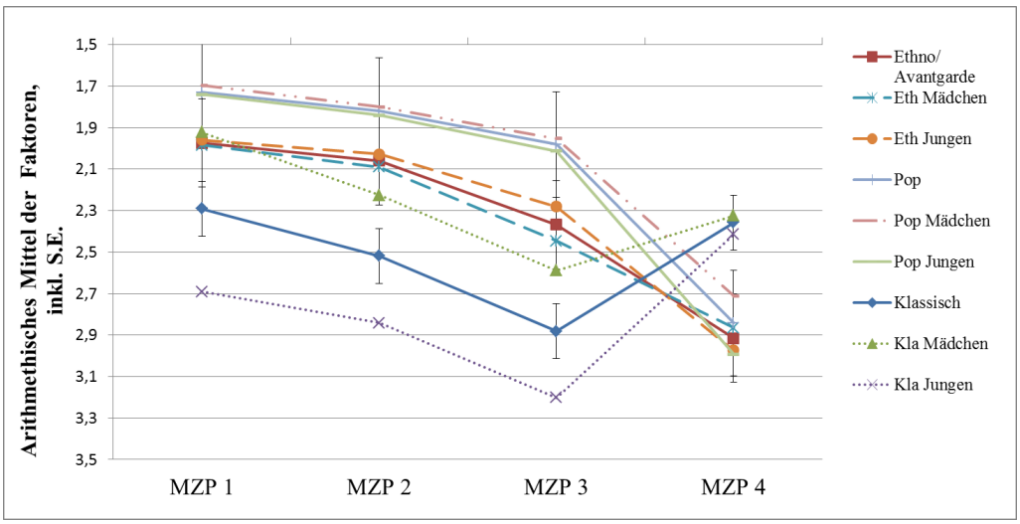


Abbildung 3. Arithmetisches Mittel der Faktoren.

Anmerkung: Die y-Achse ist zum intuitiveren Verständnis invertiert.

Für die MZP 1 bis 3 ist zu erkennen, dass die Präferenzratings für die Faktoren „Pop“ und „Ethno/Avantgarde“ generell verhältnismäßig positiv ausfallen. Ausschließlich die mittleren Präferenzratings der Jungen bezüglich des Faktors „Klassik“ zum dritten MZP befinden sich unterhalb des natürlichen Skalenmittelwertes von 3. Während sich die geschlechtsspezifischen Ratings der Faktoren „Ethno/Avantgarde“ und „Pop“ nicht signifikant unterscheiden ($p<,001$), differieren die Ratings des Faktors „Klassik“ stark. Obwohl sich die „Klassik“-Ratings der Jungen und Mädchen signifikant ($p<,001$) unterscheiden, verläuft die Entwicklung dieser Ratings über die ersten drei MZP parallel. Wie erwartet fallen die Urteile im Verlauf für alle belastbaren Faktoren deutlich linear ab. Bezüglich der zweiten explorativen Fragestellung lässt sich somit vermuten, dass ältere Kinder weniger offenohrig sind als jüngere Kinder. Auf der Basis dieser Ergebnislage kann jedoch nicht angenommen werden, dass Jungen generell weniger offen für unkonventionelle Musik sind als Mädchen, wie in der dritten explorativen Fragestellung formuliert wurde. Vielmehr erscheint eine Differenzierung nach musikalischen Stilkategorien vonnöten, um den Geschlechtseffekt zu verdeutlichen.

3.5.3 Latente Regressionen

Bei der Analyse der latenten Regressionen werden nur jene Ergebnisse vorgestellt, bei denen keine Missings auftraten, sodass die Stichprobengröße der Einzelmodelle relativ gering ist (MZP 1: $\underline{n}$=256, MZP 2: $\underline{n}$=178, MZP 3: $\underline{n}$=161). Alle Modelle weisen erwartungsgemäß einen guten Fit auf (jeder $\underline{Chi}^2$ nicht signifikant auf 5%-Niveau, RMSEA immer <,035). In diese Analysen wurden zu den MZP 1 bis 3 folgende Prädiktoren einbezogen: Persönlichkeitsdimensionen „Kultur/Intellekt/Offenheit“ und „Extraversion“, privater Instrumentalunterricht, JeKi-Teilnahme, Migrationshintergrund und höchster Bildungsabschluss im Elternhaus. Da das Geschlecht bereits als Instrumentalvariable verrechnet wurde, konnte es hier nicht mit angegeben werden (standardisierte Regressionskoeffizienten $\underline{\beta}$ siehe Tabelle 8). Zum Vergleich beläuft sich die Effektstärke der Variable Geschlecht – sofern sie nicht als Dummy, sondern als Prädiktor für den Faktor „Klassik“ verrechnet wird – auf $\underline{\beta_{stand}}$=-,840. Insgesamt konnten somit nur schwache Effekte beobachtet werden, sodass die vierte bis siebte explorative Fragestellung verneint werden sollten.

Tabelle 8

Messzeitpunkt	**1**		**2**		**3**		
Variablen	Klassik	Ethno/ Avantgarde	Klassik	Ethno/ Avantgarde	Klassik	Ethno/ Avantgarde	Pop
Migrationshinter-grund der Eltern					-0,236		-0,251
JeKi-Teilnahme							
Höchster Bildungsabchluss der Eltern		0,221					
Persönlichkeits-dimension Kultur/Intellekt/ Offenheit							
Persönlichkeits-dimension Extraversion	-0,164			-0,185	-0,137	-0,213	
Instrumental-unterricht							-0,171

Anmerkungen: Berichtet wurden Effekte die mindestens auf dem Niveau <,05 signifikant sind. In der Codierung der Musikbeispiele bedeutet eine „1“ ein positives Urteil und eine „5“ ein negatives Urteil.

3.6 Fazit zur quantitativen Erhebung

Die Erfassung kindlicher Musikpräferenz über Paper & Pencil-Untersuchungen in Klassengruppen ist problematisch. So verweisen die Daten auf einen ausgeprägten Geschlechtereffekt. Darüber hinaus konnten Deckeneffekte beobachtet werden und erst mit wachsendem Alter wird die volle Skalenspanne verwendet. Die hohen Korrelationen zwischen den latenten Faktoren implizieren die Möglichkeit eines Faktors zweiter Ordnung. Diese Möglichkeit deutet auf ein generelleres Konstrukt über alle Faktoren hinweg, welches

aber auf der Basis unserer Daten noch nicht sinnhaft abgeleitet werden konnte, da die Modelle rechnerisch nahezu äquivalent wären und keine klare Modellwahl getroffen werden könnte. Die starke Abnahme der Präferenzäußerung für die Auftragskompositionen lassen sich vermutlich auch mit der wachsenden ästhetischen Urteilskraft der Kinder erklären: So konnte zu den letzten MZP informell beobachtet werden, dass die Kinder während der Erhebung mit dem Klingenden Fragebogen die einheitlichen Kompositionsstile wiedererkannten und diese in der Wiederholung als langweilig empfanden. Auch wurde insbesondere zum MZP 4 wiederholt geäußert, dass die Pop-Musik nicht so gut sei wie diejenige, die sie privat hören würden.

Im Bewusstsein dieser methodischen Einschränkungen bestätigen die präsentierten Ergebnisse zwar die bereits in früheren Studien beobachtete generelle Abnahme der Präferenzäußerungen für unkonventionelle Musik in verschiedenen identifizierbaren Facetten. Von dieser Entwicklung sind jedoch auch die konventionellen Pop-Beispiele betroffen. Die Daten verweisen zudem auf einen ausgeprägten Geschlechtereffekt, der jedoch genre-spezifisch zu differenzieren ist und sich wesentlich durch die oben beschriebene sozial bedingte Antwortverzerrung bei den Jungen erklären lässt. Des Weiteren ergab sich im Alter von 9 bis 10 Jahren ein Bruch in den Präferenzäußerungen, der sich in der Auflösung der zuvor stabilen Faktorstruktur manifestiert. Dies verweist vermutlich auf eine beginnende Individualisierung der Präferenzäußerungen, die sich nicht mehr deutlich über die Gesamtgruppe faktoriell bündeln lässt. Um auf dieser Datenbasis Aggregate zu bilden und eindeutigere Ableitungen zu erzielen, wäre die Verwendung sparsamer Modelle mit klar definierten Subgruppen vonnöten, um beispielsweise individuell herauszustellen, was als unkonventionelle Musik angesehen wird. Zusammenfassend kann festgehalten werden, dass die verwendeten Variablen neben Alter wenig prädiktive Kraft für die verwendeten Konstrukte haben, sofern man von dem starken geschlechtsspezifischen Effekt absieht. Auch hier bietet sich die Weiterverarbeitung in kleinen Sub-Samples an, um die Individualität in dem facettenreichen Konstrukt Offenohrigkeit greifbarer zu machen.

4 Qualitative Erhebung: Kindliche Musikpräferenz als Ausdruck von Geschlechstidentität

4.1 Methodisches Vorgehen

Die qualitative Erhebung wird im Folgenden stark zusammenfassend und auf die für die quantitative Erhebung relevanten Ergebnisse fokussierend dargestellt (eine detaillierte Beschreibung und Analyse der Interview-Erhebungen wird im Rahmen des Promotionsvorhabens von Nicola Bunte präsentiert werden). Ergänzend zu der quantitativen Erhebung wurden insgesamt 28 der beforschten Kinder in neun leitfadengestützten Interviews in Kleingruppen (2-4 Kinder) zu ihren Musikpräferenzen an je zwei Interviewzeitpunkten (1. IZP: Mitte 2. Schuljahr; 2. IZP: Mitte 4. Schuljahr) befragt. Der Interviewleitfaden umfasste Fragen zur Lieblingsmusik, zu vorgespielten Musikbeispielen (Pop-Pop D und Kla-Kla) sowie zu musikalischen Geschlechtsstereotypen. Letztgenannter Aspekt wurde in den Leitfaden aufgenommen, da Erklärungsansätze für den bereits zum 1. MZP beobachteten Geschlechtereffekt erwartetet wurden. Zu beiden IZP wurden daher auch geschlechtshomogene Gruppen interviewt. Die Auswertung erfolgte zunächst inhaltsanalytisch in einer Kombination aus induktivem und deduktivem Vorgehen (vgl. Mayring, 2007). Im Rahmen der Analysen des 1. IZP, auf denen hier in der vergleichenden Perspektive der IZP aufgebaut wird, hat Beutler-Prahm (2012) ein Kategoriensystem entwickelt, das anhand der Daten des 2. IZP weiterentwickelt sowie induktiv ergänzt wurde.

4.2 Zusammenfassung ausgewählter Ergebnisse

Ein Teilbereich des entwickelten Kategoriensystems erfasst Aussagen zu „musikalischen Konzepten“ (in Anlehnung an Behne,1975), diese wurden jedoch weniger auf individueller als auf interindividueller Ebene analysiert. Die umfangreichsten Aussagen betrafen dabei zum einen die geschlechtsstereotypen Konzepte „Mädchenmusik“ und „Jungenmusik“, die auch losgelöst von expliziten Interviewfragen offenbar wurden, sowie die induktiv ermittelten Konzepte „Rockmusik“ und „Chartsmusik“.

Zum 1. IZP sind die Kinder überwiegend der Meinung, dass es so etwas wie „Jungenmusik“ und „Mädchenmusik“ gibt. Während die Jungen zumeist klare Vorstellungen zu diesen beiden Konzepten äußern, fallen die Charakterisierungen der Mädchen zurückhaltender aus. Die Aussagen der Kinder ergänzen sich jedoch zu konsistenten Konzepten. „Mädchenmusik“ wird überwiegend als „schön und leise“ oder als „ruhigere“ Musik charakterisiert und umfasst typischerweise die Instrumente Geige, Cello und Flöte sowie weiblichen Gesang. „Jungenmusik“ wird häufig mit „rockig“, „Rockgitarre“ und Schlagzeug sowie als laut und „cool“ charakterisiert. Diese Charakterisierung ist in großen Teilen identisch mit dem unabhängig davon analysierten Konzept „Rockmusik“, das eine zentrale Rolle bei der Bewertung von Musik einnimmt.

Hingegen wird über die persönliche Musikpräferenz von den Kindern nur selten in direkter Verbindung zu den beiden geschlechtsspezifischen Konzepten gesprochen. Allerdings zeigen sich geschlechtsspezifische Unterschiede, die insbesondere bei den Jungen in Richtung der Konzepte weisen. Während der überwiegende Teil der Jungen zum 1. IZP eine deutliche Vorliebe für laute und rhythmusbetonte Musik und Instrumente (Schlagzeug, Trommel und E-Gitarre) äußert, mögen Mädchen auch „mittellaute“ oder „leise Musik“. Die meisten Mädchen hören zudem gern Instrumente wie Cello, Flöte und Gitarre.

Bei der Bewertung von Musik (auch der vorgespielten Musikbeispiele) nehmen Jungen häufiger Bezug zu dem Konzept „Rockmusik“ als Mädchen. Die Gesprächsverläufe zeigen, dass Jungen überwiegend von sich aus eine starke Präferenz für Rockmusik äußern, während in den wenigen Mädchen-Aussagen zu „Rockmusik“ hauptsächlich die Bewertungen der Jungen übernommen werden. In der reinen Mädchen-Interviewgruppe geben dagegen alle an, nicht besonders gern Rockmusik zu hören.

Ihre positive Bewertung des im Interview vorgespielten Pop-Musikbeispiels begründen die Jungen am häufigsten durch den Bezug auf Rhythmusinstrumente, Rhythmus oder Rockmusik: „Ich fand die besser, weil die mehr rockig war eben“ (Schüler 2, Interview 1), „Ich finde der Rhythmus ist cool“ (Schüler 2, Interview 8). Solche Begründungen sind bei den Mädchen selten und fast ausschließlich im Anschluss an eine ähnliche Jungen-Aussage zu finden. In Bezug auf das Klassik-Musikbeispiel nehmen einige Jungen eine Bewertung in Abgrenzung zur „Rockmusik“ vor: „Mir gefällt ja mehr diese Rockmusik“ (Schüler 3, Interview 1).

Im Gegensatz zum 1. IZP weisen zum 2. IZP alle Mädchen spontan die Existenz von „Jungenmusik“ und „Mädchenmusik“ zurück oder machen keine Angaben zu dieser Frage. Lediglich in einer homogenen Mädchen-Interviewgruppe werden die Konzepte noch beschrieben – allerdings unter Einnahme einer Jungen-Perspektive: „Meine Brüder meinen immer, Mädchenmusik ist, wenn eine Frau das singt“ (Schülerin 2, Interview 9).
Die Jungen äußern zwar nur vereinzelt spontane Zustimmungen, dennoch finden sich weiterhin einige Beschreibungen der Konzepte, eine Mädchen-Perspektive wird dabei nicht eingenommen. „Jungenmusik“ wird vereinzelt beschrieben als „DJ-Mixes“, „laute und coole“ Musik mit schnellem und unverständlichem Text sowie mit „anderem Gesang“ als bei „Mädchenmusik“. Diese wird selten, dann aber konform zu den Angaben der Mädchen und den Aussagen zum 1. IZP charakterisiert. Im Vergleich zum 1. IZP fällt auf, dass weder Jungen noch Mädchen in ihrer Beschreibung von „Jungenmusik“ und „Mädchenmusik“ typische Instrumente nennen.
Bei den Angaben zu den eigenen Präferenzen sowie zu den Präferenzen anderer orientieren sich nur die Jungen an den geschlechtsspezifischen Konzepten: „Und es gibt typische Jungsmusik. Zum Beispiel Mixes. Von DJs und so, das Mädchen jetzt (…) Ich kenne außer dir kein einziges Mädchen, das die hört” (Schüler 2, Interview 2). Auf normativer Ebene sind sich die Kinder zwar einig, dass jeder hören könne, was er mag. Doch zeigen sich in einer homogenen Jungen-Interviewgruppe sprachlich äußerst hart formulierte Ablehnungen von „Mädchenmusik“: Diese sei “bestimmt für uns sicherlich scheiße“ (Schüler 2, Interview 8).
Eine enge Assoziation zwischen den Konzepten „Jungenmusik“ und „Rockmusik“ bildet sich zum 2. IZP nicht mehr ab. „Rockmusik“ wird teilweise zwar weiterhin zur Charakterisierung persönlicher Präferenzen verwendet. Als neues und zentrales Konzept für die Beurteilung von Musik zeigt sich jedoch das der „Chartsmusik“, mit dem die nun bedeutsame Aktualität und Bekanntheit von Musik vor allem von Jungen beschrieben wird. Über „Chartsmusik“ wird als die „bekanntesten“ und „die aktuellen Lieder“ gesprochen und ein Bezug zu den „Top 20“ oder „Top 100“ hergestellt. Aktualität und verbale Musikpräferenzen werden eng miteinander verknüpft: So bemängelt ein Junge, im Musikunterricht würden „leider immer nur so alte Sachen, nie mal die modernsten Lieder“ (Schüler 1, Interview 6) durchgenommen. Seine Mitschülerin fügt hinzu, dass sie dort manchmal „nur so ganz doofe Lieder“ (Schülerin 3, Interview 6) hörten, und der Junge ergänzt „so ganz alte!“ (Schüler 1, Interview 6).

Weitere vergleichende Analysen zu den Hörpräferenzen ergaben einen Rückgang der geschlechtsspezifischen Präferenzen für bestimmte Instrumente sowie insgesamt weniger Aussagen zu bestimmten Rhythmuspräferenzen bei den Jungen.

4.3 Fazit zur qualitativen Erhebung

Während die Kinder im zweiten Grundschuljahr noch relativ enge und präsente Vorstellungen von „Jungenmusik“ und „Mädchenmusik“ haben, scheinen diese Konzepte für die Mädchen im vierten Schuljahr keine Relevanz zur Beurteilung ihrer eigenen Musikpräferenzen mehr zu haben. Die Jungen geben jedoch weiterhin Auskunft über diese Stereotype und zeigen dabei insbesondere ein inhaltlich ausdifferenzierteres Konzept von „Jungenmusik“.
Obgleich sich in geschlechtsheterogenen Interviewgruppen im vierten Schuljahr eine neue Sensibilisierung für die normative Bewertung von musikalischen Geschlechtsstereotypen zeigt, offenbaren die Jungen aus geschlechtshomogenen Gruppen eine vehemente Ablehnung von „Mädchenmusik“. Gerade in dieser Abgrenzungstendenz kann eine Unterstützung der These gesehen werden, dass musikalische Präferenzäußerungen zumindest bei Jungen bereits in der Grundschule als Ausdruck sich entwickelnder sozialer Geschlechtsidentität funktionalisiert werden (vgl. Wilke 2012). So umschreibt ein Junge zum ersten IZP die von ihm rezipierte Musik folgendermaßen: „aber das ist auch /ehm/ so .. männerartig ist, so ganz, mit einer bösen Stimme und so.“

5 Zusammenführung der quantitativen und qualitativen Ergebnisse

Im Folgenden werden die Ergebnisse der quantitativen Erhebung mit denen der qualitativen zusammengeführt, wobei sich auf den wesentlich erscheinenden geschlechtsspezifischen Effekt beschränkt wird. Diese Zusammenführung der Ergebnisse erfolgt im Sinne eines „complementary model of triangulation“ (Erzberger & Kelle, 2003, S. 469). Davon ausgehend, dass die angewendeten quantitativen und qualitativen Methoden unterschiedliche Aspekte des Phänomens „Musikpräferenz“ beleuchten, dienen die herausgearbeiteten musikalischen Konzepte zur Ergänzung und Erklärung der vorgefundenen klingenden

Präferenzen. Dabei bilden die verbalen Präferenzäußerungen, die als Einstellungskomponente Teil musikalischer Konzepte sein können, den zentralen Anknüpfungspunkt zwischen musikalischen Konzepten und klingenden Präferenzen.

5.1 Beurteilung von „Klassik“ durch Jungen

Die negativeren Präferenzäußerungen der Jungen bezüglich des Faktors „Klassik“ zu MZP 1 bis 3 lassen sich durch die Interviewergebnisse auf zweifache Weise erklären:
(1) Zum einen äußern vor allem Jungen zum 1. Interviewzeitpunkt (IZP) im Rahmen des Konzeptes „Rockmusik“ persönliche Präferenzen für das Instrument Schlagzeug, das in den Musikbeispielen des Faktors „Klassik“ nicht vorkommt und somit eine Ablehnung nahelegt.
(2) Zum anderen steht das Konzept „Rockmusik“ im Zusammenhang mit den zum 1. IZP stark ausgeprägten geschlechtsspezifischen Konzepten „Jungenmusik“ und „Mädchenmusik“. So sind bei Jungen zum 1. IZP „Rockmusik“ und „Jungenmusik“ quasi deckungsgleich, während „Mädchenmusik“ von Jungen mit Charakteristika beschrieben wird, die auf Assoziationen mit „klassischer“ Musik schließen lassen (Cello, ruhiger etc.). Möglicherweise spielt somit bereits bei der negativeren Beurteilung des Faktors „Klassik“ durch die Jungen diese Beschreibung von „Mädchenmusik“ eine Rolle.

5.2 Auflösung der Faktorstruktur zu MZP 4

Für die Auflösung der an Stilkategorien angelehnten Faktorstruktur sowie für die Abnahme geschlechtsspezifisch differenzierter Präferenzurteile zum MZP 4 bietet die Interviewanalyse des 2. IZP zwei Erklärungsansätze:
(1) Die Begrenzung von „Jungenmusik“ auf „Rockmusik“ hat sich zum 2. IZP zugunsten eines ausdifferenzierteren und individuelleren Konzeptes aufgelöst. Diese Ausdifferenzierung zeigt sich auch in den verbalen Präferenzen der Jungen. Zudem lässt die veränderte normative Bedeutsamkeit von „Jungenmusik“ und „Mädchenmusik“ eine Zuordnung nach diesen beiden geschlechtsspezifischen Konzepten nicht mehr zu.

(2) Die Auflösung der Faktorstruktur fällt mit dem Aufkommen des Konzeptes „Chartsmusik" zusammen, auf das sich auffällig viele verbale Präferenzäußerungen beziehen. Die im Faktor „Pop" zusammengefassten Musikbeispiele weisen weder die für „Chartsmusik" bedeutsame Aktualität noch Bekanntheit auf und erfahren somit negativere Bewertungen. Des Weiteren bekräftigen die Interviews zum 2. IZP den vor allem bei Jungen beobachteten Peer-Gruppen-Effekt, der bereits für die quantitativen Erhebungen vermutet wurde. In Hinblick auf die sozio-kulturelle Prägung sollte daher vermutlich nicht von Geschlechter-, sondern treffender von Gender-Effekt gesprochen werden (vgl. Busch, 1998).

6 Fazit und Ausblick

Generell konnte anhand der quantitativen Daten eine an verschiedenen Stilkategorien angelehnte Strukturierung von kindlichen Präferenzäußerungen sowie ein altersabhängiger Rückgang von Offenohrigkeit zwar bestätigt werden (u.a. Gembris & Schellberg, 2007), jedoch „nicht im Sinne einer schroffen Ablehnung unbekannter Musik" (Louven, 2011, S. 54). Des Weiteren wurde im Übergang von der dritten zur vierten Schulklasse ein deutlicher Bruch dieser generellen Entwicklung beobachtet, der sich durch Individualisierungsprozesse erklären ließ. Zudem bestätigen die qualitativen Daten, dass von einer vielschichtigen Entwicklung und Ausdifferenzierung musikalischer Konzepte auszugehen ist und Präferenzäußerungen von sozio-kulturellen und individuellen Funktionen bedingt sind. Neben der Variable Alter wurde die Variable Geschlecht als stärkster Prädiktor für Präferenzäußerungen identifiziert, während die sonstigen Variablen kaum berichtenswerte Einflüsse ausübten. Der Geschlechtereffekt konnte im Wesentlichen auf die Jungen und deren Beurteilung von „Klassik" zurück geführt werden. Kopiez und Lehmann (2008) haben somit vermutlich aufgrund ihres Ausschlusses der „klassischen" Musikbeispiele keinen Geschlechtereffekt gefunden. Beutler-Prahm (2012) hat bereits anhand der Analyse des 1. IZP dargelegt, dass sich unsere Befunde gut in die allgemeinen entwicklungs- und sozialpsychologischen Erkenntnisse zur Geschlechtstypisierung einordnen lassen. Dies kann nun anhand der vergleichenden Befunde zum 2. IZP bestätigt werden. Nach diesen allgemeinen Erkenntnissen sind Stereotype zu „männlich" und „weiblich" unmittelbar vor Schuleintritt am rigidesten und werden im Verlauf der Grundschulzeit zunehmend

differenziert und flexibilisiert, wobei Jungen grundsätzlich stärkere Stereotypenfixierungen aufweisen als Mädchen (Maccoby, 2000; Ruble et al., 2006). Die vertiefende statistische Analyse dieses Effektes legt – basierend auf Zusammenhängen mit der kindlichen Persönlichkeit – eine Interpretation als jungenseitiger Effekt sozialer Erwünschtheit innerhalb der Peer-Gruppe nahe (Schurig, 2012). Fraglich ist, ob sich geschlechtsspezifische Unterschiede in der musikalischen Präferenzentwicklung bei Grundschulkindern nicht nur als Ergebnis unterschiedlicher Sozialisation, sondern auch als Grundlage für weitere Entwicklungs- und Sozialisationsprozesse und damit als sinnfälliger Schritt in der Entwicklung von Geschlechtsidentität deuten lassen. Den Musikpräferenzäußerungen käme dann vor allem bei Jungen eine identitätsbildende Funktion zu – in Anlehnung an die Beurteilung von North und Hargreaves (1999, S. 90): „music functions as a ‚badge' in adolescents' social cognitions". Diese Interpretation wird gestützt durch Wilke (2012), die eine Funktionalisierung der Musikpräferenz für „Gangsta Rap" bei Jungen mit Migrationshintergrund (4. Klasse) zur Inszenierung von Männlichkeit festgestellt hat. Schäfer und Sedlmeier (2009) haben aus musikalischen Präferenzäußerungen eine Vielzahl an Dimensionen (u.a. evaluative und behaviorale) mit jeweils spezifischen Funktionen für das Individuum abgeleitet. Aus den dargestellten Ergebnissen ließe sich jedoch folgern, dass musikbezogene Präferenzäußerungen bereits während der Grundschulzeit Ausdruck spezifischer individueller und sozialer Funktionen sind und diese Funktionalisierung somit früher einsetzt als bislang vermutet (vgl. Behne, 1997; Baacke, 1993). Fraglich erscheint zudem, inwieweit der bei Jugendlichen und Erwachsenen beobachtete Zusammenhang von Musikpräferenz und Persönlichkeits- sowie Identitätsbildung (North & Hargreaves, 1999; MacDonald et al., 2002; Rentfrow & Gosling, 2003; Rawlings & Ciancarelli, 1997; Zweigenhaft, 2008) auch bereits bei Kindern zu Beginn der Pubertät auffindbar ist und ob dieser Zusammenhang – analog zu Jugendlichen und Erwachsenen – ebenfalls geschlechtsspezifisch differenziert (u.a. Müller, 1999).

Für die eingangs zitierte Frage von Rentfrow et al. (2011), warum Musik so bedeutsam für Menschen sei, kann in Hinblick auf die bisherige Analyse geschlussfolgert werden, dass Musik Menschen bereits im jungen Kindesalter die Möglichkeit zur Darstellung und Ausbildung ihrer psychosozialen (Geschlechts-) Identität bietet und somit eine grundlegende Funktion im Entwicklungsprozess übernehmen kann. Dass solch eine frühzeitig auftretende

intensive Funktionalisierung von Musik ihre Bedeutsamkeit bis ins Erwachsenenalter beibehält, erscheint nachvollziehbar. Warum Mädchen anscheinend Musik weniger intensiv für die genannte Funktion nutzen als Jungen und wie sich die individuelle Funktionalisierung von Musik weiterentwickelt, sollte in zukünftigen Studien untersucht werden.
Im Gegensatz zu anderen empirischen Befunden (z.B. Louven, 2011) erfolgte in der vorliegenden Studie die Abnahme an Offenohrigkeit im Grundschulalter relativ unabhängig von schulischem oder privatem Instrumentalunterricht (vgl. Schurig et al., 2012). Dies wirft die musikpädagogisch relevante Frage auf, wie Instrumentalunterricht konzipiert sein sollte, damit „[k]ulturelle/musisch-ästhetische Bildung als integraler Bestandteil individueller und sozialer Identitätsentwicklung“ (Autorengruppe Bildungsberichterstattung, 2012, S.160) zu neuen Handlungsoptionen und vielschichtigen Erfahrungen führen kann.
Nicht hinreichend erforscht ist ebenfalls, welchen nachhaltigen Einfluss frühe musikalische Erfahrungen auf diese Entwicklungsprozesse haben (Hargreaves et al., 1995). Zudem stellt sich die grundlegende Frage, inwieweit Musikpräferenzäußerungen tatsächlich Musikpräferenz abbilden oder eher Ausdruck der oben genannten Funktion sind. Dies führt zur methodischen Herausforderung von Gruppenerhebungen: Der Einfluss von Klassenzugehörigkeit konnte statistisch berücksichtigt werden, während der soziale Peer-Gruppen-Effekt interessante Einblicke in das Zustandekommen der Präferenzäußerungen insbesondere der Jungen lieferte. Des Weiteren hat die Studie gezeigt, dass Kinder bereits zu Beginn der Grundschulzeit Konzepte von Musik ausgebildet haben, die nach musikalischen Stilkategorien und nach Geschlecht differenzieren, jedoch in unterschiedlichem Maße als Grundlage für Präferenzäußerungen dienen. Ungeklärt ist bislang, inwieweit individuelle und sozial erwünschte Präferenz interagieren und inwiefern diese die jeweiligen Konzepte prägen oder aber von den Konzepten geprägt werden. Für die Untersuchung der individuellen Präferenzen von Kindern sollte der musikbezogene und sozio-kulturelle Hintergrund auf Individualebene eingehend betrachtet werden, was auf der Grundlage der umfangreichen Datenbasis des SIGrun-Projektes in zukünftigen Analysen vorgenommen werden wird.
Zudem sollte der methodische Fokus zukünftiger Studien verstärkt auf die Analyse musikbezogener Verhaltensweisen gelegt werden (Downloads, CD-Sammlungen, Hörsituationen etc.; auch freiwillige Hördauern verschiedener Musikbeispiele, vgl. Software

„OpenEar“ von Louven & Ritter, 2011) und den in der vorliegenden Studie äußerst gewinnbringenden Mixed-Method-Ansatz verfolgen.

Um das Konstrukt Offenohrigkeit noch differenzierter fassen zu können, erscheint es sinnvoll, für die ersten drei Grundschuljahre die Modellierung mit Faktoren, die sich aus verschiedenen Stilkategorien zusammensetzen, und der Variable Geschlecht als Instrumentalvariable zu übernehmen und nach weiteren Prädiktorvariablen zu forschen. Für das vierte Grundschuljahr sollte nach Alternativen zu einer auf Stilkategorien basierenden Strukturierung gesucht werden. Hierfür können von der differenzierten Untersuchung musikalischer Konzepte und individueller Funktionen von Musik hilfreiche Ansatzpunkte sowie generell vertiefte Einblicke in die musikalische Urteilsbildung von Kindern erwartet werden.

Literaturverzeichnis

Asendorpf, J. B. (1998). FFFK – Fünf-Faktoren-Fragebogen für Kinder. Tests Info. Berlin: Humboldt-Universität, Institut für Psychologie.

Asendorpf, J. D. & van Aken, M. A. G. (2003). Validity of Big Five personality judgments in childhood: A 9 year longitudinal study. European Journal of Personality, 17, 1-17.

Autorengruppe Bildungsberichterstattung (2012): Bildung in Deutschland 2012. Ein indikatorengestützter Bericht mit einer Analyse zur kulturellen Bildung im Lebensverlauf. Bielefeld: Bertelsmann.

Baacke, D. (1993). Jugendkulturen und Musik. In H. Bruhn, R. Oerter & H. Rösing (Hrsg.), Musikpsychologie. Ein Handbuch (S. 228-237). Reinbek: Rowohlt.

Baumann, M. P. (Hrsg.) (1985). Musik der Türken in Deutschland. Kassel: Verlag Yvonne Landeck.

Behne, K.-E. (1975). Musikalische Konzepte. Zur Schicht- und Alterspezifität musikalischer Präferenzen. In E. Kraus (Hrsg.), Forschung in der Musikerziehung (S. 35-61). Mainz: Schott.

Behne, K.-E. (1986). Hörertypologien. Zur Psychologie jugendlichen Musikgeschmacks. Regensburg: Bosse.

Behne, K.-E. (1997). The development of „Musikerleben“ in adolescence. How and why young people listen to music. In I. Deliège & J. A. Sloboda (Hrsg.), Perception and Cognition of Music (S. 143-159). Hove, UK: Psychology Press.

Bentler, P. M. and C.-P. Chou (1987). Practical issues in structural modeling. Sociological Methods & Research 16 (1), 78-117.

Beutler-Prahm, B. (2012). Geschlechtsspezifische Aspekte in der musikalischen Präferenz bei Grundschulkindern. Unveröffentlichte Bachelor-Arbeit, Universität Bremen.

Bollen, K. A. (2002). Latent Variables in Psychology and the Social Sciences. Annual Review of Psychology, 53, 605-34.

Busch, V. (1998). Gender Studies. Eine Einführung. In S. Fragner, J. Hemming & B. Kutschke (Hrsg.), Gender Studies & Musik. Geschlechterrollen und ihre Bedeutung für die Musikwissenschaft. (S. 9-18). Regensburg: ConBrio.

Busch, V., Lehmann-Wermser, A. & Liermann, C. (2009). The Influence of Music Genre, Style of Singing, and Gender of Singing Voice on Music Preference of Elementary School Children. In J. Louhivuori et al. (Hrsg.). Proceedings of the 7th Triennial Conference of European Society for the Congitive Sciences of Music (ESCOM) in Jyväskylä, 2009 (S. 33-37), Jyvväskylä, Finland.

Costa, P. T. & McCrae, R. R. (1992). Revised NEO Personality Inventory (NEO PI-R) and NEO Five Factor Inventory. Professional Manual. Odessa, FL: Psychological Assessment Ressources.

Cremades, R., Oswaldo, L., & Lucia, H. (2010). Musical tastes of secondary school students with different cultural backgrounds: A study in the spanish north african city of Melilla. Musicae Scientiae, 14 (1), 121-141.

Delsing, M. J. M. H., Bogt, T. F. M. T., Engels, R. C. M. E. & Meeus, W. H. J. (2008). Adolescents' music preferences and personality characteristics. European Journal of Personality, 22, 109-130.

Ehmke, T., & Siegle, T. (2005). ISEI, ISCED, HOMEPOS, ESCS. Indikatoren der sozialen Herkunft bei der Quantifizierung von sozialen Disparitäten. Zeitschrift für Erziehungswissenschaft, 8, 521-539.

Eijck, K. van (2001). Social differentiation in musical taste patterns. Social Forces, 79 (3), 1163-1184.

Erzberger, C. &. Kelle U. (2003). Making inferences in mixed methods: The rules of Integration. In A. Tashakkori und C. Teddlie (Hrsg.), Handbook of mixed methods in social & behavioral research (S. 457-488). Thousand Oaks, Calif. [u.a.]: SAGE.

Gembris, H. (2005). Musikalische Präferenzen. In R. Oerter & T. H. Stoffer (Hrsg.), Enzyklopädie der Psychologie, Vol. 2, Spezielle Musikpsychologie (S. 279-342). Göttingen: Hogrefe.

Gembris, H., & Schellberg, G. (2003). Musical preferences of elemantary school children. Paper presented at the 5th Triennial Conference of the European Society for the Cognitive Sciences of Music. Hannover, 2003.

Gembris, H. & Schellberg, G. (2007). Die Offenohrigkeit und ihr Verschwinden bei Kindern im Grundschulalter. Musikpsychologie, 19, 71-92.

Greve, M. (2003). Die Musik der imaginären Türkei. Musik und Musikleben im Kontext der Migration aus der Türkei in Deutschland. Stuttgart, Weimar: J. B. Metzler.

Hargreaves, D. J. (1982). The development of aesthetic reactions to music. Psychology of Music (Special issue), 51-54.

Hargreaves, D. J. (1987). Development of liking for familiar and unfamiliar melodies. Bulletin of the Council for Research in Music Education, 91, 65-69.

Hargreaves, D. J., Comber, C. & Colley, A. (1995). Effects of age, gender, and training on musical preferences of British secondary school students. Journal of Research in Music Education, 43 (3), 242-250.

Hargreaves, D. J., North, A. C., & Tarrant, M. (2006). Musical preference and taste in childhood and adolescence. In G. E. McPherson (Hrsg.), The child as musician: A handbook of musical development (S. 135-154). New York: Oxford University Press.

Henninger, J. C. (1999). Ethnically diverse sixth graders' preferences for music of different cultures. Texas Music Education Research, 37-43.

Hu, L. & Bentler, P. M. (1998). Fit Indices in Covariance Structure Modeling: Sensitivity to Underparameterized Model Misspecification. Psychological Methods, 3, 424-453.

Jaedtke, W. (2000). Popmusik als Epochenstil. Versuch einer musikhistorischen und musiktheoretischen Aufarbeitung. In H. Rösing & T. Phleps (Hrsg.). Populäre Musik im kulturwissenschaftlichen Diskurs [=Beiträge zur Popularmusikforschung 25/26] (S. 201-216), Karben: Coda.

Kopiez, R., & Lehmann, M. (2008). The 'open-earedness' hypothesis and the development of age-related aesthetic reactions to music in elementary school children. British Journal of Music Education, 25 (2), 121-138.

Kulin, S. & Schwippert, K. (2012). Kooperationsbeziehungen im JeKi-Kontext: Beweggründe zur Kooperation und Merkmale gemeinsamer Reflexion methodischer und didaktischer Fragen. In J. Knigge & A. Niessen (Hrsg.), Musikpädagogisches Handeln. Begriffe, Erscheinungsformen, politische Dimensionen [= Musikpädagogische Forschung 33] (S. 152-171). Essen: Die Blaue Eule.

LeBlanc, A. (1991). Some unanswered questions in music preference research. Contribution to Music Education, 18, 66-73.

LeBlanc, A., Sims, W. L., Siivola, C., & Obert, M. (1996). Music style preferences of different age listeners. Journal of Research in Music Education, 44 (1), 49-59.

Lehmann-Wermser, A. & Jessel-Campos, C. (2013). Aneignung von Kultur. Wege zu kultureller Teilhabe und zur Musik. In A. Hepp & A. Lehmann-Wermser (Hrsg.), Transformation des Kulturellen. Prozesse des gegenwärtigen Kulturwandels (S. 129-144). Wiesbaden: vs-Verlag,.

Leopold, E. (2012). Urteilshomogenität und Klassengemeinschaft – Ein Beitrag zur Offenohrigkeitshypothese. Musikpsychologie, 22, 74-90.

Louven, C. (2011). Mehrjähriges Klassenmusizieren und seine Auswirkungen auf die „Offenohrigkeit" bei Grundschulkindern. Eine Langzeitstudie. Diskussion Musikpädagogik, 50 (11), 48-59.

Louven, C. & Ritter, A. (2011). Hargreaves' „Offenohrigkeit" – Ein neues, softwarebasiertes Forschungsdesign. Beitrag zur AMPF-Tagung 2011 in Stuttgart (S. 275-299).

Maccoby, E. (2000). Psychologie der Geschlechter. Sexuelle Identität in den verschiedenen Lebensphasen. Translated by E. Vorspohl. Stuttgart: Klett-Cotta.

Mayring, P. (2007). Qualitative Inhaltsanalyse. Grundlagen und Techniken. Weinheim: Beltz.

MacDonald, R., Hargreaves, D. J. & Miell, D. (2002). Musical Identities. Oxford: Oxford University Press.

MPLUS (Version 6.11). [Computer Software]. Los Angeles, CA: Muthén & Muthén.

Müller, R. (1999). Musikalische Selbstsozialisation. In J. Fromme, S. Kommer, J. Mansel & K.-P. Treumann (Hrsg.), Selbstsozialisation, Kinderkultur und Mediennutzung (S. 113-125). Oladen: Leske + Budrich.

Nonte, S. & Schwippert, K. (2012). Musikalische und sportliche Profile an Grundschulen – Auswirkungen auf Klassenklima und Selbstkonzept. Beiträge empirische Musikpädagogik, 3 (1) [Verfügbar unter: http://www.b-em.info/index.php?journal=ojs&page=article&op=view&path[]=72&path[]=208].

North, A. C. & Hargreaves, D. J. (1999). Music and adolescent identities. Music Education Research, 1, 75-92.

Peterson, R. A. (1992). Understanding audience segmentation: From elite and mass to omnivore and univore. Poetics, 21, 243-258.

Rawlings, D., & Ciancarelli, V. (1997). Music preference and the five-factor model of the NEO Personality Inventory. Psychology of Music, 25, 120-132.

Reinecke, J. (2005). Strukturgleichungsmodelle in den Sozialwissenschaften, München: Oldenbourg.

Rentfrow, P. J. & Goslings, S. D. (2003). The do re mi's of everyday life: The structure and personality correlates of music preference. Journal of Personality and Social Psychology, 84, 1236-1256.

Rentfrow, P. J., Goldberg, L. R. & Levitin, D. J. (2011). The structure model of musical preferences: A five-factor model. Journal of Personality and Social Psychology, 100 (6), 1139-1157.

Ruble, D. N., Martin, C. L. & Berenbaum, S. A. (2006). Gender development. In W. Damon & R. M. Lerner (Hrsg.), Handbook of Child Psychology, Vol. 3, 6th ed. (S. 858-932). Hoboken: Wiley.

Sakai, W. (2011). Music preferences and family language background: A computer-supported study of children's listening behavior in the context of migration. Journal of Research in Music Education, 59 (2), 174-195.

Schäfer, T., & Sedlmeier, P. (2009). From the functions of music to music preference. Psychology of Music, 37, 279-300.

Schurig, M. (2012). Response Bias und Messinvarianz in einem Urteil zu musikalischer Präferenz. Hinter der Messinvarianz. Vortrag gehalten auf der 77. Tagung der Arbeitsgruppe für Empirische Pädagogische Forschung in Bielefeld, 2012.

Schurig, M., Busch, V. & Strauß, J. (2012). Effects of Structural and Personal Variables on Children's Development of Music Preference. In E. Cambouropoulos, C. Tsougras, P. Mavromatis & K. Pastiadis (Hrsg.), Proceedings of the 12th International Conference on Music Perception and Cognition (ICMPC) and the 8th Triennial Conference of European Society for the Cognitive Sciences of Music (ESCOM) in Thessaloniki, 2012 (S. 896-902).

Statistisches Bundesamt (2009). Bevölkerung und Erwerbstätigkeit. Ausländische Bevölkerung. Ergebnisse des Ausländerzentralregisters, Fachserie 1, Reihe 2. Wiesbaden [verfügbar unter:

https://www.destatis.de/DE/Publikationen/Thematisch/Bevoelkerung/MigrationIntegration/AuslaendBevoelkerung2010200087004.pdf?__blob=publicationFile].

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

Teo, T., Hargreaves, D. J., & Lee, J. (2008). Musical preference, identification, and familiarity: A multicultural comparison of secondary students from Singapore and the United Kingdom. Journal of Research in Music Education, 56 (1), 18-32.

Wagschal, J., Schulte, K. & Busch, V. (2010). Big Five bei Grundschulkindern. Poster präsentiert auf der Tagung der Deutschen Gesellschaft für Psychologie in Bremen 2010.

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

Wilke, K. (2012). Bushido oder bunt sind schon die Wälder?! Musikpräferenz von Kindern in der Grundschule. Münster: Lit.

Wurm, M. (2006). Musik in der Migration. Beobachtungen zur kulturellen Artikulation türkischer Jugendlicher in Deutschland. Bielefeld: transcript.

Zweigenhaft, R. (2008). A Do Re Mi Encore. A Closer Look at the Personality Correlates of Music Preference. Journal of Individual Differences, 29 (1), 45-55.

BEITRAG 4

Erschienen in (Zitierweise):

Schurig, M., Wendt, H., Kasper, D. & Bos, W. (2015). Fachspezifische Stärken und Schwächen von Viertklässlerinnen und Viertklässlern in Deutschland im europäischen Vergleich. In H. Wendt, T. C. Stubbe & K. Schwippert (Hrsg.), *10 Jahre international vergleichende Schulleistungsforschung in der Grundschule. Vertiefende Analysen zu IGLU und TIMSS 2001 bis 2011* (S. 35–54). Münster: Waxmann.

Relevanz:

Dieser Beitrag fokussiert auf eine kompetenzdomänenübergreifende Betrachtungsweise von Schulleistungsergebnissen in Rahmen internationaler Schulleistungsstudien und integriert dabei Indikatoren welche ihrerseits in einem IRM abgeleitet wurden. Dabei wird ein internationales latentes Profilmodell als Referenzmodell abgeleitet und für spezifische Analysen im deutschen Kontext übertragen, um eine hohe Generalisierbarkeit herzustellen. Verschiedene Hintergrundmerkmale werden verrechnet, um deren Relevanz für die verschiedenen Profile zu bestimmen. Im Gegensatz zum Beitrag 2 ist hier kein längsschnittlicher, sondern ein querschnittlicher Referenzrahmen mit einer landesspezifischen Schachtelung gegeben.

II. Fachspezifische Stärken und Schwächen von Viertklässlern in Deutschland und im europäischen Vergleich

Michael Schurig, Heike Wendt, Daniel Kasper & Wilfried Bos

1. Einführung

Internationale Schulvergleichsstudien ermöglichen den Vergleich von schulischer Leistung zwischen Bildungssystemen. Üblicherweise fokussieren die Schulvergleichsstudien dabei auf bestimmte Kompetenzbereiche: In der *Internationalen Grundschul-Lese-Untersuchung* (IGLU) etwa wird alle fünf Jahre das Leseverständnis von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe untersucht und die Studie *Trends in International Mathematics and Science Study* (TIMSS) untersucht alle vier Jahre die mathematischen und naturwissenschaftlichen Leistungen von Schülerinnen und Schülern am Ende der vierten und achten Jahrgangsstufe (Bos, Tarelli, Bremerich-Vos & Schwippert, 2012; Bos, Wendt, Köller & Selter, 2012). Für eine vertiefende Einführung siehe Schwippert, Stubbe, Wendt & Bos (in diesem Band). Im Jahr 2011 wurden erstmals die beiden Schulleistungsstudien TIMSS und IGLU simultan in 37 Staaten und Regionen durchgeführt (Martin & Mullis, 2013). Durch diese gemeinsame Erhebung der Fachleistungen von Schülerinnen und Schülern der vierten Jahrgangsstufe in den Bereichen Mathematik, Naturwissenschaften und Leseverständnis ist es erstmals möglich, nationale Schulleistung in den Grundschulen fachübergreifend im internationalen Vergleich zu betrachten.

In der nationalen und internationalen Berichtlegung (Bos, Tarelli et al., 2012, Bos, Wendt, Köller et al., 2012; Martin & Mullis, 2013) werden die Ergebnisse einzelner Kompetenzdomänen in Kompetenzstufen beziehungsweise durch das Erreichen internationaler Benchmarks getrennt voneinander dargestellt. Ein domänenübergreifender internationaler Vergleich ist zurzeit eher unüblich. Benchmarks und Kompetenzstufen bieten eine gute Interpretationsfähigkeit der Ergebnisse (Bos, Voss & Goy, 2009), allerdings sind die Verfahren im Kontext der Darstellung substantiell verknüpfter Leistungsergebnisse nur bedingt tauglich, da es kaum möglich ist gegenseitige Abhängigkeiten der beobachteten Konstrukte zu beschreiben. Von Mullis (2013) wurden zur domänenübergreifenden Betrachtung von Leistungen in Leseverständnis, Mathematik und Naturwissenschaften Prozentanteile von Schülerinnen und Schülern berichtet, die verschiedene Kompetenzniveaustufen erreichen. Einen methodisch anderen Zugang zur Bildung von fächerübergreifenden Leistungsklassen bietet die latente Profilanalyse (LPA). Im Unterschied zu dem Vorgehen von Mullis (2013) kann mit der LPA auch die Güte der gewählten Klassifikation bestimmt werden und die abgebildeten Profile können in aufbauenden Sekundäranalysen weiterverwendet werden.

Erste Analysen mit der LPA zur domänenübergreifenden probabilistischen Klassifikation der domänenspezifischen Leistungsergebnisse wurden von Bos et al. (Bos, Wendt, Ünlü, Valtin, Euen, Kasper & Tarelli, 2012a; Bos, Wendt, Ünlü, Valtin, Euen, Kasper & Tarelli, 2012b [59]) für Leistungsdaten in Deutschland vorgenommen: Die Schülerinnen und Schüler wurden anhand der Ähnlichkeit ihrer Testwerte in den Domänen Mathematik, Naturwissenschaften und Leseverständnis sieben Gruppen (Leistungsprofilen) zugewiesen. Zum Beispiel bilden Schülerinnen und Schüler mit herausragenden Leistungen in allen drei Leistungsdomänen ein Profil. Durch den nationalen

[59] Die Beiträge sind deckungsgleich, daher wird fortfolgend nur noch einer der beiden Beiträge zitiert.

Bezugsrahmen dieser Profilbildung sind allerdings international vergleichende Aussagen nicht möglich.

Um diese Forschungslücke zu schließen, wird das von Bos et al. verwendete Verfahren auf 17 europäische Staaten erweitert: Zunächst wird ein europäisches Profilmodell über alle Staaten hinweg bestimmt, um daran anschließend für jedes Land separat ein Modell zu bestimmen, welche sich in Abweichung zu Bos et al. an den Kennwerten des europäischen Modells orientieren. Durch dieses Vorgehen kann festgestellt werden, wie sich die für Deutschland gefundenen Profile im europäischen Rahmen einordnen lassen. Durch die simultane Einbeziehung von Hintergrundmerkmalen der Schülerinnen und Schüler bei den LPA-Analysen ist es außerdem möglich, nationale Zusammenhangsmuster mit denen anderer europäischen Staaten zu vergleichen.

2. Theorie und Forschungsstand

Die individuelle Fähigkeit zum Lernen wird als fächerübergreifende Handlungskompetenz verstanden, welche kognitive Grundvoraussetzung und Kenntnisse, sowie Fertigkeiten und Strategien beim Erreichen von Ergebnissen zusammenführt (Weinert, 1999). Wenn das Lernen eine fächer- oder domänenübergreifende (Baumert, Klieme, Neubrand, Prenzel, Schiefele et al. 2000) Disposition ist, darf unterstellt werden, dass der Kompetenzerwerb im Grundschulalter in den Bereichen Lesen, Mathematik und Naturwissenschaften in Teilen parallel verläuft, schließlich ist auch die Förderung fachspezifischer sowie fächerübergreifender Kompetenzen gleichsam für die Grundschulbildung zentral. Ansätze des fächerverbindenden Unterrichts (Gudjons & Traub, 2012; Peterßen, 2000) betonen, dass jede Schülerin und jeder Schüler seine Realität nicht innerhalb eines Fachkorridors versteht, sondern mittels überfachlicher tragfähiger Konstrukte (Jonen & Jung, 2007; Lenzen, 1996). Es lässt sich also vermuten, dass es basierend auf den überfachlichen Kompetenzen und Dispositionen gemeinsame Entwicklungen in den Kompetenzdomänen Lesen, Mathematik und Naturwissenschaften gibt.

Auffällig ist, dass der an Schulen unterrichtete Fächerkanon überall in Europa vergleichbar ist (Mullis, Martin, Minnich, & Drucker et al. 2012; Mullis, Martin, Minnich, & Stanco et al., 2012). Da die administrativ vorgegebenen Instruktionszeiten entlang der Kompetenzdomänen allerdings zwischen den europäischen Staaten variieren (Tabelle II.1), könnte die Entwicklung – so lassen auch die Ergebnisse von IGLU/TIMSS vermuten – fachspezifischer Leistungen national durchaus spezifisch ausfallen, zumindest wenn die unterschiedlichen Instruktionszeiten mit einer unterschiedlichen curricularen Bedeutung der Domänen in den nationalen Schulsystemen korrespondiert.

Tabelle II.1: Domänenspezifische Instruktionszeiten pro Teilnehmerland

	Lesen	Mathematik	Naturwissenschaften
Finnland	nicht spezifiziert	nicht spezifiziert	nicht spezifiziert
Nordirland	nicht spezifiziert	nicht spezifiziert	nicht spezifiziert
Ungarn	32%-42%	17%-23%	4%-8%
Irland	18%	13%	4%
Tschechische Republik	nicht spezifiziert	nicht spezifiziert	nicht spezifiziert
Deutschland	variiert pro Bundesland (ca.19%)	variiert pro Bundesland (ca.19%)	variiert pro Bundesland (ca.9%)
Slowakei	35%	19%	12%
Portugal	32%	30%	20%
Schweden	nicht spezifiziert	nicht spezifiziert	nicht spezifiziert
Italien	nicht spezifiziert	20%	10%
Österreich	30%	ca. 18%	ca. 2%-3%
Slowenien	20%	21%	13%
Rumänien	24%-29%	15%-20%	5%-10%
Litauen	26%	19%	4%
Polen	nicht spezifiziert	nicht spezifiziert	nicht spezifiziert
Spanien	25%	15%-19%	6%-8%
Malta	15%	19%	8%

Angaben der Nationalen Forschungskoordinatoren TIMSS/IGLU (Mullis et al., 2012a & 2012b)

Zusammenhänge zwischen domänenspezifischen Leistungsergebnissen konnten in TIMSS, IGLU und in der Studie *Programme for International Student Assessment* (PISA; Prenzel, Sälzer, Klieme & Köller, 2013) beobachtet werden. Die für PISA 2003 in Deutschland (PISA Konsortium, 2004) berichteten latenten Zusammenhänge zwischen Mathematik und Lesen liegen bei $r = .77$ und zwischen Mathematik und Naturwissenschaften bzw. Naturwissenschaften und Lesen bei $r = .88$. Bei einer methodisch vergleichbaren Operationalisierung der Daten aus TIMSS und IGLU 2011 konnten latente Korrelationen von $r = .54$ zwischen Mathematik und Lesen, $r = .66$ zwischen Mathematik und Naturwissenschaften, sowie $r = .74$ zwischen Naturwissenschaften und Lesen beobachtet werden (Bos et al., 2012a).

In der erweiterten internationalen Berichtlegung zu TIMSS und IGLU (Mullis, 2013) wurden zur gemeinsamen Betrachtung der Domänen die Anteile der Schülerinnen und Schüler zusammengefasst, die internationale Benchmarks erreicht haben. Beispielhaft sind in Tabelle II.2 die Prozentanteile der Schülerinnen und Schüler abgetragen, welche die hohe internationale Benchmark erreicht haben. Es konnte nachgewiesen werden, dass die meisten Staaten heterogene Anteile pro Domäne in den hohen Benchmarks aufwiesen. Relativ wenige Staaten hatten ähnliche Anteile von Schülerinnen und Schülern innerhalb einheitlicher Benchmarks über alle drei Domänen hinweg.

Tabelle II.2: Anteile der Schülerinnen und Schüler in Deutschland die innerhalb verschiedenen Domänen die hohe internationale Benchmark erreicht haben

Domäne	Anteile (SE)
Alle drei Domänen	23% (1.3)
Lesen	46% (1.3)
Mathematik	37% (1.4)
Naturwissenschaften	39% (1.5)

Sofern ein vollständiger Vergleich der theoretisch möglichen Kombinationen von Kompetenzniveaustufen der drei Domänen vorgenommen werden soll, ergeben sich 125 Kombinationsmöglichkeiten (Bos et al., 2012a). Mittels dieser Herangehensweise konnte für Deutschland beobachtet werden, dass bei nur 2.9 Prozent der Schülerinnen und Schüler sowohl Leistungen der verschiedenen Bereiche auf den untersten und gleichzeitig auf den obersten Kompetenzstufen zugeordnet werden konnten. Nur bei 5.3 Prozent der Schülerinnen und Schüler streute die Leistung über drei Kompetenzstufen. Demnach lassen sich bei einem großen Teil der Schülerschaft in Deutschland die Leistungen in den drei Domänen auf nahezu den gleichen Kompetenzstufen einordnen.

Das Vorgehen von Bos et al. (Bos et al., 2012a) erlaubt es einzuschätzen, wie groß der Anteil leistungsheterogener Schülerinnen und Schüler generell ist. Insbesondere die Anteile von Schülerinnen und Schüler mit heterogenen fachspezifischen Stärken und Schwächen können leicht ermittelt werden. Aber die Zuordnung zu Leistungsklassen auf Basis der Kompetenzniveaustufen berücksichtigt zum einen nicht die probabilistische Zuordnung der Schülerinnen und Schüler zu den Kompetenzstufen, zum anderen wird auch nicht berücksichtigt, wie weit die Schülerinnen und Schüler von den normativ festgesetzten *cut-off*-Punkten weg liegen (Hartig, 2008). Damit sind quantitative Aussagen über die Güte dieses Klassifikationssystems (und ein Vergleich mit alternativen Klassifikationssystemen anhand dieser Gütemaße) nur schwer ableitbar. Eine parallele Betrachtung aller Domänen mit einer empirischen Profilzuordnung über die latente Profilanalyse vermeidet diese Schwierigkeiten.

Dementsprechend verwundert es nicht, dass Bos et al. (Bos et al., 2012a) neben der qualitativen Zuordnung der Schülerinnen und Schüler zu kompetenzübergreifenden Klassen auch eine quantitative Zuordnung anhand der latenten Profilanalyse vorgenommen haben. Dabei konnten sieben Profile abgeleitet werden, die sich vor allem hinsichtlich ihres Niveaus unterschieden. Fachspezifische Stärken und Schwächen der Schülerinnen und Schüler im Sinne von relativ hoher Leistung in einer oder zwei Domänen und relativ niedrigen Leistungen in einer oder zwei anderen Domänen (also sich kreuzenden Profillinien) konnten nicht beobachtet werden. Diese Ergebnisse decken sich also mit den vorangestellten Forschungsbefunden, nach denen eine hohe Konkordanz zwischen den Leistungen in den drei Domänen Mathematik, Naturwissenschaften und Lesen besteht.

Neben einer Analyse der Leistungskennwerte betrachteten Bos et al. (Bos et al., 2012a) auch den Zusammenhang zwischen den sieben Leistungsprofilen und verschiedenen Hintergrundmerkmalen. Dabei konnten signifikante Unterschiede zwischen den Profilen beobachtet werden: Nur etwa 12 Prozent der Familien von Kindern der Leistungstypen mit geringen und sehr geringen Leistungen können der dritten Sozialschicht (Akademiker, Techniker und Führungskräfte) zugeordnet werden, in den stärksten beiden Profilen sind es über 50 Prozent. Hinsichtlich der Bildungsnähe der Elternhäuser zeigt sich, dass grundsätzlich für die höheren Leistungstypen der Anteil von Kindern aus bildungsnahen Elternhäusern höher ausfällt. Die positiven Einstellungen zum Lernen differieren zwischen den Profilen nur für die Domäne Lesen, nicht aber in Mathematik und den Naturwissenschaften und das domänenspezifische Selbstkonzept unterscheidet sich nicht bei benachbarten Profilen.

3. Forschungsfragen

Insgesamt spricht die bisherige Forschung dafür Zusammenhänge zwischen den verschiedenen Kompetenzdomänen anzunehmen, die über die fachspezifischen Entwicklungen hinaus gedacht werden können und mit der Lebenswelt und den dispositionalen Lernbedingungen der Schülerinnen und Schüler zusammenhängen. Allerdings beziehen sich die hier dargestellten Ergebnisse auf die Schülerschaft in Deutschland, und der Zusammenhang zwischen den bei Bos et al. abgeleiteten sieben Profiltypen und den Hintergrundmerkmalen erfolgte sukzessiv (im ersten Schritt wurden die Profiltypen gebildet, im zweiten Schritt wurden die Profiltypen mit Hintergrundmerkmalen in Zusammenhang gebracht), weswegen die Ergebnisse aus statistisch-methodischer Sicht als weniger effizient einzustufen sind als bei einer parallelen Modellierung von Kompetenzprofilen und Zusammenhängen zu den Hintergrundmerkmalen. Damit stellen sich folgende Forschungsfragen:

1. Welche Profiltypen ergeben sich für die europäische Schülerschaft, wenn die latente Profilanalyse auf den europäischen Datensatz von TIMSS und PIRLS angewendet wird?
2. Wie lassen sich die Ergebnisse für Deutschland im Vergleich zu diesem europäischen Referenzmodell einordnen?
3. Wie lassen sich die Profile in Deutschland durch den sozio-ökonomischen und kulturellen Hintergrund sowie die positive Einstellung zum Lernen und das fachspezifische Selbstkonzept der Schülerinnen und Schüler beschreiben, wenn deren Einfluss parallel zu der Klassifikation geschätzt wird?

Aufgrund der latenten Korrelationen zwischen den Kompetenzdomänen wird erwartet, dass sich auch für die europäische Schülerschaft latente Profile abbilden, die sich vor allem durch Niveauunterschiede auszeichnen; kreuzende Profile sollten unter Berücksichtigung der mittleren Korrelationen und der geringen beobachteten Prozentanteile von Schülerinnen und Schülern deren Leistungen über mehrere Kompetenzstufen streuten hingegen eher nicht resultieren. Da sich die Verteilung der Leistungswerte im internationalen Kontext maßgeblich durch die mittlere Leistungsfähigkeit unterscheidet wird erwartet, dass sich dies in den Prozenten der Profile wiederfindet. In Deutschland werden also größere Volumina in den oberen Profilen erwartet als zum Beispiel in Rumänien oder Slowenien und geringere als zum Beispiel in Finnland oder Nordirland.

Bezüglich der Zusammenhänge zwischen den Profiltypen und den simultan berücksichtigten Hintergrundmerkmalen wird erwartet, dass sich zwar die im bivariaten Bereich bekannten Abhängigkeitsmuster auch in den Profilen wiederfinden lassen, aber aufgrund der effizienteren Schätzmethode sollten sich diese Abhängigkeitsmuster deutlich klarer ausweisen lassen.

4. Daten und Methoden

Die Stichprobe zur Ableitung des europäischen Profilmodells setzt sich aus den TIMSS und IGLU-Daten von 17 europäischen Staaten zusammen (N=74.868 Viertklässlerinnen und Viertklässler). In diese Stichprobe wurden alle europäischen Staaten eingeschlossen, die sowohl an TIMSS als auch an IGLU 2011 teilgenommen haben und die eine ausreichende Ausschöpfungsquote aufwiesen, um repräsentative Aussagen über die Gesamtheit der Grundschülerinnen und Grundschüler in ihrem Land treffen zu können. Bei den 17 Staaten handelt es sich um: Deutschland, Finnland, Irland, Italien, Litauen, Malta, Nordirland, Österreich, Polen, Portugal, Rumänien, Schweden, die Slowakei, Slowenien, Spanien, die Tschechische Republik und Ungarn.

Der Datensatz für Deutschland umfasst 3928 Schülerinnen und Schüler aus 197 Schulen. (vgl. Schwippert, Stubbe & Wendt, in diesem Band).

Zur Bildung des europäischen und des nationalen Profilmodells wurde die Latente Profilanalyse (LPA) nach Gibson (1966) und Lazarsfeld und Henry (1968) verwendet. Im Ergebnis der LPA resultieren latente Profile. Die Anzahl der Profile wird dabei über Informationskriterien und nach inhaltlichen Gesichtspunkten bestimmt. Die Güte der Modellpassung wird auf Basis von

Klassifikationsfehler, Entropie und Pseudo-Reliabilität wiedergegeben (Rost, 2004; Bacher & Vermunt, 2010). Die relative Entropie (0-1) beschreibt die Sicherheit der Klassifikation und die Pseudo-Reliabilität (0-1) die durchschnittliche Klassenzugehörigkeits-wahrscheinlichkeit. Hohe Werte deuten ein hohes Maß an Sicherheit in der Schätzung der Klassenzugehörigkeit an. Diese Werte sind wie der Reliabilitätskoeffizient α zu interpretieren. Der Klassifikationsfehler ist die die umgekehrte und für die Klassengröße relativierte durchschnittliche Klassenzugehörigkeit und kann als probabilistische prozentuale Fehlklassifikation pro Land interpretiert werden.

Als Indikatorvariablen für die LPA wird auf die auf einer internationalen Metrik skalierten *plausible-Values* der drei Leistungsdomänen Mathematik, Naturwissenschaften und Lesen zurückgegriffen (vgl. Foy, Brosman & Galia, 2012; Rubin, 1987; Schwippert, Stubbe & Wendt, in diesem Band). Der robuste *Maximum Likelihood* Schätzer (MLR) wird verwendet, um die Klassenzugehörigkeit zu schätzen (Muthén, 2004). Insgesamt wurden drei Modelle mit der LPA geschätzt:

- Modell 1: Ausgangspunkt für Modell 1 ist der europäische Gesamtdatensatz. Die Daten wurden mit dem *senate weight* gewichtet (Schwippert, Stubbe & Wendt, in diesem Band). Die hierarchische Struktur der Staaten wurde über die Einführung einer Landesebene in der Verrechnung berücksichtigt (Vermunt, 2003).
- Modell 2: Ausgangspunkt für Modell 2 sind die national spezifischen Datensätze. Auf jeden dieser Datensätze wurde separat eine LPA angewendet. Die Daten jeden Landes wurden mit dem *house weight* gewichtet (Schwippert, Stubbe & Wendt, in diesem Band). Um Vergleiche der resultierenden nationalen Profile zwischen den einzelnen Staaten zu ermöglichen, wurden sowohl die Anzahl der Profile als auch die Mittelwerte der latenten Klassen auf die Werte von Modell 1 fixiert.
- Modell 3: Ausgangspunkt für Modell 3 ist der europäische Gesamtdatensatz (Modell 3a) und der Datensatz der deutschen Teilstichprobe (Modell 3b). Auf jeden dieser Datensätze wurde separat eine LPA angewendet, wobei in Modell 3a mit dem *senate weight*, und im Modell 3b mit dem *house weight* gewichtet wurde. Unabhängig vom gewählten Datensatz wurden sowohl die Profilanzahl als auch die Profilmittelwerte auf die Werte von Modell 1 fixiert. In Ergänzung zu Modell 1 bzw. von Modell 2 wurden neben den Leistungswerten auch Hintergrundmerkmale in die LPA eingeführt. Diese Hintergrundmerkmale können im Sinne von Prädiktoren der Profilzugehörigkeit verstanden werden.

In Bezug auf die Hintergrundmerkmale für Modell 3 wurden Faktoren ausgewählt, die unterschiedliche Kapitalformen in Anlehnung an Bourdieu (1992) abbilden und die in vorangegangenen Untersuchungen zum Einfluss von Hintergrundmerkmalen auf Basis der Erhebungen von TIMSS und IGLU Verwendung gefunden haben (Schwippert, Wendt & Tarelli, 2012; Stubbe, Tarelli & Wendt, 2012; Tarelli, Schwippert & Stubbe, 2012; Wendt, Stubbe & Schwippert, 2012).

Die heimische Ausstattung mit Büchern wurde über ‚mehr als 100 Bücher im Haushalt' (1) und , maximal 100 Bücher im Haushalt' (0) und abgebildet ($\text{Europa}_{\text{Ausprägung }1}$: 27.8% (*SE* = 0.3); $\text{Deutschland}_{\text{Ausprägung }1}$: 34.9% (*SE* = 1.5)). Der sozio-ökonomische Status der Familie wurde über den beruflichen Status der Eltern operationalisiert. ‚Schülerinnen und Schüler mit mindestens einem Elternteil mit einem Berufshintergrund als Akademikerin oder Akademiker oder Führungskraft' (1) wurden ‚allen anderen Schülerinnen und Schülern' (0) entgegengestellt ($\text{Europa}_{\text{Ausprägung }1}$: 36.8% (*SE* = 0.3); $\text{Deutschland}_{\text{Ausprägung }1}$: 32.0% (*SE* = 1.4)). Der Bildungshintergrund der Eltern wurde über die Ausprägungen ‚mindestens ein Elternteil hat einen Universitätsabschluss oder Vergleichbares' (1) und ‚alle anderen Eltern' (0) gebildet ($\text{Europa}_{\text{Ausprägung }1}$: 27.8% (*SE* = 0.3); $\text{Deutschland}_{\text{Ausprägung }1}$: 28.0% (*SE* = 1.6)). Da der Migrationshintergrund nicht in allen europäischen Staaten erhoben worden ist, wurde zusätzlich die zu Hause gesprochene Sprache als Indikator der kulturellen Fremdheit herangezogen[60]. Dabei wurde darüber operationalisiert, ob die Sprache des Tests nur ‚manchmal oder

[60] Die Familiensprache wurde in Slowenien nicht erhoben.

nie‘ (1) oder ‚fast immer oder immer‘ (0) im Elternhaus gesprochen wird (Europa$_{\text{Ausprägung 1}}$: 19.8% (*SE* = 0.3); Deutschland$_{\text{Ausprägung 1}}$: 19.6% (*SE* = 1.1)).

Weiterhin wurden Indizes zu fachspezifischen Einstellungen zum Lernen und zum fachspezifischen Selbstkonzept als Hintergrundmerkmale verarbeitet, wie es bei Bos et al. (Bos et al., 2012a) vorgenommen wurde. Die Skalierungen sind bei Martin & Mullis (2012) dargestellt. Für die fachspezifischen positiven Einstellungen zum Lernen wurden jeweils die höchsten Kategorien der international gebildeten Indices abgeleitet. Zuletzt wurde das fachspezifische Selbstkonzept (median-split diskretisiert, um ein eher hohes Selbstkonzept von eher niedrigem Selbstkonzept zu trennen) eingeführt. Alle Skalen erreichen international und für Deutschland ausreichend gute Reliabilitätsmaße. In der Tabelle II.3 sind die Kennwerte der Skalen sowie die Prozentanteile im internationalen Mittel abgetragen. Zusätzlich sind die Prozentanteile in Deutschland angegeben.

Weiterhin wird die Full Information Maximum Likelihood-Methode (FIML) angewendet, die in Mplus implementiert ist, um Populationsparameter und Standardfehler basierend auf allen beobachteten Daten zu schätzen (Lüdtke, Robitzsch, Trautwein & Köller, 2007).

Tabelle II.3: Übersicht der verwendeten unabhängigen Variablen für die europäischen Stichprobe und die deutsche Teilstichprobe

Hintergrund -variable	Skalen-informationen	α_{Europa}	$\alpha_{\text{Deut-schland}}$	Operationalisierung	Inter-nationaler Prozentanteil auf der höchsten Kategorie $\%_1$ (SE)	Prozentanteil auf der höchsten Kategorie in Deutschland $\%_1$ (SE)
Positive Einstellung zum Lernen- Lesen	6 Items; Wertelabel: 1 = ‚Stimme völlig zu‘ bis 4 = ‚Stimme überhaupt nicht zu‘; Beispielitem: „Ich lese gerne.“; Internationaler cut-off der höchsten Kategorie=11.0	.87	.85	‚Die Schülerinnen und Schüler haben die höchste Kategorie des internationalen Students like Reading- Index erreicht‘ (1) vs. ‚Andere Schülerinnen und Schüler‘ (0)	48.6% (0.3)	50.2% (1.0)
Positive Einstellung zum Lernen-Mathematik	5 Items; 1 = ‚Stimme völlig zu‘ bis 4 = ‚Stimme überhaupt nicht zu‘; Beispielitem: „Ich lerne gern Mathematik.“; Internationaler cut-off der höchsten Kategorie=11.3	.86	.88	‚Die Schülerinnen und Schüler haben die höchste Kategorie des internationalen Students like learning Mathematics-Index erreicht‘ (1) vs. ‚Andere Schülerinnen und Schüler‘ (0)	40.4% (0.3)	32.9% (0.8)
Positive Einstellung zum Lernen- Natur-wissenschaften	5 Items; 1 = Stimme völlig zu bis 4 = Stimme überhaupt nicht zu; Beispielitem: „Ich lerne gern im Sachunterricht.“; Internationaler cut-off der höchsten Kategorie=9.7	.79	.89	‚Die Schülerinnen und Schüler haben die höchste Kategorie des internationalen Students like learning Science- Index erreicht‘ (1) vs. ‚Andere Schülerinnen und Schüler‘ (0)	43.1% (0.3)	44.1% (1.4)
Positives Selbstkonzept-Lesen	8 Items; 1 = ‚Stimme völlig zu‘ bis 4 = ‚Stimme überhaupt nicht zu‘; Beispielitem: „Normalerweise bin ich gut im Lesen.“	.77	.76	‚Die Schülerinnen und Schüler liegen im oberen Teil eines median-split der Skala zum fachbezogenen Schülerselbstkonzept‘ (1) vs. ‚Andere Schülerinnen und Schüler‘ (0)	53.8% (0.3)	58.4% (1.0)
Positives Selbstkonzept-Mathematik	7 Items; 1 = Stimme völlig zu bis 4 = Stimme überhaupt nicht zu; Beispielitem: „Normalerweise bin ich gut in Mathematik.“	.85	.89	‚Die Schülerinnen und Schüler liegen im oberen Teil eines median-split der Skala zum fachbezogenen Schülerselbstkonzept‘ (1) vs. ‚Andere Schülerinnen und Schüler‘ (0)	59.2% (0.3)	60.7% (0.9)
Positives Selbstkonzept-Natur-wissenschaften	6 Items; 1 = ‚Stimme völlig zu‘ bis 4 = ‚Stimme überhaupt nicht zu‘; Beispielitem: „Normalerweise bin ich gut im Sachunterricht.“	.75	.84	‚Die Schülerinnen und Schüler liegen im oberen Teil eines median-split der Skala zum fachbezogenen Schülerselbstkonzept‘ (1) vs. ‚Andere Schülerinnen und Schüler‘ (0)	61.1% (0.3)	68.5% (1.0)

5. Ergebnisse

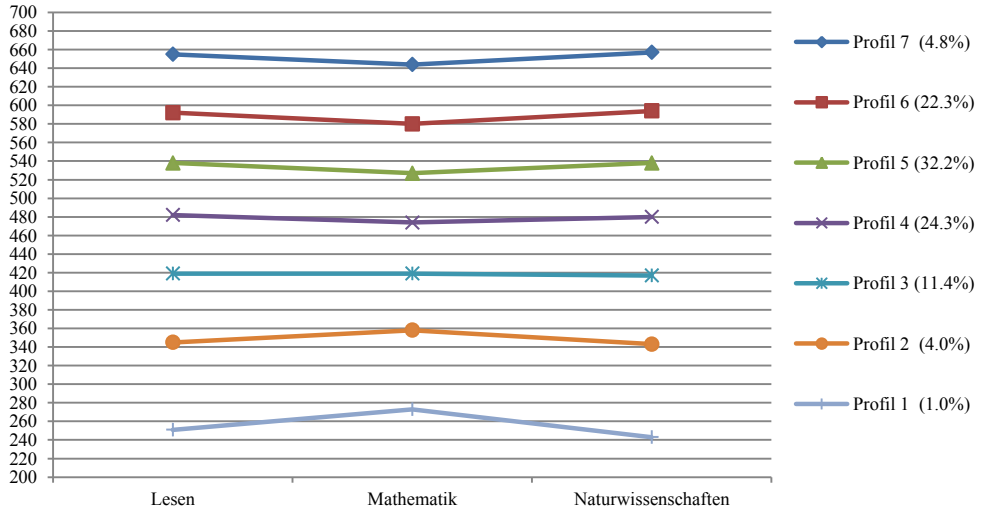
Im ersten Schritt wurde die Zahl der latenten Profile für das europäische Referenzmodell abgeleitet (Modell 1). Die Informationskriterien für Modelle mit verschiedener Anzahl an Profilen sind in Tabelle II.4 abgetragen.

Tabelle II.4: Modellvergleiche für das europäischen Referenzmodell (N=74868)

Modell	Anzahl der Leistungsprofile	Mittlere -2*Log-likelihood	BIC	CAIC	Parameter
I	3	-1243507	2487170	2487041	14
II	4	-1231360	2462922	2462756	18
III	5	-1224201	2448649	2448446	22
IV	6	-1220405	2441102	2440862	26
V	7	-1218364	2437065	2436789	30
VI	8	-1217217	2434816	2434503	34

Die Veränderung in der *Log-Likelihood* ist generell geringer als 1.0 Prozent. Die Veränderung der Informationskriterien *Bayes Information Criterion* (BIC; Schwarz, 1978) und des für Stichprobengrößeneffekte korrigierten *Akaike Information Criterion* (CAIC; Bozdogan, 1987), sowie die *Log-Likelihood* zeigen dabei einen Knick bei dem Wechsel von sieben auf acht Profile auf unter 0.1 Prozent. Deswegen und weil bei weiterer Anhebung der Profilanzahl die Zellbesetzungen in den Randprofilen auf unter 1 Prozent der Grundgesamtheit fallen würde, wurde einem europäischen Modell mit sieben Profilen der Vorzug gegeben. Die Profillinien dieses Modells sind in Abbildung II.1 abgetragen.

Abbildung II.1: Europäische Leistungsprofile der Viertklässlerinnen und Viertklässler in Europa in Lesen, Mathematik und den Naturwissenschaften



Es wird deutlich, dass die vorliegenden Profile invariant bezüglich der Rangreihe sind, das heißt es bilden sich keine Überschneidungen in den Kompetenzdomänen ab. Ähnlich wie bei den Kompetenzniveaustufen der einzelnen Domänen kann also auch in Bezug auf die generelle

Leistungsfähigkeit der Schülerinnen und Schüler von aufeinander aufbauenden Leistungsniveaus der Schülerinnen und Schüler ausgegangen werden. Die Prozentanteile und die exakten Profilmittelwerte des europäischen Profilmodells sind in Tabelle II.5 aufgezeigt. Es kann nachvollzogen werden, dass das Leistungsprofil 1 und das Leistungsprofil 2 am geringsten besetzt sind, gemeinsam erreichen sie ungefähr den Umfang des Leistungsprofils 7. Die Leistungsprofile 4, 5 und 6 sind mit einer deutlichen Häufung im Leistungsprofil 5 am stärksten besetzt. Das Leistungsprofil 5 bildet in allen Domänen die größte Nähe zu den Mittelwerten der europäischen Vergleichsgruppe aus, womit das Modell als plausibel erachtet werden kann.

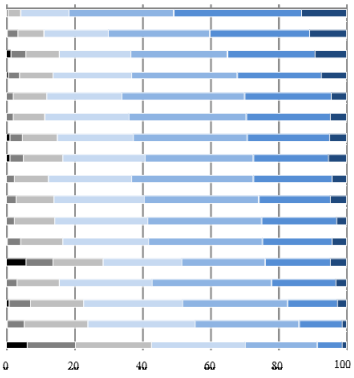
Tabelle II.5: Mittelwerte und Verteilung auf die latenten Profile von Viertklässlerinnen und Viertklässlern in Europa

Leistungs-profil	Prozentanteile		Gesamt Lesen		Gesamt Mathematik		Gesamt Natur-wissenschaften	
	n	(%)	M	(SE)	M	(SE)	M	(SE)
1	783	(1.0)	251	(27.5)	273	(31.3)	243	(26.2)
2	2993	(4.0)	345	(23.5)	358	(20.8)	343	(23.7)
3	8514	(11.4)	419	(16.9)	419	(13.6)	417	(15.8)
4	18193	(24.3)	482	(11.8)	474	(10.1)	480	(11.9)
5	24140	(32.2)	538	(9.0)	527	(9.1)	538	(9.2)
6	16681	(22.3)	592	(7.3)	580	(7.8)	594	(7.4)
7	3564	(4.8)	655	(7.0)	644	(8.2)	657	(7.2)

In Abbildung II.2 sind die prozentualen Verteilungen der Schülerinnen und Schüler auf die internationalen Profile pro Land bei Zugrundelegung der national spezifischen LPA (Modell 2) abgetragen.

Abbildung II.2: Prozentuale Verteilung auf die sieben Leistungsprofile innerhalb der Staaten mit fixierten latenten Mittelwerten

Land	n	Profil 1 %	(SE)	Profil 2 %	(SE)	Profil 3 %	(SE)	Profil 4 %	(SE)	Profil 5 %	(SE)	Profil 6 %	(SE)	Profil 7 %	(SE)
Finnland	4540	0.0	0.0	0.6	0.1	3.9	0.3	14.2	0.5	30.6	0.7	37.7	0.7	13.1	0.5
Nordirland	3469	0.4	0.1	3.1	0.3	7.7	0.5	18.9	0.7	29.8	0.8	29.4	0.8	10.6	0.5
Ungarn	5149	1.4	0.2	4.5	0.3	9.8	0.4	20.9	0.6	28.4	0.6	25.8	0.6	9.2	0.4
Irland	4383	0.7	0.1	3.5	0.3	9.6	0.4	23.2	0.6	31.1	0.7	24.8	0.7	7.2	0.4
Tschechische Republik	4105	0.2	0.1	2.0	0.2	9.8	0.5	22.3	0.6	36.0	0.7	25.3	0.7	4.5	0.3
Deutschland	3928	0.1	0.0	2.2	0.2	9.3	0.5	24.6	0.7	34.6	0.8	24.6	0.7	4.7	0.3
Slowakei	5561	1.1	0.1	3.6	0.3	10.5	0.4	22.4	0.6	33.3	0.6	24.2	0.6	4.9	0.3
International	74686	1.1	0.0	4.1	0.1	11.6	0.1	24.2	0.2	31.7	0.2	22.3	0.2	5.1	0.1
Schweden	4482	0.1	0.0	2.5	0.2	9.9	0.4	24.4	0.6	35.8	0.7	23.3	0.6	4.0	0.3
Italien	4125	0.2	0.1	2.8	0.3	11.0	0.5	26.6	0.7	33.6	0.7	21.1	0.6	4.6	0.3
Österreich	4587	0.1	0.0	2.4	0.2	11.9	0.5	27.4	0.7	33.4	0.7	22.1	0.6	2.7	0.2
Slowenien	4433	0.4	0.1	3.9	0.3	12.5	0.5	25.1	0.7	33.4	0.7	20.5	0.6	4.2	0.3
Rumänien	4643	5.8	0.3	8.1	0.4	14.6	0.5	23.2	0.6	24.3	0.6	19.3	0.6	4.6	0.3
Litauen	5284	0.1	0.1	3.1	0.2	12.4	0.5	27.3	0.6	35.0	0.7	19.0	0.5	3.1	0.2
Polen	4962	0.8	0.1	6.4	0.3	15.6	0.5	29.3	0.6	30.7	0.7	14.5	0.5	2.6	0.2
Spanien	4433	0.5	0.1	5.0	0.3	18.5	0.6	31.6	0.7	30.6	0.7	12.6	0.5	1.2	0.2
Malta	3492	6.2	0.4	14.3	0.6	22.3	0.7	27.8	0.8	20.9	0.7	7.6	0.4	1.1	0.2



% der Schülerinnen und Schüler in Profil 1
% der Schülerinnen und Schüler in Profil 2
% der Schülerinnen und Schüler in Profil 3
% der Schülerinnen und Schüler in Profil 4
% der Schülerinnen und Schüler in Profil 5
% der Schülerinnen und Schüler in Profil 6
% der Schülerinnen und Schüler in Profil 7

Hervorgehobene Fälle unterscheiden sich statistische signifikant (p<=.05) vom internationalen Mittelwert .
Die Länder sind nach der Summe der Zellvolumen in den Profilen 6 und 7 geordnet.
Die Standardfehler unterschätzen die Populationsvarianz.
Unterschiede zu 100% resultieren aus Rundungsfehlern.

Abbildung II.2 kann entnommen werden, dass eher leistungsstarke Staaten über alle Domänen hinweg hohe Prozentanteile von Schülerinnen und Schülern in hohen Profilen und eher leistungsschwache Staaten entsprechend erhöhte Prozentanteile in niedrigeren Profilen zeigen. So zeigt zum Beispiel Finnland besonders hohe Anteile von Schülerinnen und Schülern die in allen drei

Domänen auf oder über der Kompetenzniveaustufe IV liegen und Rumänien zeigt einen erhöhten Prozentanteil mit der Kompetenzniveaustufe I in allen Domänen. In der vorliegenden Operationalisierung spiegelt sich dies in besonders hohen Prozentanteilen von Schülerinnen und Schülern in Finnland in den Profilen Fünf bis Sieben und Schülerinnen und Schülern in Rumänien in den Profilen Drei bis Fünf wider. Die Profilvolumina in Deutschland liegen knapp oberhalb des internationalen Mittels und unterscheiden sich nicht signifikant von diesen.

Nationalen Prozentanteile der einzelnen Profile wurden hervorgehoben, wenn sie sich statistisch signifikant von den Prozentanteilen des europäischen Profilmodells unterscheiden. Für kein einzelnes Profil kann ein signifikanter Unterschied im Prozentanteil ausgemacht werden. Profilübergreifend kann ein geringerer Prozentanteil von Kindern in den Leistungsprofilen 2 und 3 (insgesamt 4.2% weniger) und ein erhöhter Anteil in den Leistungsprofilen 4 bis 6 beobachtet werden (insgesamt 5.7% mehr). Generell kann für Deutschland eine gute Übertragbarkeit der internationalen Mittelwerte festgestellt werden.

Insgesamt lässt sich sowohl die Güte des europäischen Modells als auch die Güte aller nationalen Modelle als gut beschreiben, wenn man als Gütekriterien die relative Entropie, den Klassifikationsfehler, und die Pseudo-Reliabilität verwendet. Das europäische Modell (Modell 1) erreicht eine Entropie von .81, einen Klassifikationsfehler von .16 (Bacher & Vermunt, 2010), und eine Pseudo-Reliabilität von .87 (Rost, 2004). Das nationale Modell (Modell 2) für Deutschland hat eine Entropie von .85, einen Klassifikationsfehler von .17 und eine Pseudo-Reliabilität von .90. Die Gütemaße der nationalen Modelle sind in Tabelle II.6 abgetragen.

Tabelle II.6: Gütemaße der nationalen latenten Profilmodelle mit fixierten Mittelwerten und fixierter Profilanzahl (7 Profile)

Land	Entropie[1]	Pseudo Reliabilität[1]	Klassifikations-Fehler[2]
Finnland	0.83	0.83	0.15
Slowenien	0.86	0.90	0.16
Ungarn	0.83	0.88	0.16
Deutschland	0.85	0.90	0.17
Österreich	0.84	0.88	0.18
Slowakei	0.84	0.89	0.18
Portugal	0.84	0.89	0.18
Litauen	0.85	0.89	0.18
Tschechische Republik	0.84	0.88	0.19
Polen	0.84	0.89	0.19
Italien	0.82	0.87	0.20
Irland	0.80	0.86	0.21
Schweden	0.82	0.86	0.21
Nordirland	0.76	0.83	0.23
Spanien	0.81	0.86	0.24
Rumänien	0.73	0.80	0.27
Malta	**0.73**	**0.79**	**0.31**

[1]: Werte nahe 1 geben einen guten Fit an; [2]: Rate der falsch klassifizierten Schülerinnen und Schüler bei gegebenen Latenten Profilen

Gemessen an diesen Gütekriterien passen die nationalen Modelle in Rumänien und Malta am wenigsten gut zu den ermittelten Daten, während die nationalen Modelle von Finnland und Slowenien

die höchsten Passungen (Fit) aufweisen.

Der Tabelle II.7 können die Ergebnisse zum Modell 3a der LPA (europäisches Modell mit Hintergrundmerkmalen) entnommen werden. Dabei wurden die Prozentanteile abgetragen, den Schülerinnen und Schüler mit den entsprechenden sozio-ökonomischen Hintergründen innerhalb der sieben beobachteten Profile einnehmen.

Tabelle II.7: Anteile von Schülerinnen und Schülern unter Berücksichtigung unterschiedlicher sozioökonomischer und sozialer Hintergründe innerhalb der europäischen Leistungsprofile

	Profil 1		Profil 2		Profil 3		Profil 4		Profil 5		Profil 6		Profil 7	
Sozioökonomischer & kultureller Hintergrund	%	(SE)	%	(SE)	%	(SE)	%	(SE)	%	(SE)	%	(SE)	%	(SE)
Familie mit hohem SES[1]	11.7	(1.2)	14.9	(0.7)	20.1	(0.4)	28.5	(0.4)	40.5	(0.5)	53.2	(0.5)	67.4	(0.9)
Familie mit hohem Bildungsniveau[2]	5.8	(0.5)	2.4	(0.6)	10.9	(0.4)	18.0	(0.3)	28.9	(0.4)	43.7	(0.4)	60.7	(0.9)
Hohe Zahl von Büchern zu Hause[3]	10.9	(1.1)	12.4	(0.6)	14.7	(0.4)	19.4	(0.3)	28.4	(0.4)	41.1	(0.4)	57.4	(0.8)
Familiensprache														
Familiensprache ist nicht die Landessprache[4]	45.0	(1.0)	38.9	(0.9)	30.5	(0.5)	23.4	(0.3)	16.8	(0.3)	11.6	(0.3)	7.3	(0.5)

1 Mindestens ein Elternteil hat einen Berufshintergrund als Akademiker oder Akademikerin oder Führungskraft

2 Mindestens ein Elternteil hat einen Universitätsabschluss

3 Im Elternhaus gibt es mehr als 100 Bücher

4 Im Elternhaus wird nur manchmal oder nie die Landessprache gesprochen

Generell kann abgeleitet werden, dass in den oberen Profilen deutlich häufiger ein vorteilhafter familiärer Hintergrund beobachtet werden kann. Es gibt dabei jedoch keine scharfen Sprünge in den prozentualen Anteilen. Nur bezüglich des Bildungsniveaus zwischen den Leistungsprofilen 1 und 2 ist keine Steigerung der Volumina zwischen höheren Profilen zu beobachten. Der Berufshintergrund der Eltern ist dabei der trennschärfste Indikator für eine Zugehörigkeit zu den oberen Profilen. Zu betonen ist hierbei, dass das Bildungsniveau eine besonders scharfe Trennung für die Zugehörigkeit zu den unteren und mittleren Profilen aufweist. Kontrastierend zum europäischen Modell sind in der Tabelle II.8 die entsprechenden Ergebnisse abgetragen, wenn Modell 3b (national spezifisches Modell für Deutschland mit Hintergrundmerkmalen) angenommen wird.

Tabelle II.8: Anteile von Schülerinnen und Schülern unter Berücksichtigung unterschiedlicher sozioökonomischer und sozialer Hintergründe innerhalb der Leistungsprofile in Deutschland

	Profil 1		Profil 2		Profil 3		Profil 4		Profil 5		Profil 6		Profil 7	
Sozioökonomischer & kultureller Hintergrund	%	(SE)	%	(SE)	%	(SE)	%	(SE)	%	(SE)	%	(SE)	%	(SE)
Familie mit hohem SES[1]	9.1	(1.9)	12.2	(1.3)	17.7	(5.3)	33.6	(1.8)	30.8	(3.5)	47.3	(2.3)	64.3	(2.1)
Familie mit hohem Bildungsniveau[2]	3.7	(0.9)	7.1	(4.5)	10.9	(1.1)	19.0	(3.6)	32.6	(1.8)	49.2	(3.9)	66.0	(2.1)
Hohe Zahl von Büchern zu Hause[3]	14.0	(4.5)	16.0	(1.2)	17.1	(1.6)	27.5	(3.9)	42.7	(1.6)	55.0	(2.0)	67.6	(3.0)
Familiensprache														
Familiensprache ist nicht die Landessprache[4]	34.5	(6.0)	40.3	(3.1)	32.0	(1.8)	19.4	(1.3)	13.2	(1.1)	9.0	(2.7)	5.8	(1.1)

1 Mindestens ein Elternteil hat einen Berufshintergrund als Akademiker oder Akademikerin oder Führungskraft

2 Mindestens ein Elternteil hat einen Universitätsabschluss

3 Im Elternhaus gibt es mehr als 100 Bücher

4 Im Elternhaus wird nur manchmal oder nie die Landessprache gesprochen

Für Deutschland ist, in Abweichung zu dem europäischen Modell, die Wahrscheinlichkeit dem Profil mit dem niedrigsten generellen Leistungsniveau (Leistungsprofil 1) zugeordnet zu werden für Kinder von Migranten (Familiensprach ist nicht deutsch) um knapp 10 Prozent geringer. Kinder aus Familien, in denen die Familiensprache nicht deutsch ist haben in Deutschland also eine deutlich geringere Wahrscheinlichkeit ein generell niedriges Leistungsniveau aufzuweisen als Kinder mit äquivalenten Charakteristiken in Europa insgesamt. Differenzen zum europäischen Modell können außerdem bezüglich der Ausstattung mit Büchern ausgemacht werden. In den Profilen Sechs und Sieben sind die Prozentanteile in Deutschland deutlich erhöht, wenn mehr als 100 Bücher im elterlichen Haushalt vorliegen.
In Tabelle II.9 finden sich die Ergebnisse von Modell 3a und Modell 3b in Bezug auf die fachspezifischen Selbstkonzepte und die positiven Einstellungen zum fachspezifischen Lernen.

Tabelle II.9: Anteile von Schülerinnen und Schülern mit positiver Einstellung zum Lernen und hohem akademischem Selbstkonzept innerhalb der europäischen Leistungsprofile

	Profil 1		Profil 2		Profil 3		Profil 4		Profil 5		Profil 6		Profil 7	
Positive Einstellung zum Lernen	%	(SE)	%	(SE)	%	(SE)	%	(SE)	%	(SE)	%	(SE)	%	(SE)
Lesen[1]	10.1	(1.1)	15.2	(1.2)	20.5	(0.4)	24.4	(0.3)	29.7	(0.3)	39.3	(0.4)	52.3	(0.8)
Mathematik[2]	24.7	(1.5)	36.2	(1.6)	44.7	(0.5)	46.8	(0.4)	48.3	(0.3)	50.3	(0.4)	54.4	(0.8)
Naturwissenschaften[3]	25.2	(1.6)	36.2	(0.9)	45.9	(0.5)	51.4	(0.4)	54.4	(0.3)	56.6	(0.4)	58.1	(0.8)
Positives Selbstkonzept[4]														
Lesen	12.7	(1.2)	22.7	(0.7)	33.5	(0.5)	44.9	(0.4)	53.3	(0.3)	71.6	(0.4)	84.6	(0.6)
Mathematik	23.2	(1.5)	34.8	(0.9)	43.1	(0.5)	51.6	(0.4)	61.8	(0.3)	73.2	(0.4)	84.5	(0.6)
Naturwissenschaften	24.8	(1.5)	35.8	(0.9)	45.8	(0.5)	56.1	(0.4)	65.0	(0.3)	72.5	(0.4)	78.1	(0.7)

1 Die Schülerinnen und Schüler haben die höchste Kategorie des internationalen Students like Reading- Index erreicht

2 Die Schülerinnen und Schüler haben die höchste Kategorie des internationalen Students like learning Mathematics- Index erreicht

3 Die Schülerinnen und Schüler haben die höchste Kategorie des internationalen Students like learning Science- Index erreicht

4 Die Schülerinnen und Schüler liegen im oberen Teil eines median-split der Skala zum fachbezogenen Schülerselbstkonzept.

Für das internationale Modell kann nachvollzogen werden, dass alle Werte in einer Rangreihe stehen, die Prozentanteile bei höheren Profilen also auch immer höhere Werte annehmen. Die positiven Einstellungen zum Lernen in allen Domänen differieren in den oberen Profilen in den Naturwissenschaften und Mathematik nur gering, also weisen europäische Schülerinnen und Schüler ab einem durchschnittlichen fachübergreifenden Leistungsniveau kaum mehr Differenzen auf. Die höchsten Differenzen zwischen den Profilen erreicht das positive Selbstkonzept im Lesen. Dieses sagt eine Profilzugehörigkeit über alle Profile gleichmäßig gut voraus. Die Intervalle zwischen den Profilen liegen dabei immer circa 10 Prozent.
Der Tabelle II.10 können die Ergebnisse von Modell 3a und Modell 3b entnommen werden, wenn die Lernfreude als Prädiktor in die LPA eingeführt wird.

Tabelle II.10: Anteile von Schülerinnen und Schülern mit positiver Einstellung zum Lernen und positivem fachspezifischem Selbstkonzept in Deutschland

	Profil 1		Profil 2		Profil 3		Profil 4		Profil 5		Profil 6		Profil 7	
Positive Einstellung zum Lernen	%	(SE)	%	(SE)	%	(SE)	%	(SE)	%	(SE)	%	(SE)	%	(SE)
Lesen[1]	13.8	(1.4)	17.5	(1.2)	22.6	(2.6)	28.4	(5.5)	38.5	(4.4)	51.9	(1.6)	62.3	(2.1)
Mathematik[2]	46.4	(2.0)	37.9	(1.6)	39.5	(3.1)	40.1	(6.2)	41.7	(4.4)	45.2	(1.6)	49.7	(2.1)
Naturwissenschaften[3]	50.4	(2.0)	53.1	(2.2)	51.3	(1.6)	55.8	(6.3)	61.1	(1.6)	63.5	(3.1)	64.3	(4.2)
Positives Selbstkonzept[4]														
Lesen	28.9	(5.6)	33.0	(1.9)	40.8	(1.6)	53.6	(4.5)	67.5	(1.5)	79.5	(2.7)	89.0	(1.4)
Mathematik	37.9	(6.2)	36.0	(3.1)	45.8	(2.0)	56.8	(1.6)	67.5	(1.5)	78.4	(3.7)	90.3	(1.3)
Naturwissenschaften	51.7	(2.0)	50.8	(1.7)	56.3	(3.2)	65.5	(6.0)	74.1	(4.0)	85.8	(1.2)	87.9	(1.4)

1 Die Schülerinnen und Schüler haben die höchste Kategorie des internationalen Students like Reading- Index erreicht

2 Die Schülerinnen und Schüler haben die höchste Kategorie des internationalen Students like learning Mathematics- Index erreicht

3 Die Schülerinnen und Schüler haben die höchste Kategorie des internationalen Students like learning Science- Index erreicht

4 Die Schülerinnen und Schüler liegen im oberen Teil eines median-split der Skala zum fachbezogenen Schülerselbstkonzept.

Bezüglich der positiven Selbstkonzepte und der positiven Einstellungen zum Lernen kann für die deutsche Stichprobe vermerkt werden, dass die Einstellung zum Lesenlernen in den oberen Profilen in Deutschland besonders hoch ist. Für Mathematik und die Naturwissenschaften ist hingegen die Einstellung in den niedrigen Profilen deutlich höher als im europäischen Durchschnitt. Es gelingt in Deutschland also trotz fachübergreifend unterdurchschnittlicher Leistung eine positive Einstellung zum Lernen zu erhalten. Die Einstellung zum Mathematiklernen variiert in Deutschland kaum zwischen den Profilen. Die fachspezifischen Selbstkonzepte sind in Deutschland generell eher hoch. Dies trifft insbesondere auf die niedrigeren Profile in den Bereichen Mathematik und Naturwissenschaften zu.

6. Fazit und Ausblick

Die Aufgabe von Schulleistungsstudien internationale Vergleiche von Schulleistung zu ermöglichen, kann sich nicht darauf beschränken einzelne Domänen zu betrachten, sondern muss auch domänenübergreifende Leistung und deren Abhängigkeitsstruktur in den Fokus nehmen. Um Vergleiche zu ermöglichen, wurden internationale Kompetenzprofile als Referenz abgeleitet (Modell 1), landesspezifische Ergebnisse mit diesen Referenzwerten abgeglichen (Modell 2) und Hintergrundmerkmale innerhalb der Modelle verrechnet (Modell 3). Multikompetenzprofile erlauben eine vertiefende und umfassendere Betrachtung der Wirkung schulischer Bildung jenseits des Spektrums einzelner Domänen. Komplexe Wirkungen zwischen den Leistungsprofilen und den

Hintergrundmerkmalen konnten so ökonomisch abgebildet und parallel international verglichen werden. Die Güte der Trennung kann statistisch eingeschätzt werden und die Einflüsse können übergreifender und nicht weniger differenziert betrachtet werden.

Ein zentraler Aspekt von Schulleistungsstudien ist es die Ergebnisse zieladäquat zusammengefasst und valide darzustellen. Die Ableitung von Leistungsprofilen ist eine Möglichkeit dies in angemessener Weise vor zu nehmen.

Auch im europäischen Kontext zeichnen sich sieben plausible Profile ab, wie auch in den Beiträgen von Bos et al. (Bos et al., 2012a). Im Kontrast zu den Ergebnissen von Mullis (2013) kann basierend auf den mittleren Leistungswerten pro Profil eine geringere Variabilität festgestellt werden als wenn die fachübergreifende Leistung durch über Kompetenzniveaustufen beobachtet wurden.

Für die Prozentanteile der Profile können für Deutschland substanzielle Differenzen zu dem Modell von Bos et al. (Bos et al., 2012a) ausgemacht werden, die insbesondere im europäischen Vergleich Schlussfolgerungen erlauben. Die hohen drei europäischen Leistungprofile sind in der Teilstichprobe in Deutschland geringer besetzt als bei Bos et al. und die niedrigen drei Leistungsprofile stärker, womit sich ein im europäischen Verglich erhöhtes Leistungsmittel abbildet. Referentiell zeigen sich für Deutschland im internationalen Vergleich vergleichsweise geringe Anteile an leistungsschwachen Kindern jedoch auch vergleichsweise geringe Anteile an sehr leistungsstarken Schülerinnen und Schülern.

Bezogen auf die Hintergrundvariablen kann von einer guten Trennung über die Profile gesprochen werden. Bekannte Zusammenhänge (Schwippert et al., 2012; Stubbe et al., 2012; Tarelli et al., 2012; Wendt et al., 2012) zwischen vorteilhaften Hintergrundmerkmalen zeichnen sich in den Profilen gut ab und unterstreichen damit die Validität dieser. Über die bivariaten Verrechnungen hinaus können weiterführende Ableitungen gemacht werden. In Bezugnahme auf den sozialen Gradienten des elterlichen Berufshintergrundes kann sowohl für Deutschland als auch für die europäische Stichprobe ein linearer Zusammenhang vermutet werden, wie bei Stubbe et al. (2012) und Wendt et al. (2012) unterstellt wird. Für die Anzahl an Büchern im Haushalt kann über die Ableitungen von Stubbe et al. (2012) und Wendt et al. (2012) hinaus beobachtet werden, dass das erhöhte Leistungsmittel für Schülerinnen und Schüler mit vorteilhafterem Hintergrund, in Deutschland insbesondere für die höheren Profile zutrifft, der Zusammenhang von Buchbesitz und Leistung also bei leistungsstärkeren Schülerinnen und Schülern besonders ausgeprägt ist.

Der von (Schwippert et al., 2012) und (Tarelli et al., 2012) beobachtete Leistungsvorsprung von Schülerinnen und Schülern, bei denen zu Hause häufiger die Testsprache gesprochen wird, konnte dahingehend ausdifferenziert werden, als dass es für das untere Profil in Deutschland vermutlich besser gelingt Kompensationsarbeit an Schulen zu leisten.

Interessant sind ebenso die Differenzen zwischen den Profilen in Deutschland und den europäischen Profilen bezüglich der Selbstkonzepte und der Lerneinstellungen. Generell ist die positive Einstellung zum Lernen in Deutschland in nahezu allen Profilen höher als in der europäischen Referenz. Leistungsstarke Schülerinnen und Schüler weisen in Deutschland eine besonders hohe positive Einstellung auf. Die Einstellungen in Mathematik und den Naturwissenschaften in den niedrigen Profilen sind besonders positiv. Die positiven Einstellungen zum Lernen in Mathematik und den Naturwissenschaften differieren dabei besonders wenig über die Profile hinweg, so dass davon ausgegangen werden kann, dass es im Schulsystem in Deutschland generell besser als im europäischen Vergleich gelingt, Schülerinnen und Schülern unabhängig von deren Leistung Lernfreude zu vermitteln. Es gelingt ebenso besser die Selbstkonzepte trotz geringerer individueller Leistung positiv zu erhalten. An dieser Stelle kann von einem positiven Zeugnis für das Schulsystem in Deutschland gesprochen werden. Positive Lerneinstellungen und Selbstkonzepte werden trotz eher geringer Leistungen erhalten. Dies gilt insbesondere für die Naturwissenschaften und Mathematik. Es muss aber angemerkt werden, dass es auch möglich ist, dies als eine mangelhafte Selbstreflexionsfähigkeit der Schülerinnen und Schüler zu deuten. Diese wichtige Differenzierung sollte in kommenden Zyklen

vertiefend inspiziert werden.

Die Wechselwirkungen zwischen Hintergrundmerkmalen und fachübergreifenden Leistungen differieren deutlich zwischen den Staaten: Einigen Bildungssystemen gelingt es anscheinend sehr gut, domänenübergreifend Leistung relativ unabhängig von Hintergrundmerkmalen zu ermöglichen.

Problematisch bleiben verschiedene Punkte. Zum ersten wird die hierarchische Schachtelung der Daten nicht vollständig berücksichtigt. Die Schul- und die Klassenebene bleiben unberücksichtigt. Zum zweiten wird beim europäischen Vergleich von Hintergrundmerkmalen und affektiv-behavioralen Skalen bisher forschungspragmatisch davon ausgegangen, dass diese über Staaten messinvariat sind, diese also innerhalb der Staaten vergleichbare Konstrukte erfassen. Eine Prüfung dessen steht noch aus. Zum dritten muss darauf hingewiesen werden, dass die Zusammenhänge nicht als kausal verstanden werden dürfen, da keine zeitliche Verlagerung der Erfassung der Hintergrundmerkmale und der Leistungsergebnisse vorliegt.

Als Rahmen für einen weiter vertiefenden modellbasierten europäischen Vergleich wären Daten über realisierte Instruktionszeiten von hohem Interesse. Relative Instruktionszeiten in der Landessprache, Mathematik und Naturwissenschaften könnten herangeführt werden, um Mittelwertsunterschiede innerhalb der latenten Profile zu erklären. Hier ist die vorliegende Datenbasis zu den Instruktionszeiten noch nicht ausreichend für eine vertiefende Analyse (Weinert & Helmke, 1997; Mullis, Martin, Minnich, & Drucker et al. 2012; Mullis, Martin, Minnich, & Stanco et al., 2012). Dazu müssten einzelne nationale Leistungsprofile mit frei variierenden Mittelwerten abgeleitet werden. In explorativen Analysen konnte bereits beobachtet werden, dass die Mittelwerte in anderen Staaten deutlich höhere Ausmaße annehmen als in Deutschland, wo die Profile verhältnismäßig homogen verlaufen.

Von Interesse kann es außerdem sein, Leistungsprofile einzelner Schulen herauszuarbeiten, entweder um die Variabilität, die auf einzelne Schulen zurückgeführt werden kann aus den Schätzungen zu entfernen, oder auch um spezifische Schulprofile abzuleiten und somit die Wirksamkeit pädagogischer Arbeit an Schulen besser beobachten und vergleichen zu können. Hierzu müssten die Schulen als Untersuchungseinheiten mit aufgenommen werden.

Wie Bos et al. (2012a) berichteten wäre es denkbar die Analyse fachspezifischer Stärken und Schwächen auch auf Subdimensionen, wie zum Beispiel Arithmetik und Geometrie in Mathematik oder Biologie und Physik/Chemie in den Naturwissenschaften, auszudehnen, aber eine Betrachtung der drei Hauptdomänen erschien bislang angemessen ohne das Modell überkomplex zu gestalten.

Während bislang (a) drei konditionale Scores abgeleitet werden (b) ein latentes Klassenmodell geschätzt wird und dies mit Regressoren verrechnet wird, ist es technisch möglich die latenten Klassen direkt aus einem multidimensionalen IRT Modell abzuleiten (Bacci, Bartolucci & Gnaldi, 2014). Hier gilt es zukünftig zu prüfen, ob die Modellfamilie auch für den Umgang mit fehlenden Werten und imputierten Datensätzen umgesetzt werden kann.

7. Literatur

Bacci, S., Bartolucci, F. & Gnaldi, M. (2014). A Class of Multidimensional Latent Class IRT Models for Ordinal Polytomous Item Responses. *Communications in Statistics - Theory and Methods, 43* (4), 787–800.

Bacher, J. & Vermunt, J. K. (2010). Analyse latenter Klassen. In C. Wolf & H. Best (Hrsg.), *Handbuch der sozialwissenschaftlichen Datenanalyse* (1. Aufl, S. 553–574). Wiesbaden: VS Verl. für Sozialwissenschaften.

Baumert, J., Klieme, E., Neubrand, M., Prenzel, M., Schiefele, U., Schneider, W. et al. (2000). *Die Fähigkeit zum Selbstregulierten Lernen als fächerübergreifende Kompetenz*. Berlin: Max-Planck-Institut für Bildungsforschung.

Bos, W., Tarelli, I., Bremerich-Vos, A. & Schwippert, K. (Hrsg.), (2012). *IGLU 2011. Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich*. Münster [u.a.]: Waxmann.

Bos, W., Voss, A. & Goy, M. (2009). Leistung und Leistungsmessung. In S. Andresen, R. Casale, T. Gabriel, R. Horlacher, S. Larcher Klee & J. Oelkers (Hrsg.), *Handwörterbuch Erziehungswissenschaft* (S. 563–576). Weinheim, Germany: Beltz.

Bos, W., Wendt, H., Köller, O. & Selter, C. (Hrsg.), (2012). *TIMSS 2011. Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich*. Münster: Waxmann.

Bos, W., Wendt, H., Ünlü, A., Valtin, R., Euen, B., Kasper, D. & Tarelli, I. (2012a). Leistungsprofile von Viertklässlerinnen und Viertklässlern in Deutschland. In W. Bos, I. Tarelli, A. Bremerich-Vos & K. Schwippert (Hrsg.), *IGLU 2011. Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* (S. 227–257). Münster [u.a.]: Waxmann.

Bos, W., Wendt, H., Ünlü, A., Valtin, R., Euen, B., Kasper, D. & Tarelli, I. (2012b). Leistungsprofile von Viertklässlerinnen und Viertklässlern in Deutschland. In W. Bos, H. Wendt, O. Köller & C. Selter (Hrsg.), *TIMSS 2011. Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* (S. 269–299). Münster: Waxmann.

Bourdieu, P. (1992). Ökonomisches Kapital – Kulturelles Kapital – Soziales Kapital. In P. Bourdieu (Hrsg.), *Die verborgenen Mechanismen der Macht* (S. 49–80). Hamburg: VSA-Verlag.

Bozdogan, H. (1987). Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika, 52* (3), 345–370.

Foy, P., Brosman, B. & Galia, J. (2012). Scaling TIMSS and PIRLS 2011 achievement data. In M. O. Martin & I. V. Mullis (Hrsg.), *Methods and Procedures in TIMSS and PIRLS 2011.* Chestnut Hill, Mass.: International Study Center, Lynch School of Education, Boston College. Zugriff am 02.01.2015. Verfügbar unter http://timssandpirls.bc.edu/methods/index.html

Gibson, W. A. (1966). Latent structure analysis and test theory. In P. F. Lazarsfeld & N. W. Henry (Hrsg.), *Readings in mathematical social science* (S. 78–88). Chicago: Science Research Associates.

Gudjons, H. & Traub, S. (2012). *Pädagogisches Grundwissen. Überblick - Kompendium - Studienbuch* (UTB Pädagogik, Bd. 3092, 11., grundlegend überarb. Aufl). Bad Heilbrunn: Klinkhardt.

Hartig, J. (2008). Psychometric models for the assessment of competencies. In J. Hartig, E. Klieme & D. Leutner (Hrsg.), *Assessment of competencies in educational contexts* (1. Aufl, S. 69–90). Cambridge, Mass.: Hogrefe.

Jonen, A. & Jung, J. (2007). *SINUS-Transfer Grundschule Naturwissenschaften Modul G 6: Fächerübergreifend und fächerverbindend unterrichten,* IPN. Zugriff am 05.03.2015. Verfügbar unter http://www.sinus-transfer.uni-bayreuth.de/fileadmin/MaterialienIPN/G6_gesetzt.pdf

Lazarsfeld, P. F. & Henry, N. W. (1968). *Latent structure analysis*. New York: Houghton, Mifflin.

Lenzen, D. (1996). *Handlung und Reflexion. Von pädagogischen Theoriedefizit zur reflexiven Erziehungswissenschaft* (Reihe Pädagogik). Weinheim: Beltz.

Martin, M. O. & Mullis, I. V. (Hrsg.), (2012). *Methods and Procedures in TIMSS and PIRLS 2011*. Chestnut Hill, Mass.: International Study Center, Lynch School of Education, Boston College.

Martin, M. O. & Mullis, I. V. (Hrsg.), (2013). *TIMSS and PIRLS 2011: Relationship Among Reading, Mathematics and Science Achievement at the Fourth Grade. Implications for Early Learning*. Chestnut Hill, MA: International Study Center, Lynch School of Education, Boston College.

Martin, M. O., Mullis, I. V., Foy, P. & Arora, A. (2012). *The PIRLS 2011 Students Like Reading Scale.* Zugriff am 12.02.2015. Verfügbar unter http://timssandpirls.bc.edu/methods/pdf/P11_R_Scales_SLR.pdf

Mullis, I. V. (2013). Profiles of Achievement Across Reading, Mathematics, and Science at the Fourth Grade. In M. O. Martin & I. V. Mullis (Hrsg.), *TIMSS and PIRLS 2011: Relationship Among Reading, Mathematics and Science Achievement at the Fourth Grade. Implications for Early Learning* (S. 13–58). Chestnut Hill, MA: International Study Center, Lynch School of Education, Boston College.

Mullis, Ina V. S., Martin, M. O., Minnich, C. A., Drucker, K. T. & Ragan, M. A. (Hrsg.), (2012). *PIRLS 2011 encyclopedia. Education policy and curriculum in reading*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center.

Mullis, Ina V. S., Martin, M. O., Minnich, C. A., Stanco, G. M., Arora, A., Centurino, V. A. et al. (Hrsg.), (2012). *TIMSS 2011 encyclopedia. Education policy and curriculum in mathematics and science ; Trends in International Mathematics and Science Study ; TIMSS*. Chestnut Hill, Mass: TIMSS & PIRLS International Study Center.

Muthén, B. O. (2004). Latent variable analysis. Growth mixture modeling and related techniques for longitudinal data. In D. Kaplan (Hrsg.), The Sage Handbook of quantitative methodology (S. 345-368). Thousand Oaks: Sage.

Peterßen, W. H. (2000). *Fächerverbindender Unterricht. Begriff - Konzept - Planung - Beispiele ; ein Lehrbuch* (EGS-Texte, 1. Aufl). München: Oldenbourg.

PISA Konsortium (2004). *PISA 2003. Der Bildungsstand der Jugendlichen in Deutschland : Ergebnisse des zweiten internationalen Vergleichs*. Münster: Waxmann.

Prenzel, M., Sälzer, C., Klieme, E. & Köller, O. (Hrsg.), (2013). *PISA 2012. Fortschritte und Herausforderungen in Deutschland*. Münster: Waxmann.

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

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys* (Wiley classics library). Hoboken, N.J: Wiley-Interscience.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics, 6* (2), 461–464.

Schwippert, K., Wendt, H. & Tarelli, I. (2012). Lesekompetenzen von Schülerinnen und Schülern mit Migrationshintergrund. In W. Bos, I. Tarelli, A. Bremerich-Vos & K. Schwippert (Hrsg.), *IGLU 2011. Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* (S. 191–207). Münster [u.a.]: Waxmann.

Stubbe, T. C., Tarelli, I. & Wendt, H. (2012). Soziale Disparitäten der Schülerleistungen in Mathematik und Naturwissenschaften. In W. Bos, H. Wendt, O. Köller & C. Selter (Hrsg.), *TIMSS 2011. Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* (S. 231–244). Münster: Waxmann.

Tarelli, I., Schwippert, K. & Stubbe, T. C. (2012). Mathematische und naturwissenschaftliche Kompetenzen von Schülerinnen und Schülern mit Migrationshintergrund. In W. Bos, H. Wendt, O. Köller & C. Selter (Hrsg.), *TIMSS 2011. Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* (S. 247–267). Münster: Waxmann.

Vermunt, J. K. (2003). Multilevel Latent Class Models. *Sociological Methodology, 33* (1), 213–239.

Weinert, F. E. (1999). *Konzepte der Kompetenz*. Paris: OECD.

Weinert, F. E. & Helmke, A. (Hrsg.), (1997). *Entwicklung im Grundschulalter*. Weinheim: Beltz, Psychologie Verlags Union.

Wendt, H., Stubbe, T. C. & Schwippert, K. (2012). Soziale Herkunft und Lesekompetenzen von Schülerinnen und Schülern. In W. Bos, I. Tarelli, A. Bremerich-Vos & K. Schwippert (Hrsg.), *IGLU 2011. Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* (S. 175–190). Münster [u.a.]: Waxmann.

Wendt, H., Tarelli, I., Bos, W., Frey, K. & Vennemann, M. (2012). Ziele, Anlage und Durchführung der Trends in International Mathematics and Science Study (TIMSS 2011). In W. Bos, H. Wendt, O. Köller & C. Selter (Hrsg.), *TIMSS 2011. Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* (S. 27–68). Münster: Waxmann.

BEITRAG 5

Erschienen in (Zitierweise):

Schurig, M., Glesemann, B. & Schröder, J. (2016). Dimensionen von Unterrichtsqualität. Die Generalisierbarkeit von Schülerurteilen über Fächer und Zeit. In R. Strietholt, W. Bos, H. G. Holtappels & N. McElvany (Hrsg.), *Jahrbuch der Schulentwicklung Band 19. Daten, Beispiele und Perspektiven* (S. 30–56). Weinheim: Beltz Juventa.[61]

Relevanz:

Die Qualität von Unterricht wird, unabhägig von den Beobachtungseinheiten, üblicherweise mehrdimensional betrachtet. Unklar ist, wie stabil latente Strukturen zwischen Fächern und über die Zeit sind, ob also Vergleiche in diesen Dimensionen vorgenommen werden können. Dieser Beitrag bearbeitet die Frage nach der Generalisierbarkeit einer einheitlichen mehrdimensionalen latenten Struktur über vier Messzeitpunkte und zwischen vier verschiedenen Fächern. Dafür werden Analysen der Messinvarianz vorgenommen. Damit wird insbesondere der Frage nach der Evidenz für externale Konstruktvalidität nachgegangen.

[61] Es hier abgedruckte Fassung ist eine vorläufige Version. Die finale Version kann unter der angegebenen Zitation gefunden werden.

II Dimensionen von Unterrichtsqualität

Die Generalisierbarkeit von Schülerurteilen über Fächer und Zeit

Michael Schurig, Birte Glesemann & Jan Schröder

Die Qualität des Unterrichts ist ein bedeutender Faktor der Schuleffektivitätsforschung. Die Beobachtung guten Unterrichts steht dabei vor den Fragen nach Kriterien und validen Instrumenten zur Beobachtung dieser. Hinsichtlich der Erfassung der Kriterien gibt es verschiedene Ansätze, bei denen fragebogenbasierte Abfragen der Schülerwahrnehmung dominieren. Die Analysen der Schülerdaten können je nach Forschungsinteresse auf Klassen- oder auf Schülerebene vorgenommen werden. In der vorliegenden Studie beschäftigen wir uns mit den Einschätzungen von Schülern auf der Individualebene. Hier verbleiben Forschungslücken zu der Dimensionalität von Instrumenten sowie der Übertragbarkeit über Fächer und das Alter hinweg. Diese Studie greift dabei auf Daten aus dem Projekt *„Ganz In"* zu sieben strukturell äquivalenten Dimensionen der Unterrichtsqualität in den Fächern Biologie, Deutsch, Englisch und Mathematik zu drei Messzeitpunkten zwischen der fünften und neunten Klasse zurück. Auf Grundlage der Daten zu Kriterien guten Unterrichts wurde ein mehrdimensionales Modell mit sieben latenten Dimensionen erarbeitet und auf Messinvarianz geprüft. Es konnten innerhalb der Fächer zwischen den drei Messzeitpunkten sowie zwischen den Fächern innerhalb der Messzeitpunkte ausreichend hohe Grade an Messinvarianz festgestellt werden, sodass aus diesem Modell abgeleitete latente Mittelwerte vergleichbar sind.

Schlüsselwörter: Generalisierbarkeit; Schülerebene; Messinvarianz; Modellvergleiche; Unterrichtsqualität

1 Einleitung

Chancenungleichheiten im Bildungssystem und eher unterdurchschnittliche Leistungen der Schülerinnen und Schüler in Deutschland stellen die gravierendsten Befunde in der ersten *Trends in International Mathematics and Science Study* (TIMSS), der *Internationalen Grundschul-Leseuntersuchung* (IGLU) und des ersten *Programme for International Student Assessment* (PISA) Zyklus in Deutschland dar (Klieme et al., 2010). In vertiefen-

den Studien konnte aufgezeigt werden, dass mangelnde Lernerfolge auch auf mangelhafte Unterrichtsprozesse zurückgeführt werden können (Köller, 2008). Unter anderem diese Befunde führten dazu, dass das Thema Unterrichtsqualität seitdem in Kontexten der Bildungspolitik und -forschung in vielfältiger Weise diskutiert wird. Bereits 2002 wurden seitens der Ständigen Konferenz der Kultusminister der Länder (KMK) verschiedene Handlungsansätze formuliert, um die in den internationalen Vergleichsstudien identifizierten Schieflagen des Schul- und Unterrichtswesens zu bearbeiten (KMK, 2002). Die „Weiterentwicklung und Sicherung der Qualität von Unterricht und Schule" (KMK, 2002, S. 7) wurde neben weiteren Handlungsansätzen zur zentralen Aufgabe aller Akteure im deutschen Schul- und Bildungssystem.

Um dieser Aufgabe entsprechen zu können, gilt es, neben der Konzeption von unterschiedlichen Unterrichtsentwürfen und theoretischen Vorschlägen zur Neugestaltung des Unterrichts, eine Bewertung des Unterrichts zur Identifizierung des Ist-Standes vorzunehmen (Praetorius, 2014; Altrichter et al., 2004). Erst daran anschließend können Veränderungsvorschläge formuliert, Handlungsansätze konzeptioniert und Maßnahmen implementiert werden (ebd.). Eine valide und reliable Erfassung der Unterrichtsqualität muss demzufolge angestrebt werden, stellt aber zugleich Forschung und Schulpraxis vor besondere Herausforderungen. Neben der Frage nach geeigneten Zugängen und Perspektiven zur Erfassung der Unterrichtsqualität stellt sich auch die Frage nach geeigneten Kriterien anhand derer die Qualität von Unterricht bestimmt werden soll (Clausen, 2002). Auch ist aus Sicht der empirischen Bildungsforschung noch nicht umfassend geklärt, ob sich Kriterien guten Unterrichts über Fächer und Zeitverläufe vergleichen lassen und ob diese eine entsprechende Konstruktvalidität aufweisen.

Mit der Frage nach der Vergleichbarkeit von theoretischen Konstrukten hat sich insbesondere die interkulturelle Forschung beschäftigt (z. B. Davidov et al., 2014; Helfrich, 1993). Im Rahmen der Forschung zur Unterrichtsqualität beschäftigten sich Scherer et al. (2016) bereits mit der Invarianz von Dimensionen der Unterrichtsqualität auf Basis der Daten aus PISA 2012 zwischen Australien, Kanada und den USA. Wagner et al. (2013) beschäftigten sich in diesem Kontext mit der Invarianz eines Modells der Unterrichtsqualität zwischen den Fächern Deutsch und Englisch. Der Grad der Vergleichbarkeit von Messwerten lässt sich bei gegebener theoretischer Konstruktäquivalenz und formaler operationaler Äquivalenz (Helfrich, 1993) über die Überprüfungen der Messinvarianz (MI; Mellenbergh, 1989) bestätigen. Da theoretische Konstrukte nicht direkt beobachtbar sind, wer-

den die Parameter der verwendeten statistischen Modelle auf Gruppen(un-)gleichheiten geprüft. Traditionell wird MI im Rahmen von konfirmatorischen Faktoranalysen im Mehrgruppenansatz geprüft (Jöreskog, 1971). Es erfolgten Erweiterungen zur partiellen Invarianzprüfung (Byrne et al., 1989) und zu der Schätzung latenter Mittelwertdifferenzen (Sörbom, 1974). Die formale Behandlung von verschiedenen Stufen der MI wurde von Meredith (1993) eingeführt.

Die Folgen mangelnder Messinvarianz und die Konsequenzen daraus wurden an verschiedenen Stellen thematisiert (z. B. van de Schoot et al., 2015; Guenole & Brown, 2014; Chen, 2007). Zusammengefasst liegt in diesem Fall keine methodische Grundlage für den Vergleich von latenten Strukturen und Mittelwerten vor. Es bleibt dann unklar, ob ähnliche Ausprägungen einer latenten Variable zwischen zwei Gruppen inhaltlich gleichbedeutend sind oder sich auf die Verschiedenheit von weiteren Modellparametern zurückführen lassen.

2 Kriterien guten Unterrichts

Die Allgemeine Pädagogik und die Schulpädagogik beschäftigen sich seit langem mit den Fragen danach, was konkret eine gute Lehrperson und was guten Unterricht auszeichnet. Diese Fragen sind ebenfalls zum zentralen Gegenstand der empirischen Unterrichtsforschung und der Lehr-Lern-Forschung geworden (Helmke, 2012). Um diese grundlegenden Fragestellungen hinsichtlich der Beurteilung der Qualität des Unterrichts bearbeiten zu können, sollen zwei sich ergänzende Perspektiven Aufschluss geben (Helmke, 2012; Ditton, 2009). Zum einen gibt es die produktorientierte Sichtweise, die die Frage nach Effektivitätskriterien guten Unterrichts stellt und von gutem Unterricht spricht, wenn die anvisierten Wirkungen erzielt worden sind (ebd.). Die andere Perspektive bezieht sich hingegen auf den Unterrichtsprozess, der dann eine gute Qualität aufweist, wenn er gültigen Qualitätsprinzipien und demnach auch unterrichtsmethodischen Forderungen entspricht (Helmke, 2012). Um aber letztlich von gutem und wirkungsvollem Unterricht sprechen zu können, ist eine Verschränkung dieser Perspektiven unumgänglich (Helmke, 2012; Einsiedler, 2002). So ist zusammengefasst nach Einsiedler (2002, S. 195) Unterrichtsqualität ein „[…] Bündel von Unterrichtsmerkmalen, die sich als ‚Bedingungsseite' (oder Prozessqualität) auf Unterrichts- und Erziehungsziele (Kriterienseite oder Produktqualität) positiv auswirken, wobei die Kriterienseite überwiegend von normativen Festlegungen bestimmt ist und der Zusammenhang von

Unterrichtsmerkmalen und Zielerreichung von empirischen Aussagen geleitet ist".

Anhand dieser Definition wird deutlich, dass sich guter Unterricht nicht durch die Erreichung eines einzelnen Merkmals auszeichnet, sondern durch vielerlei Merkmale und deren sinnvollem Zusammenwirken, welche letztlich in Beziehung zu den Zielen des Unterrichts zu setzen sind. Folglich ist das „Gesamtmuster des Unterrichts" (Helmke, 2012, S. 27) entscheidend, sodass die Unterrichtsqualität nicht nur durch die Wirksamkeit des Unterrichts definiert wird, sondern auch auf Grundlage von Prozessen und Merkmalen, die sowohl die Oberflächenstruktur als auch die Tiefenstruktur des Unterrichts umfassen (Reusser & Pauli, 2010).

Welche Merkmale umfassend und geeignet genug erscheinen, um die Frage nach gutem Unterricht beantworten zu können, ist eine Fragestellung, die im Laufe der letzten Jahrzehnte eine Vielzahl an Kriterienkatalogen und Merkmalslisten zur Folge hatte, die im Vergleich miteinander hohe Überschneidungen aufweisen (Helmke, 2012; Meyer, 2004). Als einflussreichste Klassifikationen lassen sich die Ausführungen von Slavin (1997), Brophy (2000) und Meyer (2004) benennen, die durch die Zusammenstellung von Helmke (2003, 2012) komplettiert werden. Die Kriterienkataloge variieren stark im Hinblick auf die Anzahl (minimal vier und maximal zwölf Kriterien) und in Bezug auf die Benennung der Kriterien, sodass sowohl weiter als auch enger gefasste Merkmale benannt werden (Willems, 2016) und sich somit Deutungsspielräume eröffnen. Merkmale wie der ‚Umgang mit Heterogenität' oder ein ‚lernförderliches Klima' (Helmke, 2012) sind nur zwei Beispiele, die vielfältige Gestaltungs- und Deutungsmöglichkeiten zulassen, während Merkmale wie ‚Klarheit und Strukturiertheit' oder ‚Klassenführung' weniger Verständnisspielraum eröffnen.

Trotz der Variabilität hinsichtlich Anzahl und Benennung der Kriterien weisen die Kriterienkataloge der Autoren inhaltliche Überschneidungen auf, die sich so zusammenfassen lassen, dass sie wichtige fachübergreifende Komponenten der Unterrichtsqualität herausstellen (Meyer, 2004; Helmke, 2012;) und innerhalb der Merkmale keine Gewichtung vorgeben. Jede Zusammenstellung ist dabei als individuelle Herangehensweise anzusehen, sodass keine allgemeingültigen Richtlinien hinsichtlich Anzahl und Benennung formuliert werden können und nur die Inhalte der Kriterienkataloge Aufschluss darüber geben, wann guter Unterricht als solcher definiert werden kann (ebd.). Unter Einbezug aller benannten Merkmalskataloge, realisiert sich guter Unterricht, wenn er durch folgende Merkmale geprägt und im Sinne dieser realisiert wird: eine klare Strukturierung des Unterrichts und eine entsprechende Klassenführung (ebd.), Aktivierung, Motivierung

und individuelle Förderung der Schüler (ebd.; Brophy, 2000; Slavin, 1997); die Schaffung eines lernförderlichen Klimas und Methodenvielfalt zum geeigneten Umgang mit heterogenen Lerngruppen (ebd.).

Im Rahmen unterschiedlicher Studien sind, vor dem Hintergrund der Fülle an Qualitätsmerkmalen, Modelle entstanden, die auf eine Reduzierung auf wesentliche Kernaspekte guten Unterrichts abzielen und versuchen, die jeweils wichtigsten Aspekte zu vereinen. Die bekannteste Einteilung von Merkmalen guten Unterrichts in grundlegende Basisdimensionen haben Klieme et al. (2001) im Kontext der TIMS-Video-Studie vorgenommen. Sie unterteilen die Merkmale in die drei Bereiche ‚kognitive Aktivierung', ‚Klassenführung' und ‚unterstützendes Klima' (Klieme et al., 2009; Creemers & Kyriakides, 2008; Klieme & Rakoczy, 2008). Während mit dem Bereich der ‚kognitiven Aktivierung' Inhalte und Varianten zur Förderung und Unterstützung und letztlich auch Möglichkeiten der Differenzierung anklingen, beziehen sich die Bereiche ‚Klassenführung' und ‚unterstützendes Klima' eher auf das Setting des Unterrichts und die Gestaltung der Lehr-Lernatmosphäre.

Durch eine solche Einteilung von Kriterien guten Unterrichts in übergreifende Basisdimensionen lässt sich die Komplexität der Merkmale guten Unterrichts reduzieren und trägt durch die pragmatische Zusammenfassung der Merkmale dazu bei, „[…] eine empirisch valide Betrachtung und theoretisch fundierte Evaluation der Unterrichtsqualität [zu] ermöglichen" (Schwanenberg et al., 2015, S. 506). Für alle diese übergeordneten Faktoren konnte in verschiedenen Studien positive Wirksamkeit nachgewiesen werden. Martin et al. (2013) und Fauth et al. (2014) zeigten positive Effekte für die Leistungsebene auf, Fauth et al. (2014) verwiesen zudem auf Zusammenhänge zwischen fachspezifischem Interesse und Dimensionen der Unterrichtsqualität. Klieme et al. (2009) entdeckten wiederum positive Zusammenhänge zur Lernmotivation und affektiven Einstellungen der Schülerinnen und Schüler.

2.1 Erfassung guten Unterrichts – Perspektiven und Operationalisierungen

Um die Unterrichtsqualität zu erfassen, bieten sich unterschiedliche Perspektivenzugänge an (Clausen, 2002). So ist eine Bewertung des Unterrichts aus Perspektive der Lehrkräfte, der Schülerinnen und Schüler oder externer Beobachter möglich (ebd.). Für die jeweiligen Zugänge lassen sich jeweils Vor- und Nachteile benennen, die es hinsichtlich des Forschungsinteresses abzuwägen gilt (ebd.).

Beurteilungen der Unterrichtsqualität unter Hinzunahme externer Beobachter, z. B. durch Videostudien (z. B. Seidel et al., 2005), sind zumeist aufwändig und kostspielig (Clausen, 2002), bieten aber, wenn sie umsetzbar sind, viele differenzierte Möglichkeiten (Seidel et al., 2005). Zudem wird Urteilen externer Beobachter eine hohe Validität attestiert (Waldis et al., 2010) und durch mögliche verschiedene Unterrichtsstunden, die von den externen Beobachtern beurteilt werden, ergeben sich vielerlei Vergleichsmöglichkeiten (Rakoczy, 2008). Jedoch sind Aufwand und Kosten ausschlaggebende Faktoren dafür, dass oftmals auf andere Beurteilerquellen zurückgegriffen wird (Clausen, 2002).

Einen weiteren Zugang zur Erfassung der Unterrichtsqualität eröffnet die Lehrerperspektive. Von dieser Perspektive ausgehend, kann, auf Grundlage von erstelltem Videomaterial der eigenen Stunde oder über Einschätzungsbögen nach Unterrichtsende, eigenes Verhalten reflektiert, das der Schüler bewertet und schließlich beides in einen Gesamtzusammenhang gebracht werden (ebd.). Als Nachteil von Beurteilungen des Unterrichts aus Sicht der Lehrkräfte verweist Clausen (2002) jedoch darauf, dass Lehrkräfte ihren eigenen Unterricht tendenziell positiver bewerten und es häufig zu Wahrnehmungsverzerrungen kommt (Terhart, 2006).

Einen letzten Zugang bietet die Schülerperspektive, die bisher der am häufigsten verwendete Zugang zur Erfassung der Unterrichtsqualität ist (Fauth et al., 2014; Kunter et al., 2008; Clausen, 2002). Schülerurteile zur Unterrichtsqualität bieten die Vorteile, dass durch die hohe Heterogenität innerhalb der Klassen ein mehrperspektivisches Abbild des Unterrichts möglich ist (ebd.) und die Urteile eine hohe Validität in Bezug auf die Erfassung der Unterrichtsqualität und die sozialen Interaktionen im Klassenraum aufweisen (Kunter et al., 2005). Darüber hinaus werden sie in der Forschung als sehr stabil bewertet (Ditton & Arnoldt, 2004; Ditton et al., 2002). Einschränkung erfährt dieser Zugang zur Erfassung der Unterrichtsqualität dahingehend, dass Schüler nur über ein geringes methodisch-didaktisches Wissen verfügen und sich diesbezügliche Einschätzungen demzufolge verzerren können (Kunter & Baumert, 2006). Trotz der unterschiedlichen Zugänge und der damit verbundenen Vor- und Nachteile hinsichtlich der Qualität der Urteile liegt nach Clausen (2002, S. 186) „keine der drei Sichtweisen generell näher an einer „Unterrichtswirklichkeit" als die anderen Perspektiven", sodass letztlich jede Perspektive zur Erfassung der Unterrichtsqualität angemessen ist. Es kommt vielmehr auf das Erkenntnisinteresse und auf die Abwägung der Vor- und Nachteile für das entsprechende Forschungsinteresse an.

Unabhängig von der Perspektive, ist das Verfahren, welches am häufigsten eingesetzt wird, um die Unterrichtsqualität abzubilden, die standardisierte Fragebogenerhebung, bei der die Kriterien guten Unterrichts die Grundlage für die eingesetzten Skalen stellen. Die Vorteile einer Fragebogenerhebung liegen darin, dass diese Variante nicht nur kostengünstig ist, sondern auch längere Zeiträume als Grundlage der Bewertung genommen werden können, sodass es möglich ist, trotz Kostenneutralität einen hohen Erkenntnisgewinn zu erzielen (Wagner et al., 2013; Clausen, 2002). Bei der Erfassung guten Unterrichts mittels standardisierter Fragebögen „ist es von hoher Bedeutung, dass die gewonnenen empirischen Daten eine hinreichende Reliabilität und Validität aufweisen: Für die Unterrichtsforschung sind zuverlässige und valide Messungen wichtig, da nur dann Ergebnisse zur Qualität von Unterricht sinnvoll interpretiert werden können" (Praetorius, 2014, S. 13).

Die Unterrichtsqualität kann sowohl auf der Individualebene als auch auf der Klassen- oder Schulebene untersucht werden. Lüdtke und Kollegen (2009, S. 121 f.) stellten heraus, dass die Wahl der Untersuchungsebene direkt mit der Forschungsfrage verknüpft ist. Insofern individuelle Wahrnehmungen oder interindividuelle Unterschiede im Forschungsinteresse stehen ist die Schülerebene als zentral zu erachten. Wenn Abhängigkeiten zu kontextuellen Umgebungsmerkmalen untersucht werden sollen ist die Klassen- oder Schulebene heranzuziehen und die Angaben der Schülerinnen und Schüler würden nur als Aggregat verwendet.

Für die hier eingesetzten Skalen zu Dimensionen der Unterrichtsqualität konnte bereits in vorangegangenen Analysen zur Unterrichtsqualität im selben Projektkontext festgestellt werden, dass die Skalen reliabel sind und inhaltlich verschiedenartig kategorisiert werden können. Bei Willems und Glesemann (2015) und Schwanenberg et al. (2015) wurden dieselben Daten und Skalen verwendet, aber inhaltlich verschiedenartig kategorisiert, wobei die theoretischen Grundlegungen und internen Konsistenzen beider Operationalisierungen als angemessen erschienen. Zudem wurde festgestellt, dass die untersuchten Dimensionen der Unterrichtsqualität hoch miteinander korrelieren (Schwanenberg et al., 2015). Dies ist, vor dem Hintergrund der Aussage von Helmke (2012), dass Anzahl und Bezeichnungen der Kategorien bis zu einem gewissen Grade willkürlich sind, zu erwarten gewesen. Auf der Basis dieser beiden Ansätze war es allerdings nicht möglich, die rechnerische Angemessenheit der Operationalisierungen gegeneinander zu testen und die angemessenere Lösung zu identifizieren. Bei einem modellbasierten Vorgehen im Rahmen des Strukturgleichungsansatzes besteht die Möglichkeit, die Passung der Modelle auf die vorliegenden Daten zu be-

werten und zu vergleichen, Relationen zwischen theoretischen Konstrukten abzubilden und die Messfehler explizit zu berücksichtigen (z. B. Reinecke & Pöge, 2010; Bühner, 2010; Bollen, 1989). Zudem wird es möglich, die Angemessenheit der Operationalisierung zwischen verschiedenen Beobachtungen zu quantifizieren. Wenn dies nicht geprüft wird, bleibt unklar, ob die Kriterien guten Unterrichts über verschiedene Fächer und die Zeit hinweg generalisierbar sind, wie von Helmke (2012) angenommen. Untersuchungen von Wagner und Kollegen (2015, 2013) zeigten, dass eine mehrdimensionale Struktur von Schülerurteilen zur Unterrichtsqualität zwischen verschiedenen Schulklassen und den Fächern Englisch und Deutsch angemessen und generalisierbar ist und dass einzelne Skalen im Fach Mathematik zeitstabile und zeitvariable Anteile aufweisen. Offen blieb jedoch, ob beobachtete mehrdimensionale Konstrukte im Zeitverlauf über mehrere Jahre stabil genug bleiben, um Vergleiche zwischen ihnen vornehmen zu können und ob ein weiter gefasster Fächervergleich möglich ist (ebd.). Ein sinnvoller und bedeutsamer Vergleich von Beobachtungen hypothetischer Konstrukte über mehrere Gruppen setzt voraus, dass jeweils dieselben Eigenschaften oder Fähigkeiten gemessen werden, also faktorielle Validität und Testfairness vorliegen (z. B. Bühner, 2010, S. 64).

2.2 Forschungsfragen

Da sich diese Studie maßgeblich mit Fragen der Konstruktvalidität von Instrumenten beschäftigt, die individuelle Urteile der Schülerinnen und Schüler abbilden und zudem keine Kontextvariablen verwendet und verglichen werden sollen, liegt das zentrale Forschungsinteresse hier vorläufig auf der Individualebene.

Aufgrund der Beobachtung von hohen Korrelationen zwischen den zugrunde liegenden Skalen bei Schwanenberg et al. (2015) erscheinen mehrdimensionale Modellierungen, welche die Zusammenhänge zur Schätzung von Modellparametern explizit miteinbeziehen, vorteilhafter als die bisherigen Operationalisierungen. Unklar bleibt, ob die angelegte dimensionale Struktur auf der Individualebene passend ist, ob diese Passung über Klassenstufen und verschiedene Fächer aufrechterhalten werden kann und ob die Konstrukte inhaltlich gleichbedeutend sind.

Damit ergibt sich für den vorliegenden Beitrag das folgende Forschungsinteresse:

I. Kann eine einzelne mehrdimensionale Struktur der Unterrichtsqualität zu jedem Messzeitpunkt und in jedem Fach angenommen werden?
II. Weist die angenommene mehrdimensionale Struktur eine ausreichende Ähnlichkeit zwischen den Fächern innerhalb der Messzeitpunkte auf, um Vergleiche zwischen den Strukturen und den Ausprägungen der Dimensionen vornehmen zu können?
III. Weist die angenommene mehrdimensionale Struktur eine ausreichende Ähnlichkeit zwischen den Messzeitpunkten innerhalb der Fächer auf, um Vergleiche zwischen den Strukturen und den Ausprägungen der Dimensionen vornehmen zu können?

Dazu werden die folgenden Hypothesen formuliert. (I.) Es wird, basierend auf den Aussagen von Helmke (2012) und den vorangegangenen empirischen Arbeiten von Willems und Glesemann (2015) und Schwanenberg et al. (2015), angenommen, dass sich eine mehrdimensionale Struktur auf Basis der vorliegenden Daten abbilden lässt, die über alle Fächer und Messzeitpunkte angenommen werden kann. (II.+III.) Es wird erwartet, dass sich die Modellparameter punktuell signifikant unterscheiden. Also wird nicht erwartet, dass Invarianz von Faktorladungen und Itemmittelwerten in allen Vergleichen nachgewiesen werden kann, da Wagner und Kollegen (2013) diese in längsschnittlichen Vergleichen von einzelnen Dimensionen zurückweisen mussten.

3 Methoden

Auf Grundlage der Zusammenstellungen von Kriterien guten Unterrichts werden, auf der Basis der Daten des Schulentwicklungsprojekts *Ganz In*, Modelle generiert und hinsichtlich ihrer Passung bewertet. Die Modelle werden über die Zeit und zwischen den Fächern schrittweisen Prüfungen auf MI zwischen den Fächern und den Messzeitpunkten unterworfen, um die Generalisierbarkeit des Modells festzustellen.

3.1 Daten

Datengrundlage bilden die Schülerdaten zur Unterrichtsqualität an Ganztagsgymnasien der ersten Projektphase des Projekts *Ganz In – Mit Ganztag*

mehr Zukunft. Das neue Ganztagsgymnasium NRW. Im Fokus des Projekts steht die Schul- und Unterrichtsentwicklung an 31 Gymnasien in NRW, die mit Evaluationen, Netzwerkbetreuung, sowie fachdidaktischen und themenspezifischen Angeboten bei der Einführung des gebundenen Ganztags unterstützt wurden (Wendt & Bos, 2015; Berkemeyer et al., 2010).

Es wurden zwei zeitlich versetze Schülerkohorten begleitet (Schwanenberg et al., 2015). In dieser Studie werden nur die Daten der ersten Kohorte verwendet, da die zweite Kohorte nur über zwei Messzeitpunkte von der Klassenstufe fünf bis zur Klassenstufe sieben begleitet worden ist. Die erste Kohorte wurde von der Klassenstufe fünf bis in die Klassenstufe neun hinein begleitet. Die Befragung zur Unterrichtsqualität fand dabei jeweils am Ende der Schuljahre 2010/11, 2012/13 und 2014/15 statt. Im Schuljahr 2010/11 wurde die Befragung mittels Fragebögen durchgeführt. In den Schuljahren 2012/13 und 2014/15 wurde auf eine Onlineumfrage mit der Software *EFS Survey* umgestellt (ebd.). 2011 nahmen 2812 Schülerinnen und Schüler teil, 2013 waren es 2845 und 2015 beteiligten sich 2367 Schülerinnen und Schüler. Die Ausschöpfung lag generell bei über 85 %. Insgesamt sind 1501 Schülerinnen und Schüler über alle drei Messzeitpunkte verfolgbar.

Die Unterrichtsqualität wurde zur Evaluation fachspezifischer Interventionen und der Evaluation der Arbeit an den Schulen über verschiedene Fächer hinweg untersucht. Die Befragung bezog sich auf den Unterricht der jeweils aktuellen Schuljahre in den Fächern Deutsch, Mathematik, Englisch und Biologie. In wenigen Schulen wurde das Fach Biologie in einzelnen Klassenstufen nicht unterrichtet, daher ist die Stichprobe für das Fach systematisch geringer. In der Jahrgangsstufe sieben wurde die Befragung zur Unterrichtsqualität um die Fächer Chemie und Physik erweitert. Da auch hier keine Beobachtungen über drei Messzeitpunkte vorlagen, werden diese Fächer hier nicht mitbetrachtet.

3.2 Instrumente

Auf Basis der wesentlichen Merkmale von Unterrichtsqualität nach Helmke (2012) wurden insgesamt elf Skalen für die Befragung ausgewählt. Sieben der elf Skalen *(Klassenmanagement, Diagnostische Kompetenz/Angemessenheit, Schüler-Lehrer-Verhältnis, Unterstützung und Anregung, Empathie, Interessantheit* sowie *Adäquatheit)* entstammen der Studie *Qualitätssicherung in Schule und Unterricht* (QuaSSU) (Ditton, 2001, 2002), zwei Skalen *(Strukturiertheit* und *Leistungsbezogene Differenzierung)* der *Studie zur Entwicklung von Ganztagsschulen* (SteG) (Quellenberg 2009) und zwei weitere Skalen *(Lernmotivation und Fachbezogene Amotivation)* aus IGLU (Bos

et al., 2005). Diese Dimensionen bilden ihrerseits eine mögliche Grundlage für die drei Basisdimensionen guten Unterrichts nach Klieme et al. (2001). Die verwendeten Items wurden jeweils für alle Fächer semantisch angepasst, analog wurde also beispielsweise von dem/der Biologielehrer/-in oder dem/der Deutschlehrer/-in gesprochen. Es wurde von einer funktionalen Äquivalenz der Konstrukte ausgegangen (Helfrich, 1993). Da bei den Analysen Unterrichtsmerkmale, nicht aber individuelle Lernbedingungen im Fokus stehen, wurden die Skalen *Lernmotivation* und *Fachbezogene Amotivation* im Rahmen dieser Studie nicht berücksichtigt.

Im ersten Schritt wurden alle Dimensionen der Unterrichtsqualität mittels konfirmatorischer Faktoranalysen einzeln innerhalb jedes Messzeitpunktes und jedes Faches analysiert. Items mit geringer Trennschärfe wurden für die weiteren Analysen ausgeschlossen.g[1] Die Skalen *Adäquatheit* und *Interessantheit* wurden an dieser Stelle wegen unzureichender Güte ausgeschlossen. Der untenstehenden Tabelle 1 sind die deskriptiven Werte der Skalen zu entnehmen. Die Metrik der Einzelitems reicht jeweils von eins bis vier. Die Skala *Leistungsbezogene Differenzierung* weist zum ersten Messzeitpunkt unzureichende Reliabilitätswerte über alle Fächer hinweg auf. Diese unzureichenden Werte können jedoch für die weiteren Analysen hingenommen werden, da die Messmodelle eine annehmbare Güte aufweisen.

Tabelle 1: Deskriptive Beschreibung der verwendeten Skalen

Skala	Itemanzahl	Beispielitem	Fach	K5			K7			K9		
				M	*SD*	*α*	M	*SD*	*α*	*M*	*SD*	*α*
Klassenmanagement	4	„Wir fangen erst lange nach dem Beginn der Stunde an zu arbeiten." (recodiert)	B	2,88	0,76	,85	2,79	0,87	,89	2,97	0,83	,89
			D	2,98	0,69	,89	2,96	0,85	,89	2,97	0,74	,86
			E	2,97	0,69	,83	2,96	0,84	,90	2,97	0,84	,89
			M	3,02	0,67	,81	2,80	0,96	,92	2,94	0,86	,89
Diagnostische Kompetenz/ Angemessenheit	4	„Die Aufgaben, die uns unsere [Fach]lehrerin/unser [Fach]lehrer stellt, sind ganz schön schwierig." (recodiert)	B	3,33	0,57	,77	3,18	0,71	,83	3,00	0,75	,84
			D	3,38	0,50	,72	3,11	0,74	,85	3,25	0,59	,77
			E	3,32	0,56	,77	3,08	0,76	,86	3,14	0,72	,83
			M	3,22	0,54	,74	2,81	0,82	,86	2,8	0,74	,84

1 Die exakten Werte, sowie Zwischenergebnisse und der verwendete Code können bei den Autoren erfragt werden.

Skala	Item-anzahl	Beispielitem	Fach	K5			K7			K9		
				M	*SD*	α	M	*SD*	α	*M*	*SD*	α
Schüler-Lehrer-Verhältnis	5	„Wir haben großes Vertrauen zu unserer [Fach]lehrerin/unserem [Fach]lehrer."	B	2,90	0,82	,87	2,52	0,87	,89	2,67	0,83	,89
			D	3,14	0,73	,86	2,8	0,89	,90	2,79	0,85	,89
			E	3,01	0,79	,88	2,71	0,87	,90	2,64	0,87	,90
			M	3,03	0,76	,87	2,50	0,95	,92	2,56	0,87	,89
Strukturiertheit	3	„Am Ende der Stunde fasst unsere [Fach]lehrerin/unser [Fach]lehrer das Wichtigste zusammen."	B	2,44	0,85	,77	2,28	0,88	,83	2,48	0,88	,85
			D	2,49	0,81	,76	2,36	0,91	,85	2,34	0,80	,76
			E	2,37	0,82	,77	2,3	0,89	,85	2,26	0,86	,84
			M	2,46	0,82	,76	2,21	0,95	,88	2,31	0,88	,83
Anregung	4	„Unsere [Fach]lehrerin/unser [Fach]lehrer stellt uns interessante Aufgaben."	B	3,04	0,72	,78	2,64	0,85	,85	2,70	0,81	,85
			D	3,15	0,65	,76	2,79	0,83	,86	2,83	0,76	,79
			E	3,12	0,68	,78	2,73	0,84	,86	2,68	0,83	,85
			M	3,16	0,63	,73	2,55	0,90	,87	2,68	0,82	,83
Empathie	5	„Unsere [Fach]lehrerin/unser [Fach]lehrer merkt, wenn der Unterricht zu schwer ist."	B	2,84	0,74	,82	2,47	0,79	,87	2,62	0,73	,85
			D	3,12	0,65	,80	2,79	0,81	,82	2,67	0,73	,84
			E	3,08	0,68	,81	2,76	0,81	,88	2,63	0,80	,88
			M	3,16	0,64	,79	2,57	0,90	,90	2,67	0,84	,88
Leistungsbezogene Differenzierung	4	„Unsere [Fach]lehrerin/unser [Fach]lehrer gibt je nach Leistung unterschiedlich schwere Aufgaben."	B	1,99	0,59	,54	1,99	0,72	,74	2,09	0,78	,82
			D	2,16	0,55	,48	2,18	0,76	,77	1,96	0,60	,64
			E	2,20	0,57	,48	2,22	0,79	,79	2,09	0,74	,78
			M	2,32	0,56	,47	2,24	0,84	,81	2,48	0,75	,73

Anmerkungen: M: Mittelwert; SD: Standardabweichung; α:Cronbachs Alpha; B: Biologie; D: Deutsch; E: Englisch; M: Mathematik; K5: Klassenstufe fünf; K7: Klassenstufe sieben; K9: Klassenstufe neun

3.3 Datenanalyse

Alle statistischen Analysen wurden mit der Software MPlus 5.1 durchgeführt (Muthén & Muthén, 1998–2007). Es wurde ein robuster Maximum Likelihood Schätzer verwendet. Fehlenden Werte wurde mit dem Full-Information-Maximum-Likelihood (FIML) Verfahren begegnet. Da sich die Gruppenvergleiche in dieser Studie immer auf Vergleiche innerhalb derselben Stichprobe beziehen, wurden die Daten für diese Analysen so umstrukturiert, dass ein langes Datenformat mit den Faktoren „Klassenstufe" und „Fach" vorlag. Die Chi^2-Statistiken wurden mit Satorra-Bentler-Kor-

rektur für erhöhte Schiefe und Kurtosis (Reinecke & Pöge, 2010) ermittelt. Die erste Variable jedes Messmodells wurde als Markervariable verwendet, also wurde das jeweilige Ladungsgewicht auf 1 fixiert. Als Cut-Off-Werte für die Modellgütekriterien wurden die von Hu und Bentler (1999) vorgeschlagenen Schwellen gewählt. Die hierarchische Struktur der Daten auf Klassenebene wurde mittels des Analysetyps „complex" unter MPlus berücksichtigt. Hierbei werden die Standardfehler und Chi^2-Tests der Modellgüte für die geschachtelte Stichprobe adjustiert. Da das zentrale Forschungsinteresse auf der Individualebene und weniger auf den Charakteristika der Lernumgebung oder den Effekten von Unterrichtsqualität liegt, erscheint dieses Vorgehen als angemessen (Lüdtke et al., 2009). Um einen Vergleich mit anderen Studien vornehmen und die Plausibilität der Ergebnisse einschätzen zu können, wurden die mittleren Intra-Klassenkorrelationskoeffizienten der Skalen bestimmt. Bei allen Modellen wurde volle Kovariation zwischen den sieben herangeführten latenten Dimensionen der Unterrichtsqualität zugelassen. Die latenten Korrelationen werden vollständig dargestellt.

Um der Frage nachzugehen, ob die dimensionale Struktur (I.) zwischen den latenten Konstrukten für die Daten angemessen ist, wurden äquivalente Modelle innerhalb der Fächer und Messzeitpunkte auf der Individualebene gefittet. Ergänzend wurden Modellvergleiche zu einem g-Faktormodell und einem Modell mit drei latenten Konstrukten zweiter Ordnung, welche die Basisdimensionen nach Klieme et al. (2001) abbilden, vorgenommen.

Nachdem ein mathematisch angemessenes Modell, welches zu allen Messzeitpunkten und innerhalb aller Fächer angenommen werden kann, abgeleitet wurde, wird über die Prüfung der MI festgestellt, ob die Messbeziehungen zwischen den Indikatoren und den latenten Variablen gleich sind.

So kann festgestellt werden, ob Verzerrungen in der Zusammensetzung der latenten Mittelwerte vorliegen. Bei Little (2013, S. 140), Widaman et al. (2010) und Widaman und Reise (1997) wurde in Anlehnung an Horn und McArdle (1992), Meredith (1993) und andere (für eine Übersicht siehe Vandenberg und Lance 2000) eine hierarchische Abfolge statistischer Prüfungen empfohlen, der an dieser Stelle gefolgt werden soll. Wenn eine signifikante Verschlechterung durch die Einführung von Restriktionen beobachtet werden kann, wird der geprüfte Grad der MI zurückgewiesen.

Im ersten Schritt wird dafür jeweils ein Mehrgruppenmodell als Basismodell spezifiziert und auf seine Passung geprüft. Dann werden steigend restriktive Modelle erstellt und jeweils gegen das Basismodell getestet. Für die Prüfung der konfiguralen Invarianz werden die Faktorladungen, bis auf

die Ladung der Markervariable, in beiden Gruppen frei geschätzt. Die Mittelwerte der latenten Variablen wurden dazu auf 0 fixiert. Bei gegebener konfiguraler Invarianz ist die Voraussetzung für die Prüfung weiterer Grade hergestellt. Für die Prüfung der (schwachen) metrischen Invarianz wurden die Faktorladungen zwischen den Modellen gleichgesetzt. Bei metrischer Invarianz weisen die gemessenen Konstrukte eine vergleichbare Ladungsstruktur auf, sind also inhaltlich ausreichend ähnlich, um Vergleiche zwischen ihnen vorzunehmen. Für die Prüfung der skalaren Invarianz wurden die Intercepts (Regressionskonstanten) und die Faktorladungen gleichgesetzt. Auf der Basis der vorliegenden Invarianz wird es möglich, Mittelwertunterschiede in den latenten Variablen sinnvoll zu interpretieren. Für die vorliegende Studie kann dies also als wichtigste Schwelle begriffen werden. Zuletzt wurden die Faktormittelwerte der ersten Gruppe auf 0 fixiert und die Differenzen zu der anderen Gruppe frei geschätzt. Auf diese Weise können Unterschiede zwischen den latenten Mittelwerten in den Gruppen beobachtet und bewertet werden. Aus Identifikationsgründen werden dabei nicht die latenten Mittelwerte, sondern nur die Mittelwertdifferenzen geschätzt (Reinecke, 2005). Auf der Basis der Prüfungen auf Messinvarianz werden ausgewählte Differenzen der latenten Mittelwerte dargestellt. Die Metrik der latenten Variablen entspricht dabei der Metrik der Markervariablen (1–4). Auf eine Prüfung strikter oder residualer Invarianz wurde an dieser Stelle verzichtet, da diese Prüfung nicht zwingend notwendig ist, um die gewünschten Vergleiche vorzunehmen. Falls keine Messinvarianz festgestellt werden konnte, wurde vertiefend auf partielle Messinvarianz geprüft. Bei Prüfungen auf partielle Invarianz werden einzelne Parameter von den generellen Restriktionen befreit, um punktuelle Abweichungen von den Modellannahmen prüfen und gegebenenfalls zulassen zu können.

Als Entscheidungskriterium für die Tests auf MI schlagen Cheung und Rensvold (2002) und Meade et al. (2008), begründet auf den Verzerrungen des Chi^2-Wertes, die Differenz der Comparative Fit Indices (ΔCFI) vor. Eine Differenz des CFI zwischen dem Basismodell und den restriktiveren Modellen von etwa $\Delta CFI = 0{,}02$ wird als Indikation einer bedeutsamen Abweichung eingeführt (Meade et al., 2008, S. 586). Dieses Vorgehen findet sich beispielhaft auch bei Analysen der MI von Instrumenten der Studie PIRLS bei Schulte, Nonte und Schwippert (2013). Andere Autoren haben eine strengere Schwellen von $\Delta CFI = 0{,}01$ für die Interpretation vorgeschlagen (Chen, 2007; Wu et al., 2007).

4 Ergebnisse

In der Tabelle 2 sind die Fit-Werte der generellen Modelle pro Messzeitpunkt und Fach abgetragen. Alle Modelle weisen ausreichende oder gute Anpassungswerte auf. Zwar werden die Chi^2-Werte signifikant, dies wird aber auf die Verzerrung aufgrund der großen Stichprobe zurückgeführt. Die Modelle in Biologie in der Klassenstufe neun und in Mathematik in den Klassenstufen sieben und neun weisen dabei insgesamt die unvorteilhafteste Modellanpassung auf. Die mittlere Intra-Klassenkorrelation auf der Klassenebene (*ICC*; 128 Klassen) lag bei $\overline{ICC}$ = ,17 (*SD* = ,07). Die auf die Schulebene (31 Schulen) zurückführbare mittlere Intra-Klassenkorrelation lag bei $\overline{ICC}$ = ,06 (*SD* = ,03).

Tabelle 2: Modellanpassungen der Modelle innerhalb der Fächer und Klassenstufen

Klassenstufe fünf							
	Chi²	*df*	*p*	*CFI*	*TLI*	*RMSEA*	*SRMR*
Biologie	1506.5	356	<,01	0,952	0,945	,036	,046
Deutsch	1369.8	356	<,01	0,954	0,948	,033	,041
Englisch	1400.1	356	<,01	0,957	0,951	,034	,039
Mathematik	1536.3	356	<,01	0,943	0,935	,036	,040
Klassenstufe sieben							
	Chi²	*df*	*p*	*CFI*	*TLI*	*RMSEA*	*SRMR*
Biologie	1711,7	356	<,01	0,952	0,946	,040	,043
Deutsch	1848,1	356	<,01	0,960	0,955	,039	,044
Englisch	1878,8	356	<,01	0,959	0,953	,039	,041
Mathematik	1938,0	356	<,01	0,961	0,956	,040	,039
Klassenstufe neun							
	Chi²	*df*	*p*	*CFI*	*TLI*	*RMSEA*	*SRMR*
Biologie	1506,5	356	<,01	0,929	0,929	,047	,047
Deutsch	1369,8	356	<,01	0,942	0,942	,040	,040
Englisch	1400,1	356	<,01	0,945	0,945	,042	,042
Mathematik	1536,3	356	<,01	0,939	0,939	,045	,040

Anmerkungen: *Chi²*: *Chi²*-Wert des Tests der Modellanpassung; *df*: Freiheitsgrade; *p*: Signifikanzniveau der *Chi²*-Tests; *CFI*: Comparative Fit Index; *TLI*: Tucker Lewis Index; *RMSEA*: Root Mean Square Error of Approximation; *SRMR*: Standardized Root Mean Square Residual

In Modellvergleichen gegenüber g-Faktor Modellen und Modellen mit einer latenten Struktur zweiter Ordnung auf Basis der Basisdimensionen nach Klieme et al. (2001) wurde überdies festgestellt, dass in jedem Fach und zu jedem Messzeitpunkt das Modell ohne latente Variablen zweiter Ordnung den Daten am besten entsprach oder es das einzige Modell war, das das Konvergenzkriterium erreichte.

Um die Zusammenhangsstruktur zwischen den latenten Dimensionen betrachten zu können sind in der Tabelle 3 die Korrelationen abgetragen.

Tabelle 3: Korrelationen zwischen den latenten Dimensionen in allen Modellen

		Klassenstufe 5				Klassenstufe 7				Klassenstufe 9			
zwischen	**und**	β_B	β_D	β_E	β_M	β_B	β_D	β_E	β_M	β_B	β_D	β_E	β_M
Klassen-man.	Diag. Kompetenz/Ang.	.33	.35	.42	.38	.29	.31	.39	.36	.46	.32	.27	.34
	Schüler-Lehrer-Verh.	.32	.37	.33	.29	.19	.25	.18	.13	n.s.	.34	.13	.33
	Strukturiertheit	.26	.23	.23	.13	.13	.18	.08	.08	n.s.	.32	.11	.27
	Anregung	.29	.33	.27	.25	.13	.23	.17	.14	.11	.34	.17	.32
	Empathie	.34	.38	.35	.30	.24	.30	.24	.20	.15	.40	.20	.36
	Leistungsb. Diff.	.11	.16	.13	.17	-.19	n.s.	-.15	-.11	-.29	.15	n.s.	.15
Diag. Kompetenz/ Ang.	Schüler-Lehrer-Verh.	.31	.35	.30	.39	n.s.	.13	n.s.	n.s.	n.s.	.30	.10	.29
	Strukturiertheit	.08	n.s.	.12	.09	-.22	-.13	-.22	-.18	-.17	n.s.	-.12	.11
	Anregung	.34	.34	.34	.36	n.s.	.15	.07	n.s.	n.s.	.35	.15	.31
	Empathie	.28	.33	.34	.43	n.s.	.14	.13	.13	n.s.	.31	.14	.35
	Leistungsb. Diff.	n.s.	n.s.	n.s.	n.s.	-.33	-.15	-.21	-.14	-.35	n.s.	-.25	.16
Schüler-Lehrer-Verh.	Strukturiertheit	.47	.45	.45	.46	.63	.54	.56	.60	.56	.57	.58	.66
	Anregung	.89	.86	.85	.88	.86	.88	.87	.90	.85	.87	.89	.91
	Empathie	.84	.84	.83	.84	.79	.82	.79	.81	.78	.82	.81	.86
	Leistungsb. Diff.	.48	.48	.53	.49	.49	.49	.51	.60	.40	.50	.54	.64
Strukturiert-heit	Anregung	.52	.51	.56	.54	.66	.59	.61	.67	.62	.61	.63	.68
	Empathie	.65	.57	.60	.54	.67	.59	.59	.63	.64	.70	.67	.70
	Leistungsb. Diff.	.62	.57	.64	.56	.59	.59	.61	.66	.53	.60	.62	.67
Anregung	Empathie	.86	.88	.91	.88	.78	.83	.83	.83	.79	.83	.85	.88
	Leistungsb. Diff.	.49	.51	.58	.55	.48	.49	.54	.66	.39	.46	.56	.67
Empathie	Leistungsb. Diff.	.65	.63	.65	.62	.53	.52	.55	.62	.46	.68	.61	.72

Anmerkungen: ß: Standardisierter Korrelationskoeffizient in den spezifischen Modellen 0; n.s.: Nicht Signifikant auf einem Niveau von 95 %; B: Biologie; D: Deutsch; E: Englisch; M: Mathematik

Es kann beobachtet werden, dass die Größe der Zusammenhänge zwischen den einzelnen latenten Dimensionen zwischen den Fächern und den Klassenstufen zumeist verhältnismäßig gering variieren. Die generell höchsten Zusammenhänge und geringsten Differenzen lassen sich zwischen den Dimensionen *Empathie*, *Lehrer-Schüler-Verhältnis* und *Anregung* beobachten, die geringsten Zusammenhänge und höchsten Differenzen liegen zwischen *Klassenmanagement*, *Leistungsbezogener Differenzierung* und *Diagnostischer Kompetenz* vor.

Dass die einzelnen Modelle angemessen angepasst sind, konnte bereits der Tabelle 2 entnommen werden. In der Tabelle 4 sind die Ergebnisse der Tests auf MI abgetragen. Differenzen, die die Schwelle von $\Delta CFI = 0{,}01$ erreichten oder überschritten, sind hervorgehoben worden.

Tabelle 4: Ergebnisse der Testungen auf Messinvarianz

Vergleiche			Konfigurale Invarianz				Metrische Invarianz				Skalare Invarianz			
In	Zwischen		Chi²	df	p	CFI	ΔChi²	Δdf	p	ΔCFI	ΔChi²	Δdf	p	ΔCFI
B	5 zu 7		3219,2	712	<,01	0,952	147,8	22	<,01	0,002	733.8	44	<,01	0,013
	7 zu 9		3687,5	712	<,01	0,945	78,4	22	<,01	0,001	186.9	44	<,01	0,002
D	5 zu 7		3218,3	712	<,01	0,958	141,6	22	<,01	0,002	991.1	44	<,01	0,016
	7 zu 9		3493,9	712	<,01	0,956	141,4	22	<,01	0,002	471.0	44	<,01	0,007
E	5 zu 7		3284,3	712	<,01	0,958	125,0	22	<,01	0,001	961.1	44	<,01	0,015
	7 zu 9		3627,8	712	<,01	0,956	65,9	22	<,01	0,001	175.2	44	<,01	0,002
M	5 zu 7		3481,8	712	<,01	0,955	160,8	22	<,01	0,002	1257.0	44	<,01	0,020
	7 zu 9		3948,7	712	<,01	0,955	105,1	22	<,01	0,001	416.6	44	<,01	0,005
5	B	zu D	2873,1	712	<,01	0,953	36,8	22	<,01	0,001	138.7	44	<,01	0,002
		zu E	2903,7	712	<,01	0,954	9,7	22	<,01	0,001	202.6	44	<,01	0,003
		zu M	3038,8	712	<,01	0,948	30,8	22	<,01	0,001	256.4	44	<,01	0,005
7	B	zu D	3558,5	712	<,01	0,957	57,8	22	<,01	0,001	270.5	44	<,01	0,004
		zu E	3590,1	712	<,01	0,956	61,1	22	<,01	0,001	295.2	44	<,01	0,004
		zu M	3652,3	712	<,01	0,958	94,6	22	<,01	0,001	397.2	44	<,01	0,005
9	B	zu D	3625,7	712	<,01	0,943	213,2	22	<,01	0,004	755.7	44	<,01	0,014
		zu E	3724,3	712	<,01	0,945	64,0	22	<,01	0,001	285.9	44	<,01	0,004
		zu M	3987,0	712	<,01	0,942	140,0	22	<,01	0,002	752.1	44	<,01	0,013
5	D	zu E	2766,2	712	<,01	0,956	8,3	22	<,01	0,000	131.1	44	<,01	0,002
		zu M	2902,5	712	<,01	0,949	29,1	22	<,01	0,000	291.9	44	<,01	0,006
7	D	zu E	3726,9	712	<,01	0,959	49,7	22	<,01	0,000	102.8	44	<,01	0,000
		zu M	3787,7	712	<,01	0,961	81,9	22	<,01	0,001	244.0	44	<,01	0,003
9	D	zu E	3395,1	712	<,01	0,951	134,7	22	<,01	0,002	474.7	44	<,01	0,008
		zu M	3660,3	712	<,01	0,948	81,5	22	<,01	0,002	710.6	44	<,01	0,012
5	E	zu M	2933,5	712	<,01	0,951	23,6	22	<,01	0,000	177.9	44	<,01	0,003
7	E	zu M	3817,4	712	<,01	0,960	75,1	22	<,01	0,001	165.1	44	<,01	0,002
9	E	zu M	3758,8	712	<,01	0,949	78,3	22	<,01	0,001	567.2	44	<,01	0,009

Anmerkungen: B: Biologie; D: Deutsch; E: Englisch; M: Mathematik; 5: Klassenstufe fünf; 7: Klassenstufe sieben; Die hervorgehobenen Werte überschreiten die Schwelle von $\Delta CFI \geq 0.020$.
Chi²: Chi^2-Wert des Tests der Modellanpassung; *df*: Freiheitsgrade; *p*: Signifikanzniveau der Chi^2-Tests; *CFI*: Comparative Fit Index; ΔChi^2: Chi^2-Wert des Differenztests; Δdf: Differenz der Freiheitsgrade; *ΔCFI*: Differenz der Comparative Fit Indices

Die generellen Modellspezifikationen erreichen in allen explizierten Mehrgruppenmodellen eine angemessene Modellanpassung. Für alle Vergleiche, sowohl über die Zeit als auch über die Fächer, kann konfigurale Invarianz festgestellt werden. Metrische Invarianz kann ebenfalls in jeder der Prüfungen angenommen werden, also verschlechtert sich die Passung der Modelle bei der Einführung einer Gleichheitsrestriktion der Faktorladungen zwischen den Gruppen nicht signifikant. Die latenten Variablen haben also inhaltlich vermutlich die gleiche Bedeutung. Im nächsten Schritt wurde die skalare Invarianz geprüft, indem auch die Intercepts zwischen den Gruppen fixiert wurden. Mit erfolgreicher Feststellung skalarer Invarianz sind die Ausprägungen der latenten Mittelwerte generell vergleichbar. Dies kann in einem der Tests, im Fach Mathematik, zwischen der Klassenstufe fünf und

sieben, eher nicht angenommen werden (ΔCFI = 0,020). Die latenten Mittelwertunterschiede zwischen den Klassenstufen, die auf Basis dieses Modells beobachtet werden, können also nicht ohne weiteres interpretiert werden, da es substantielle Abweichungen in den Mittelwerten der Indikatoritems gibt. Nach Inspektion der einzelnen Indikatormittelwerte und bei einer Prüfung auf partielle metrische Invarianz (Byrne et al., 1989) konnte beobachtet werden, dass starke MI angenommen werden kann (ΔCFI = 0,019), wenn das Item „*Unsere Mathematiklehrerin/unser Mathematiklehrer bemüht sich, dass alle im Unterricht mitkommen.*" ein Indikator der latenten Dimension Empathie, zwischen den Gruppen freigesetzt wird.

Weitere Vergleiche erreichten die kritische Schwelle von ΔCFI = 0,02 nicht, überschritten aber die strengere Schwelle ΔCFI = 0,01. Dies gilt für alle längsschnittlichen Gruppenvergleiche zwischen den Klassenstufen fünf und sieben. Weitere Vergleiche, bei denen dies zutrifft, sind Deutsch und Biologie, Deutsch und Mathematik, sowie Biologie und Mathematik.

Es konnte beobachtet werden, dass zwischen den Klassenstufen fünf und sieben innerhalb aller Fächer dieselben latenten Dimensionen relativ hohe oder geringe Differenzen (κ) aufwiesen. Die Differenzen basieren auf der Metrik der Markervariablen und werden hier inklusive des Standardfehlers der Differenz dargestellt. In den Fächern Biologie, Deutsch und Englisch sind die Differenzen jeweils so ähnlich hoch ausgeprägt, dass diese im Mittel berichtet werden. Es liegen keine signifikanten Unterschiede zwischen Deutsch und Englisch und nur punktuelle geringe Unterschiede (die maximale Differenz beträgt 0,10; die mittlere Differenz beträgt 0,03) zwischen den sprachlichen Fächern und Biologie vor. Im Fach Mathematik fallen alle Differenzen deutlich höher aus. Die Differenzen sind in der Tabelle 5 abgetragen.

Die höchsten Differenzen können in den Dimensionen *Empathie* und *Anregung* beobachtet werden, aber generell liegt in allen Dimensionen mit Ausnahme der *Leistungsbezogenen Differenzierung* eine signifikante Abweichung vor ($p \leq$,01).

Tabelle 5: Zusammengefasste Differenzen der latenten Mittelwerte zwischen Klassenstufe 5 und Klassenstufe 7

	Dimension	κ_{bde}	$s.e._{bde}$	κ_m	$s.e._m$
Mittelwerte der Fächer Biologie, Deutsch und Englisch	Klassenmanagement	-0.03	0.04	0.07	0.05
	Diag. Kompetenz/Ang.	-0.25	0.03	0.03	0.03
	Schüler-Lehrer-Verh.	-0.33	0.05	0.03	0.05
	Strukturiertheit	-0.10	0.04	0.04	0.04
	Anregung	-0.37	0.05	-0.03	0.05
	Empathie	-0.39	0.04	0.03	0.05
	Leistungsbezogene Diff.	0.04	0.04	-0.08	0.03
Mathematik	Klassenmanagement	-0.19	0.04	0.12	0.05
	Diag. Kompetenz/Ang.	-0.44	0.03	0.02	0.03
	Schüler-Lehrer-Verh.	-0.49	0.05	0.07	0.06
	Strukturiertheit	-0.22	0.04	0.09	0.04
	Anregung	-0.63	0.05	0.10	0.05
	Empathie	-0.62	0.04	0.13	0.05
	Leistungsbezogene Diff.	-0.07	0.04	0.24	0.04

Anmerkungen: Die Mittelwerte der Klassenstufe 5 sind auf 0 fixiert;
κ: Mittelwertdifferenz zwischen den Klassenstufen 5 und 7; s.e.: Standardfehler der Mittelwertdifferenz;
B: Biologie; D: Deutsch; E: Englisch; M: Mathematik

5 Diskussion und Ausblick

Das Ziel dieses Beitrages war es, eine theoretisch fundierte und empirisch angemessene Struktur der gemeinsamen Beobachtung verschiedener Dimensionen der Unterrichtsqualität aus der Perspektive der Schülerinnen und Schüler abzuleiten und zu prüfen, inwieweit ein als empirisch passend bewertetes Modell über verschiedene Fächer und Klassenstufen hinweg verglichen werden kann und welche Differenzen in den beobachteten Werten vorliegen. Dieser Weg der Erfassung der Unterrichtsqualität ist für Schulen im Prozess der Schul- und Unterrichtsentwicklung ebenso hilfreich wie für Interventionsforscher; zielen beide doch auf die Verbesserung der Prozessqualität des Unterrichts.

Wie unter anderem bei Schwanenberg et al. (2015) festgestellt, sind inhaltlich durchaus abgrenzbare theoretische Konstrukte in einer Operationalisierung als einzelne Faktoren hoch korreliert. In diesem Beitrag wurde ein den vorliegenden Daten entsprechend angemessenes Modell erstellt, um diese Korrelationen zu berücksichtigen (I.). Das Modell kann, basierend auf den reinen Modellanpassungen, gut auf unterschiedliche Fächer und Klassenstufen übertragen werden. Gleich spezifizierte Modelle konnten also technisch in jedem Fach und zu jedem Messzeitpunkt angenommen werden. Die Muster der Korrelationen zwischen den latenten Dimensionen

deuten darauf hin, dass die Wertungen der zwischenmenschlichen Kompetenzen der Lehrkräfte durch die Schülerinnen und Schüler besonders stark zusammenhängen. Dies gilt insbesondere in der Klassenstufe 5. Auf der anderen Seite hängen Dimensionen, welche die Unterrichtsgestaltung betreffen, verhältnismäßig gering zusammen, was den Befund bisheriger Studien stützt, dass das methodisch-didaktische Wissen von Schülerinnen und Schülern Einschränkungen aufweist (Ditton & Arnold, 2004; Ditton et al., 2002) und folglich Schülerurteile diese Dimensionen betreffend als störanfällig einzuschätzen sind. Die Variabilität der Korrelationen über die Zeit ist dabei als verhältnismäßig gering zu bewerten.

Um zu bestimmen, ob die Modelle auch inhaltlich vergleichbar sind, wurden in einem zweiten Schritt Testungen auf MI über die Zeit und zwischen den Fächern vorgenommen. Es kann beobachtet werden, dass generell ein hohes Maß an MI vorliegt. Zwischen allen vier Fächern innerhalb aller drei Messzeitpunkte (II.) konnte Invarianz der Faktorladungen und der Intercepts der Indikatorvariablen festgestellt werden. Damit ist die Übertragbarkeit der Modelle von einem Fach auf ein anderes Fach grundsätzlich gegeben, wie von Helmke (2012) angenommen wurde. Somit könnte also beispielsweise die Wirksamkeit von fachunspezifischen Interventionen auch über verschiedene Fächer hinweg beobachtet werden. Dabei ist hervorzuheben, dass das Modell sowohl für das Fach Mathematik, als auch für ein naturwissenschaftliches und zwei sprachliche Fächer Gültigkeit hat, was als ein breites Spektrum von unterschiedlichen Unterrichtsfächern und somit als Form der Validierung der Modellkomposition begriffen werden kann. Diese hohe Übertragbarkeit ermöglicht wertvolle querschnittliche Vergleiche, wie sie bislang nicht abgesichert vorgenommen werden konnten. Studien, die die Unterrichtsqualität in den Mittelpunkt des Erkenntnisinteresses stellen und vergleichbare Skalen verwenden, könnten somit auf erweiterte Vergleichsmöglichkeiten zurückgreifen.

Für die Vergleiche über die Zeit (III.) konnte in drei der vier geprüften Fächer skalare Invarianz und im Fach Mathematik partielle skalare Invarianz beobachtet werden. Dies bedeutet, dass die aus diesem Modell heraus beobachteten Mittelwerte und Strukturen generell zwischen den Klassenstufen fünf und sieben sowie sieben und neun verglichen werden können. Aber auf der Basis nur knapp erreichter Schwellenwerte konnte ein Muster ausgemacht werden, welches darauf hin deutet, dass die latenten Mittelwerte zwischen den Klassenstufen fünf und sieben bedeutsam voneinander abweichen. Zwischen den Klassenstufen sieben und neun liegt keine derartige Abweichung vor. Bei der konkreten Betrachtung der Differenzen der latenten Mittelwerte konnten wir feststellen, dass sich die höchsten Ver-

schiebungen in den Dimensionen zeigen, welche zwischenmenschliche Kompetenzen der Lehrkräfte abbilden. Eine Zäsur an dieser mittleren Altersschwelle wäre dahingehend vorhersagbar gewesen, da von zunehmend kritischen Urteilen der Schülerinnen und Schüler über ihre Lehrkräfte über das steigende Alter auszugehen ist, was sich insbesondere während des Übergangs von der Kindheit ins Teenageralter zeigt (z. B. Czerwenka et al., 1990). Zudem scheint das Fach Mathematik in allen Dimensionen bedeutsam größere Differenzen zwischen der Klassenstufe fünf und der Klassenstufe sieben aufzuweisen, als alle anderen beobachteten Fächer. Dies trifft in besonderem Maße für die Dimensionen *Empathie* und *Diagnostische Kompetenz* zu. Auch hier wäre nun, basierend auf den vorliegenden Ergebnissen, zu prüfen, inwieweit dieses Phänomen generell vorliegt.

Die Alterspanne der befragten Schülerinnen und Schüler reicht grob von 10 bis 16 Jahren, also von der Kindheit bis ins Jugendalter, was als breites Spektrum begriffen werden kann, für welches es keinesfalls als sicher gelten konnte, dass die Ausprägungen der latenten Konstrukte inhaltlich vergleichbar sind. Es konnte aber festgestellt werden, dass die Ausprägungen der latenten Dimensionen über das Alter zwar differieren, aber die Konstrukte größtenteils inhaltlich vergleichbar sind. Auf Basis dieser Ergebnisse kann für zukünftige Forschungsarbeiten die wichtige Voraussetzung der metrischen Äquivalenz, z. B. für den Einsatz von latenten Wachstumskurvenmodellen oder Latent-Change-Score Modellen, als gegeben angesehen werden.

Eine substantielle Interpretation der einzelnen Dimensionen, des abgeleiteten generellen Strukturmodells und der strukturellen Zusammenhänge können an dieser Stelle nicht erfolgen, da dies den Umfang des Beitrags überschreiten würde. Die Ergebnisse liefern aber wertvolle Erkenntnisse über die Voraussetzungen derartiger Interpretationen. Der folgende Schritt wäre nun, die generellen latenten Strukturen zwischen den Dimensionen der Unterrichtsqualität in den Blick zu nehmen, zu interpretieren und deren Zusammenhänge mit Drittvariablen zu betrachten, also beispielhaft konditional für verschiedene Schulprogramme oder die Beschulung in innovativen Lernarrangements. Auch längsschnittliche Analysen unter Einbezug von Drittvariablen sind möglich. Es muss aber festgehalten werden, dass skalare Invarianz, als Voraussetzung für inhaltsvolle Interpretierbarkeit der Wachstumskomponente (Widaman et al. 2010), im Fach Mathematik nur partiell gegeben ist. Die Frage danach, warum ein Item im Fach Mathematik anscheinend anders funktioniert als in anderen Fächern, zu beantworten, liegt leider jenseits der Möglichkeiten dieses Beitrags.

Unklar bleibt auch, warum die Wahrnehmung der Unterrichtsqualität durch die Schülerinnen und Schüler im Fach Mathematik im Mittel zwischen den Klassenstufen fünf und sieben besonders stark zurückgeht. Denkbar ist, dass Leistungsdifferenzen oder Differenzen in den Lerngeschwindigkeiten im Fach Mathematik deutlicher wahrgenommen werden. In folgenden Analysen wäre es hier von Interesse, die Differenzen in den latenten Mittelwerten zwischen den Fächern vertiefend zu betrachten.

Es muss betont werden, dass die vorliegenden Daten aus Ganztagsgymnasien stammen. Da hier ausschließlich der curriculare Fachunterricht bewertet werden sollte, ist anzunehmen, dass der Ganztagsaspekt kein Gewicht hatte, allerdings konnte nicht geprüft werden, ob die angenommenen Strukturen auf andere Schulformen zu übertragen sind.

Es gibt verschiedene technische Bereiche, die kritisch zu vermerken sind. Aufgrund der hohen Zahl der Einzelergebnisse, musste im Rahmen dieses Beitrags häufig auf Aggregate zurückgegriffen werden. Wenn die Streuungen der Ergebnisse auffällig gewesen sind, wurde dies berichtet.

Eine Überprüfung der residualen Invarianz wurde nicht vorgenommen. Somit muss festgehalten werden, dass die latenten Werte unter Umständen mit unterschiedlich großen Fehleranteilen gemessen wurden.

Die Mehrebenenstruktur wurde bei der Modellierung nur über die Korrektur der Standardfehler und des Chi^2-Wertes berücksichtigt, obwohl in vorangegangenen Beiträgen (z. B. Kunter, 2005 oder Wagner et al., 2015) und in diesem Beitrag nachgewiesen werden konnte, dass ein substantieller Teil der Varianz von Konstrukten der Unterrichtsqualität auf der Klassenebene liegt. Die hier festgestellten *ICC* liegen im Rahmen der von Wagner et al. (2013) und Kunter (2005) festgestellten Variabilitäten auf der Klassenebene. Da hier nur auf der Individualebene modelliert wurde, ist davon auszugehen, dass beobachtete Korrelationen systematisch überschätzt wurden (Lüdtke et al., 2009), da diese durch Effekte auf der zweiten (Klassen-)Ebene konfundiert, also nicht unabhängig sind und die Effekte auf der Klassenebene als höher anzunehmen sind.

Gleichzeitig wird aber betont, dass im Fall des vorliegenden Beitrags das zentrale Forschungsinteresse auf der Individualebene angesiedelt war, womit diese Analyseebene als angemessen erschien. Zudem gab es eine hohe Zahl von Klassenwechseln im Längsschnitt, sodass eine persistente hierarchische Zuteilung über den Längsschnitt nicht passend gewesen wäre. Stattdessen wären Modelle für multiple Mitgliedschaften (*Cross-Classified* oder *Multiple-Membership* Modelle; z. B. Skrondal und Rabe-Hesketh, 2004) angemessener (Ditton, 2002). Dieses Vorgehen wurde hier als überkomplex bewertet. Nichtsdestotrotz zeigen vorangegangene Studien und die

vorliegenden *ICC* auf, dass bedeutsame Varianzanteile auf der Klassenebene liegen und diese nun, also nachdem die generelle Vergleichbarkeit nachgewiesen worden ist, für Wirkungsanalysen in den Blick genommen werden sollten.

Die Auswahl des Kriteriums für die Annahme von MI und die Abfolge der Schwellen ist keinesfalls als *goldene* Regel zu verstehen. Weitere Fit Indices kämen in Abhängigkeit zur Modellkomplexität, der Anzahl der Indikatoren, dem Ausmaß der Faktorladungen und dem Stichprobenumfang ebenso in Frage (Marsh et al., 2004). Um eine bessere Vergleichbarkeit zwischen verschiedenen Studien bei der Beobachtung von MI zu erreichen, wäre es hilfreich, wenn an dieser Stelle klare Regularien erarbeitet werden könnten.

Im Anschluss an Helmke (2012) ist anzunehmen, dass guter Unterricht nicht durch positive Werte auf allen Dimensionen definiert ist, sondern die Dimensionen komplementär zueinander zu betrachten sind. Hier wäre im Anschluss an die vorliegenden Ergebnissen zu überprüfen, inwieweit sich verschiedene Typen von Unterrichtswahrnehmung abbilden lassen und ob diese innerhalb einzelner Schulklassen stabil sind. Der vorliegende Datensatz bietet eine gute Grundlage, um Fragen nach der Kategorisierbarkeit von Unterricht nachzugehen.

Zusammenfassend haben wir eine rechnerisch angemessene Struktur von Dimensionen der Unterrichtsqualität abgeleitet und deren Vergleichbarkeit über verschiedene Fächer und eine weite Altersspanne belegt. Als Grundlage dessen wurde auf eine übliche Befragung aus Schülerperspektive und bewährte Instrumente zur Bewertung der Unterrichtsqualität zurückgegriffen. Die Ergebnisse liefern auch Hinweise darauf, dass Unterrichtsqualität anhand vielfältiger, inhaltlich ähnlicher Kriterien zu bewerten ist, wie auch schon die unterschiedlichen Merkmalskataloge andeuten, und die Bewertung der Unterrichtsqualität unabhängig vom Fach erfolgen kann. Dies impliziert aber ebenso, dass bei weiteren Untersuchungen die Zielperspektiven des Fachunterrichts und die fachdidaktischen Zugänge eine ebenso bedeutende Rolle bei der Bewertung des Unterrichts spielen und für konkrete Bewertungen mit aufgenommen werden müssen.

Lehrerhandeln kann also laut der vorliegenden Ergebnisse multikriterial fachübergreifend und über eine breite Altersspanne aus der Perspektive von Gymnasiasten der Klassenstufen fünf bis neun bewertet werden.

Referenzen

Altrichter, H., Messner, E. & Posch, P. (2004). *Schulen evaluieren sich selbst. Ein Leitfaden.* Seelze: Kallmeyer.

Berkemeyer, N., Bos, W., Holtappels, H. G., Meetz, F & Rollett, W. (2010). „Ganz In": das Ganztagsgymnasium in Nordrhein-Westfalen: Bestandsaufnahmen und Perspektiven eines Schulentwicklungsprojekts. In N. Berkemeyer, W. Bos, H. G. Holtappels, N. McElvaney & R. Schulz-Zander (Hrsg.), *Jahrbuch der Schulentwicklung Band 16. Daten, Beispiele und Perspektiven* (S. 131–152). Weinheim: Beltz.

Bollen, K. (1989). *Structural Equations with Latent Variables.* New York: Wiley.

Bos, W., Lankes, E.-M., Prenzel, M., Schwippert, K., Valtin, R., Voss, A. & Walther, G. (2005). *IGLU: Skalenhandbuch zur Dokumentation der Erhebungsinstrumente.* Münster u. a.: Waxmann.

Bühner, M. (2010). *Einführung in die Test- und Fragebogenkonstruktion* (3. akt. Aufl.). München: Pearson.

Brophy, J. E. (2000). *Teaching (Educational Practices Series, Vol. 1).* Brussels: International Academy of Education & International Bureau of Education (www.ibe.unesco. org).

Byrne, B. M., Shavelson, R. J. & Muthén, B. (1989). Testing for the Equivalence of Factor Covariance and Mean Structures: The Issue of Partial Measurement Invariance. *Psychological Bulletin, 105*(3), 456–466.

Chen, F. F. (2007). Sensitivity of Goodness of Fit Indexes to Lack of Measurement Invariance. *Structural Equation Modeling A Multidisciplinary Journal, 14*(3), 464–504.

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

Clausen, M. (2002). *Unterrichtsqualität – eine Frage der Perspektive?* Münster: Waxmann.

Creemers B. & Kyriakides L. (2008). *The Dynamics of Educational Effectiveness.* London: Taylor & Francis.

Czerwenka, K., Nölle, K., Pause, G., Schlotthaus, W., Schmidt, H. J. & Tessloff, J. (1990). *Schülerurteile über die Schule. Bericht über eine internationale Untersuchung.* Frankfurt am Main, New York: P. Lang.

Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P. & Billiet, J. (2014). Measurement equivalence in cross-national research. *Annual Review of Sociology, 40*, 55–75.

Ditton, H. (2001). *DFG-Projekt „Qualität von Unterricht und Schule" – QuaSSU Skalenbildung Hauptuntersuchung.* www.quassu.net/SKALEN_1.pdf. Zugegriffen: 30. März 2016.

Ditton, H. (2002). Unterrichtsqualität – Konzeptionen, methodische Überlegungen und Perspektiven. *Unterrichtswissenschaft, 30*(3), 197–212.

Ditton, H. (2009). Unterrichtsqualität. In K.-H. Arnold, U. Sandfuchs & J. Wiechmann (Hrsg.), *Handbuch Unterricht* (S. 235–243). Bad Heilbrunn: Klinkhardt.

Ditton, H. & Arnoldt, B. (2004). Wirksamkeit von Schülerfeedback zum Fachunterricht. In J. Doll & M. Prenzel (Hrsg.), *Bildungsqualität von Schule: Lehrerprofessionalisierung, Unterrichtsentwicklung und Schülerförderung als Strategien der Qualitätsverbesserung* (S. 152–172). Münster: Waxmann.

Ditton, H., Arnoldt, B. & Bornemann, E. (2002). Entwicklung und Implementation eines extern unterstützenden Systems der Qualitätssicherung an Schulen – QuaSSu. In M. Prenzel & J. Doll (Hrsg.), Bildungsqualität von Schule: Schulische und außerschulische Bedingungen mathematischer, naturwissenschaftlicher und überfachlicher Kompetenzen. *Zeitschrift für Pädagogik. 45. Beiheft* (S. 374–389). Weinheim: Beltz.

Einsiedler, W. (2002). Das Konzept „Unterrichtsqualität". *Unterrichtswissenschaft, 30*(3), 194–196.

Fauth, B., Decristan, J., Rieser, S., Klieme, E. & Büttner, G. (2014). Student ratings of teaching quality in primary school: dimensions and prediction of student outcomes. *Learning and Instruction, 29*, 1–9.

Guenole, N. & Brown, A. (2014). The consequence of ignoring measurement invariance for path coefficients in structural equation models. *Frontiers in Psychology*, *5*, 980.

Helfrich, H. (1993). Methodologie kulturvergleichender psychologischer Forschung. In A. Thomas (Hrsg.), *Kulturvergleichende Psychologie* (S. 53–102). Göttingen: Hogrefe.

Helmke, A. (2003). *Unterrichtsqualität – erfassen, bewerten, verbessern*. Seelze: Kallmeyer.

Helmke, A. (2012). *Unterrichtsqualität und Lehrerprofessionalität. Diagnose, Evaluation und Verbesserung des Unterrichts* (4. Überarbeitete Aufl.). Seelze: Klett-Kallmeyer.

Horn, J. L. & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging Research. *Experimental Aging Research*, *3*, 117–144.

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

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, *36*, 409–426.

Klieme, E. & Rakoczy, K. (2008). Empirische Unterrichtsforschung und Fachdidaktik. Outcome-orientierte Messung von Prozessqualität des Unterrichts. *Zeitschrift für Pädagogik*, *54*(2), 222–237.

Klieme, E., Pauli, C. & Reusser, K. (2009). The Pythagoras Study: Investigating effects of teaching and learning in Swiss and German mathematics classrooms. In T. Janik & T. Seidel (Hrsg.), *The power of video studies in investigating teaching and learning in the classroom* (S. 137–160). Münster: Waxmann.

Klieme, E., Schümer, G. & Knoll, S. (2001). Mathematikunterricht in der Sekundarstufe I: „Aufgabenkultur" und Unterrichtsgestaltung. In E. Klieme & J. Baumert (Hrsg.), *TIMSS – Impulse für Schule und Unterricht. Forschungsbefunde, Reforminitiativen, Praxisberichte und Video-Dokumente* (S. 43–57). München: BMBF.

Klieme, E., Jude, N., Baumert, J. & Prenzel, M. (2010). PISA 2000–2009: Bilanz der Veränderungen im Schulsystem. In E. Klieme, C. Artelt, J. Hartig, N. Jude, O. Köller, M. Prenzel, W. Schneider & P. Stanat (Hrsg.), *PISA 2009. Bilanz nach einem Jahrzehnt* (S. 277–279). Münster: Waxmann

KMK – Ständige Konferenz der Kultusministerien der Länder in der Bundesrepublik Deutschland (2002). *PISA 2000 – Zentrale Handlungsfelder': Zusammenfassende Darstellung der laufenden und geplanten Maßnahmen*. https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2002/2002_10_07-Pisa-2000-Zentrale-Handlungsfelder.pdf Zugegriffen: 30. März 2016.

Köller, O. (2008). Bildungsstandards in Deutschland: Implikationen für die Qualitätssicherung und Unterrichtsqualität. In M. A. Meyer, M. Prenzel & S. Hellekamps (Hrsg.), *Perspektiven der Didaktik* (S. 47–59). Wiesbaden: VS Verlag für Sozialwissenschaften.

Kunter, M. (2005). *Multiple Ziele im Mathematikunterricht*. Münster: Waxmann.

Kunter, M. & Baumert, J. (2006). Who is the expert? Construct and criteria validity of student and teacher ratings of instruction. *Learning Environments Research*, *9*, 231–251.

Kunter, M., Brunner, M., Baumert, J., Klusmann, U., Krauss, S., Blum, W., Jordan, A. & Neubrandt, M. (2005). Der Mathematikunterricht der PISA-Schülerinnen und -Schüler. Schulformunterschiede in der Unterrichtsqualität. *Zeitschrift für Erziehungswissenschaft*, *8*, 502–520.

Kunter, M., Tsai, Y.-M., Klusmann, U., Brunner, M., Krauss, S. & Baumert, J. (2008). Students' and mathematics teachers' perceptions of teacher enthusiasm and instruction. *Learning and Instruction*, *18*(5), 468–482.

Little, T. D. (2013). *Longitudinal Structural Equation Modeling*. New York: Guilford.

Lüdtke, O., Trautwein, U., Schnyder, I. & Niggli, A. (2007). Simultane Analysen auf Schüler- und Klassenebene. Eine Demonstration der konfirmatorischen Mehrebenen-Faktorenanalyse zur Analyse von Schülerwahrnehmungen am Beispiel der Hausaufgabenvergabe. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, *39*, 1–11.

Lüdtke, O., Robitzsch, A., Trautwein, U. & Kunter, M. (2009). Assessing the impact of learning environments: how to use student ratings of classroom or school characteristics in multilevel modeling. *Contemporary Educational Psychology, 34*, 77–88.

Marsh H. W., Hau K-T. & Wen Z. (2004). In Search of Golden Rules: Comment on Hypothesis-Testing Approaches to Setting Cutoff Values for Fit Indexes and Dangers in Overgeneralizing Hu and Bentler's (1999) Findings. *Structural Equation Modeling: A Multidisciplinary Journal, 11*(3), 32–41.

Martin, M. O., Foy, P., Mullis, I. V. S. & O'Dwyer, L. M. (2013). Effective schools in reading, mathematics, and science at fourth grade, In M. O. Martin & I. V. S. Mullis, (Hrsg.), *TIMSS and PIRLS 2011: Relationships among Reading, Mathematics, and Science Achievement at the Fourth Grade-Implications for Early Learning* (S. 109–178). Chestnut Hill, MA: TIMSS & PIRLS International Study Center.

Meade, A. W., Johnson, E. C. & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology*, 93, 568–592

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research, 13*, 127–143.

Meredith, W. (1993). Measurement Invariance, Factor Analysis and Factorial Invariance. *Psychometrika, 58*(4), 525–543.

Meyer, H. (2004). *Was ist guter Unterricht?*. Berlin: Cornelsen.

Muthén, L. K. & Muthén, B. O. (1998–2007). Mplus User's Guide. Fifth Edition. Los Angeles, CA: Muthén & Muthén.

Praetorius, A.-K. (2014). *Messung von Unterrichtsqualität durch Ratings*. Münster: Waxmann.

Quellenberg, H. (2009). *Studie zur Entwicklung von Ganztagsschulen (StEG) – ausgewählte Hintergrundvariablen, Skalen und Indices der ersten Erhebungswelle*. Band 24. Frankfurt am Main: Materialien zur Bildungsforschung.

Rakoczy, K. (2008). *Motivationsunterstützung im Mathematikunterricht: Unterricht aus der Perspektive von Lernenden und Beobachtern*. Münster: Waxmann.

Reinecke, J. (2005). *Strukturgleichungsmodelle in den Sozialwissenschaften*. München: Oldenbourg.

Reinecke, J. & Pöge, A. (2010). Strukturgleichungsmodelle. In: C. Wolf & H. Best (Hrsg.), *Handbuch der Sozialwissenschaftlichen Datenanalyse* (S. 775–804). Wiesbaden: VS Verlag für Sozialwissenschaften.

Reusser, K. & Pauli, C. (2010). Unterrichtsgestaltung und Unterrichtsqualität – Ergebnisse einer internationalen und schweizerischen Videostudie zum Mathematikunterricht: Einleitung und Überblick. In: K. Reusser, C. Pauli & M. Waldis (Hrsg.), *Unterrichtsgestaltung und Unterrichtsqualität. Ergebnisse einer internationalen und schweizerischen Videostudie zum Mathematikunterricht* (S. 9–32). Münster: Waxmann.

Scherer, R., Nilsen, T. & Jansen, M. (2016). Evaluating individual students' perceptions of instructional quality: An investigation of their factor structure, measurement invariance, and relations to educational outcomes. *Frontiers in Psychology, 7*: 110.

Schulte, K., Nonte, S. & Schwippert, K. (2013). Die Überprüfung von Messinvarianz in international vergleichenden Schulleistungsstudien am Beispiel der Studie PIRLS. *Zeitschrift für Bildungsforschung, 3*(2), 99–118.

Schwanenberg, J., Winkelsett, D. & Schurig, M. (2015). Methoden der wissenschaftlichen Begleitforschung im Projekt Ganz In. In: H. Wendt & W. Bos (Hrsg.), *Auf dem Weg zum Ganztagsgymnasium* (S. 414–443). Münster: Waxmann.

Seidel, T., Meyer, L. & Dalehefte, I. M. (2005). „Das ist mir in der Stunde gar nicht aufgefallen…" – Szenarien zur Analyse von Unterrichtsaufzeichnungen. In M. Welzel & H. Stadler (Hrsg.), *Nimm doch mal die Kamera! Zur Nutzung von Videos in der Lehrerbildung – Beispiele und Empfehlungen aus den Naturwissenschaften* (S. 133–154). Münster: Waxmann.

Skrondal, A. & Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models*. Boca Raton, FL: Chapman & Hall/CRC.

Slavin, R. E. (1997). *Educational Psychology* (5. Aufl.). Boston: Allyn and Bacon.

Sörbom, D. (1974). A general method for studying differences in factor means and factor structure between groups. *British Journal of Mathematical and Statistical Psychology, 27*, 229–239.

Terhart, E. (2006). 'Was wissen wir über gute Lehrer?'. *PÄDAGOGIK, 58*(5), 42–47.

van de Schoot, R., Schmidt, P., de Beuckelaer, A., Lek, K. & Zondervan-Zwijnenburg, M. (2015) Editorial: Measurement Invariance. *Frontiers in Psychology*, 6: 1064.

Vandenberg, R. J. & Lance, C. (2000). A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research. *Organizational Research Methods, 3*(1), 4–70.

Wagner, W., Göllner, R., Helmke, A., Trautwein, U. & Lüdtke, O. (2013). Construct validity of student perceptions of instructional quality is high, but not perfect: Dimensionality and generalizability of domain-independent assessments. *Learning and Instruction, 28*, 1–11.

Wagner, W., Göllner, R., Werth, S., Voss, T., Schmitz, B. & Trautwein, U. (2015). Student and Teacher Ratings of Instructional Quality: Consistency of Ratings Over Time, Agreement, and Predictive Power. *Journal of Educational Psychology*, Sept. 21, Keine Seitenzahlen vergeben.

Waldis, M., Grob, U., Pauli, C. & Reusser, K. (2010). Der schweizerische Mathematikunterricht aus der Sicht von Schülerinnen und Schülern und in der Perspektive hochinferenter Beobachterurteile. In K. Reusser, C. Pauli & M. Waldis (Hrsg.), *Unterrichtsgestaltung und Unterrichtsqualität. Ergebnisse einer internationalen und schweizerischen Videostudie zum Mathematikunterricht* (S. 171–208). Münster: Waxmann.

Wendt, H. & Bos, W. (Hrsg.). *Auf dem Weg zum Ganztagsgymnasium*. Münster: Waxmann.

Widaman, K. F. & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In: Bryant, K. J. & Windle, M. (Hrsg.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (S. 281–324). Washington, DC: American Psychological Association.

Widaman, K. F., Ferrer, E. & Conger, R. D. (2010). Factorial invariance within longitudinal structural equation models: Measuring the same construct across time. *Child Development Perspectives, 4*(1), 10–18.

Willems, A. S. (2016). Unterrichtsqualität und professionelles Lehrerhandeln. Prozesse und Wirkungen guten Unterrichts aus dem Blickwinkel der empirischen Schul- und Unterrichtsforschung. In R. Porsch (Hrsg.), *Einführung in die Allgemeine Didaktik. Ein Lehr- und Arbeitsbuch für Lehramtsstudierende* (S. 289–338). Stuttgart: UTB.

Willems, A. S. & Glesemann, B. (2015). Individuelle Förderung und der Umgang mit Heterogenität im Fachunterricht an Ganztagsgymnasien. In: H. Wendt & W. Bos (Hrsg.), *Auf dem Weg zum Ganztagsgymnasium* (S. 414–443). Münster: Waxmann.

Wu, A. D., Li, Z. & Zumbo, B. D. (2007). Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: A demonstration with TIMSS data. *Practical Assessment, Research and Evaluation, 12*(3), 1–26.

BEITRAG 6

Erschienen in (Zitierweise):

Jaekel, N., Schurig, M., Florian, M. & Ritter, M. (im Erscheinen [2017]). From Early Starters to Late Finishers? A Longitudinal Study of Early Foreign Language Learning in School. *Language Learning*.[62]

Relevanz:

Der Beitrag beschäftigt sich mit den Vergleichen von mehrdimensionalen Leistungsergebnissen im Fach Englisch. Die Leistungsergebnisse von Kohorten mit unterschiedlichem Umfang von Englischunterricht in der Grundschule werden über IRM gemeinsam skaliert und deren spezifischer Trend wird deskriptiv und in Abhängigkeit zu verschiedenen Kovariablen in einem SEM beobachtet. Dabei werden die Güte der IRM und des SEM getrennt evaluiert. Analog zum Beitrag 4 wird die Kombination verschiedener Modelle vorgenommen, indem IRM geschätzt und dann als manifeste Größen in ein SEM übertragen werden.

[62] Es hier abgedruckte Fassung ist eine vorläufige Version. Die finale Version kann unter der angegebenen Zitation gefunden werden. (Transl.: *This version is a preleminary draft. The final version can be acquired via the Journal Language Learning.*)

From Early Starters to Late Finishers? A Longitudinal Study of Early Foreign Language Learning in School

Nils Jaekel [a,b], Michael Schurig [c], Merle Florian [b], Markus Ritter [b]

a University of Tennessee, Knoxville, b Ruhr-University Bochum, c TU Dortmund University

Keywords: early foreign language learning, receptive language skills, learner characteristics, longitudinal, structural equation modelling

1. Introduction

Foreign language learning in pre- and elementary schools has seen a rapid development in Europe. Parents, educators and politicians in Europe have long been advocating for early foreign language education in mainstream schooling believing in the credo "the earlier, the better". This movements' impact can be recognized in Europe's aim of fostering a plurilingual, multicultural society throughout the continent. The "2+1" policy is an ambitious agenda to promote language learning to the extent that every European is fluent in at least two languages in addition to their mother tongue (Council of the European Union, 2002; European Commission, 1995). In hope of achieving this bold goal, early foreign language education was identified as a potentially viable approach (Council of the European Union, 1997). Consequentially, across Europe, foreign languages education in elementary school has become the rule rather than the exception with a total enrollment surpassing 78% in 2010 (Education, Audiovisual & Culture Executive Agency, 2012).

Subsequently, research into early foreign language education has been growing steadily over the past decade (e.g. (Larson-Hall, 2008; Muñoz, 2006; Pfenninger, 2014a).This surge of research interest can, at least partially, be attributed to the

European Union's (EU) language agenda. Countries throughout Europe have been adapting their language policies following EU encouragement. Primary aims are the promotion of multilingualism and multicultural understanding as well as catering to the growing demand of fluent second language (L2) speakers for the job market.

The principal research objective of the present study is to investigate the sustainability of an early EFL onset in Year 1 (age 6-7 years) as opposed to Year 3 (age 8-9 years) and its impact on students' English language proficiency in Years 5 and 7. The study investigates the effects of moving English Foreign Language (EFL) education from Year 3 to Year 1 of elementary school in Germany. Two cohorts of students, one cohort of early starters (ES, age 6-7 years) beginning in the second half of the first school year and one cohort of late starters (LS, age 8-9 years) commencing EFL in Year 3. Additionally, individual student characteristics such as gender, parents' socioeconomic status, cultural capital, home language and cognitive abilities will be included as background variables in statistical analyses to determine their impact on students' language proficiency. Two crucial, inextricably linked factors are central in the present study: the age of onset of EFL teaching and the amount of exposure to English. As the earlier start results in more English lessons, it is impossible to separate the effects either measure has on language proficiency.

2. Early foreign language learning: brief background & research perspectives

The assumption that the earlier language learning starts, the better the results will be, still prevails in the minds of many parents and policy makers (European Commission, 2011). While the roots of this belief are manifold, a certain disconnect between SLA

research and early foreign language advocates is evident: either research did not inform policy making well enough or has been misinterpreted. DeKeyser and Larson-Hall (2005, p. 88) argue that

> [i]n the practical realm, the younger is better argument has been both used and abused, both refuted and misunderstood by advocates of early intervention from the very beginning of formal immersion education to this day .

2.1.A critical look at early foreign language learning research

2.1.1. The age issue in the school context

Research into early foreign language learning, more specifically the age factor, has experienced a significant surge in recent years (de Bot, 2014; Enever, 2011; Muñoz, 2006; Pfenninger & Singleton, 2016). Before, much of what was known about second and foreign language acquisition at an early age was based on research from language acquisition at home or through immersion programs. These findings have been applied to the context of learning a language in preschool or elementary school despite considerable differences, particularly regarding the amount of exposure to the L2 (Muñoz, 2010).

In the context of early foreign language research, a crucial distinction needs to be made between age of onset versus amount of exposure. Comparing early with late starters with a focus on age of onset, one needs to take into consideration that the amount of exposure to the L2 differs considerably and thus skews results and their interpretation. Yet, research in the field of early foreign language education has often neglected this distinction in study designs and analyses (Muñoz, 2006), which widens the margin of error due to inclusion of confounds.

The benefits propagated for moving early SLA into elementary or preschools are manifold. First and foremost, the increased exposure to the L2, which is expected to lead to linguistic benefits in form of higher language proficiency and potentially more native-like pronunciation (Flege & MacKay, 2011), advantages in academic achievement (Taylor & Lafayette, 2010) and (inter-) cultural advantages (Nikolov & Mihaljević Djigunović, 2011), as well as moderate advantages on phonological and morphosyntactic levels (Larson-Hall, 2008), which may however disappear quickly (Pfenninger, 2014b). Furthermore, younger learners show fewer problems with language anxiety (Johnstone, 2009), higher levels of motivation and positive attitudes towards language learning (Börner, Engel & Groot-Wilken, 2013; Graham, Courtney, Tonkyn & Marinis, 2016; Mihaljevic Djigunovic & Lopriore, 2011), better future employability is another potential advantage. These attributed benefits are however not always applicaple, for example due to differences in the age of students, contextual differences of studies or a lack of rigouros research. Reviewing SLA research of the 1960s and 1970s, considering L2 acquisition within and outside of formal instruction, Krashen, Long and Scarecella (1979) concluded that research had consistently pointed to advantages for older learners, adults or children, who acquired language at a faster rate while onset of L2 acquisition during early childhood lead to higher proficiency. Amount of exposure trumped age of onset, although early starters were able to catch up relatively quickly (Krashen et al., 1979). More recent studies provide evidence that older learners (12 years and older), particularly in the beginning stages, consistently learn at a faster rate than younger learners (Mihaljevic Djigunovic, Nikolov & Otto, 2008; Muñoz, 2008; Pfenninger, 2014b). In her longitudinal project Muñoz (2006), for example,

comprehensively assessed L2 development with different groups of language learners who received from 200 to 726 hours of language instruction. Groups differed in the age of onset, i.e. 10-12, 12-14, 14-18 years and adults. The later in life learners started, i.e. as adolescents or adults, the more progress they had made initially after 200 hours. It took the youngest group of learners (10 - 12 years) 726 hours to catch up to the older learners without outperforming them (Muñoz, 2006). Based on these findings, it has been argued that the level of cognitive maturity favors older learners (Cummins, 1981), especially regarding methodology based on explicit instruction applied in secondary education (Muñoz, 2006). Older learners' (13 vs. 8 years) advantages have been attributed to their ability to learn explicitly, i.e. make use of their metalinguistic knowledge such as applying rules, a more distinct goal- and future-goal motivation in SLA (Pfenninger & Singleton, 2016), their more advanced L1 literacy and oracy skills while benefiting from the cognitive advantages of a more mature brain which has implications for test-taking strategies. Once younger learners' cognitive development catches up to that of older learners, younger learners close the proficiency gap. To achieve similar levels of language proficiency, younger learners would thus require increased amount of exposure to achieve what older learners accomplish in a shorter time. In conclusion, under non-immersive conditions or without increased exposure in school environments, amount of exposure is more important than time of onset (Curtain, 2000; Muñoz, 2006; Steinlen, Hâkansson, Housen & Schelletter, 2010; Unsworth, Persson, Prins & Bot, 2015).

An early onset of foreign language education potentially increases both length and overall amount of exposure to the L2. . Current curricula, however, do not always

reflect these results. The curriculum of North-Rhine-Westphalia, Germany, for example, moved foreign language learning to the second term of Year 1 with two 45 minute lessons per week while the number of lessons at the secondary level in grammar schools was reduced by one lesson per week to four for Years 5 and 6 (MSW - NRW, 2015b; Verband Bildung und Erziehung Landesverband NRW, 2006). By moving EFL into Year 1, the overall number of hours of EFL across elementary and secondary schooling was roughly retained. In light of the available research this adjustment, however, cannot be viewed as a serious sign of interest in promoting EFL education. As described above, research into early foreign language learning has shown that to achieve equal or higher levels of language proficiency, younger learners, compared to learners that start later, require longer and more intense exposure. This means that minimal input of an hour or two per week does not suffice (DeKeyser & Larson-Hall, 2005; Lightbown & Spada, 2006). Thus, traditional foreign language education in school can hardly provide the required contact time with an L2.

2.1.2. Implementation of EFL

Learning an L2 through methodology focusing on implicit learning, i.e. unconscious, playful-like acquisition of the L2 through meaningful exposure and communicative activities, contextualized L2 and an absence of metalanguage (Housen & Pierrard, 2005). Younger learners require more exposure to target language structures

"to infer rules without awareness" and "internalize the underlying rule/pattern without their attention being explicitly focused on it" (Ellis, 2009, S. 16; Mihaljevic Djigunovic et al., 2008; Muñoz, 2008). The transition to secondary school methodology

shifts from a stronger emphasis on implicit learning to explicit learning and teaching, i.e. the overt teaching and learning of the L2 such as grammar or vocabulary and the use of metalinguistic skills (Housen & Pierrard, 2005). Metalanguage slowly builds in students' L1 during elementary school years and thus students cannot yet benefit from a transfer to L2. Explicit and implicit learning rely on different neural systems in the brain, whereas the latter requires significantly fewer cognitive resources (Ellis, 2009; Ellis, 2011). It is important to note that although we acquire our L1 implicitly, an ability that diminishes with age, it takes considerably longer to gain the ability to communicate in L1 compared to classroom based L2 (Birdsong, 2006).

Methodological recommendations in elementary school curricula, that are in line with the idea of implicit learning, however, have received criticism With in Germany, specifically Baden-Wuerttemberg. Schmelter criticizes policy makers' understanding that second language learning should resemble L1 language acquisition (Schmelter, 2010; Standing Conference of the Ministers of Education, 2013). However, the allocated time of around 90 minutes per week can hardly be compared to the amount of exposure in L1 acquisition. In addition, L1 acquisition processes are characterized by constant exposure, frequent repetition and supportive one-to-one settings from the family as well as community support and interaction with peers in preschool. Thus, these circumstances cannot simplybe compared to L2 acquisition in an institutional setting with limited exposure to the L2 provided by non-native teachers. Furthermore, while methodology may build on students' L1 acquisition experience, the L2 does not

necessarily have the same instrumental significance for students (Nicholas & Lightbown, 2008). For students to benefit from an early start, more exposure to the L2 in a meaningful context would thus be required. Following the cognition hypothesis which argues that increasing the cognitive demand of tasks will foster interaction and "push learners to greater accuracy and complexity of L2 production" (Robinson, 2003, S. 45, 2007), immersion or content-based instruction could provide this context while increased cognitive demands should provide the added benefit of processing the L2 more deeply, i.e. through elaboration rehearsal linking language to content, images or in-depth analyses, as opposed to traditional second language learning, particularly as these approaches rely on implicit learning (de Graaf & Housen, 2009).

A further point of criticism involves the hasty implementation of foreign language education into the elementary school curriculum without a sufficient number of qualified teachers (Edelenbos, Johnstone & Kubanek, 2006; Piske, 2013). The few research studies available have highlighted that the better qualified and more proficient EFL teachers in elementary schools are, the higher EFL students score (de Bot, 2014; May, 2007; Unsworth et al., 2015). Elementary school teachers are "the most important stakeholders," but Nikolov and Djigunovic (2006) have also criticized the lack of research on their abilities, level of training, methodological knowledge and language proficiency.

Despite the criticisms, there are adjustments that can help ensure success in SLA in and beyond elementary education. First, increasing the amount of classroom exposure beyond an hour or two per week (Larson-Hall, 2008; Muñoz, 2008) and enhanced L2

exposure through extra-mural English outside of school, e.g. listening to music, reading books or watching movies in the L2 (Mihaljevic Djigunovic et al., 2008). Second, ensuring EFL teachers' language proficiency and language specific training for younger student populations is imperative (Marinova-Todd, Marshall & Snow, 2000; Mihaljevic Djigunovic et al., 2008). Lastly, facilitating a smooth transition, e.g. acknowledging and building on L2 skills, adjusting methodology to that used in elementary school and communicating with elementary school teachers, from elementary to secondary education will greatly improve early SLA (e.g. (Kolb & Mayer, 2010; Marinova-Todd et al., 2000).

2.2. Reading and listening in early foreign language learning

In the context of the present study receptive language skills, i.e. listening and reading, were assessed. Particularly listening skills are essential as they constitute the main source of language input in the communicative language classroom. Learning implicitly, elementary school students need ample input to master this complex process that involves neurological, linguistic, semantic, pragmatic processes that work interdependently to decode spoken language (Rost, 2011). Transforming sound to meaning learners rely on both "bottom-up", the discrimination of sounds into meaningful units, and "top-down" processes, established knowledge and context (Vandergrift & Goh, 2012). With increased levels of language proficiency and therewith a growing body of knowledge, listening becomes progressively more automated (Vandergrift & Goh, 2012).

Although L2 reading is only gradually introduced in early foreign language learning (MSW - NRW, 2008), it promotes the overall language development, especially phonological and phonemic awareness, thus linking listening to reading and ultimately writing skills (Dlugosz, 2000) and orthography. Similar to listening, reading is an interactive process involving various simultaneous "top-down" (integrating knowledge and expectations) and "bottom-up" processes (decoding letters and words; (Birch, 2015; Grabe & Stoller, 2011).

3. Learner characteristics and their relationship to L2 proficiency in the school context

A second focus of this study is on the investigation of individual learner characteristics, i.e. gender, parents' socio-economic status, cultural capital, home language and cognitive abilities, and their contribution to L2 proficiency in the context of early EFL education. In large-scale assessments such as PISA, individual differences have often been shown to have significant repercussions for academic achievement, e.g. gender or parents' socio-economic status (SES; (Ehmke & Jude, 2010; OECD, 2015; Sälzer, Reiss, Schiepe-Tiska, Prenzel & Heinze, 2013). Disentangling the age-factor from other individual difference variables such as motivation, attitudes or environmental variables, e.g. teaching quality or teacher proficiency, poses a considerable challenge (see (Pfenninger & Singleton, 2016). Despite their general importance for learning across different subjects, research on foreign language acquisition has often failed to incorporate individual difference variables (Csapó & Nikolov, 2009; Nikolov & Djigunovic, 2006). This might lead to enlarged margins of error and the overestimation of effects attributable to the variables under scrutiny.

Our understanding of gender differences in SLA is often based on "common knowledge" (van der Slik, van Hout & Schepens, 2015, S. 1) and the "widespread belief … that females tend to be better L2 learners than males…" (Saville-Troike, 2006, S. 90). Although focusing on adult learners, van der Slik et al.'s investigation of gender differences in SLA of immigrants from 88 countries and with 49 mother tongues in the Netherlands (N=27,119) contributes important insights to the discussion of gender differences in L2 language learning and proficiency. Women's speaking and writing skills consistently surpassed the scores achieved by men, while the latter showed slight advantages in reading scores (van der Slik et al., 2015). No differences were found for listening comprehension. Studies on EFL attainment either in elementary or secondary education in Germany generally tend to confirm the "widespread belief" of a gender advantage even for younger learners: The large-scale study KESS 4 (*Kompetenzen und Einstellungen von Schülerinnen und Schülern [Competences and attitudes of students]*) focused on English language proficiency just before students' transition to secondary education after Year 4 in Hamburg, Germany. For listening comprehension, the data showed slightly higher scores for girls than for boys (May, 2007). In the follow-up studies KESS 8 and 10/11 in years 8, 10 and 11 at high school, girls maintained their lead over boys in listening comprehension and outperformed boys with respect to general language proficiency measured through literacy based C-Tests (Nikolova, 2011). The DESI (*Deutsch-Englisch-Schülerleistungen-International [Assessment of Student Achievements in German and English as a Foreign Language]*) study (N=10,543) comprehensively evaluated Year 9 students' English skills across secondary school types in Germany. Boys performed almost as well as girls in English listening comprehension,

whereas reading and particularly writing skills were significantly better performed by girls (Hartig & Jude, 2008). These findings of girls' superior language performance, particularly with regard to literacy-based tests, are corroborated by other large-scale studies (Courtney, Graham, Tonkyn & Marinis, 2015; Hartig & Jude, 2008; Jaekel, 2015; Nikolova & Ivanov, 2010; Rumlich, 2016). For speaking skills, however, slight advantages have been reported for boys regarding pronunciation, fluency and sentence structure (Nold & Rossa, 2008).

Home language constitutes another key variable in SLA. Teachers today need to cope with cultural and linguistic heterogeneity in their EFL lessons, ideally incorporating this diverse environment into their teaching. Research on the impact of non-majority L1 on foreign language learning (L3) is only growing slowly. However, the ability to speak two or more languages alone does not seem to positively impact L3 acquisition whereas evidence has been presented that biliteracy in L1 and L2 does (Rauch, Naumann & Jude, 2012; Rauch, 2014). Following the threshold hypothesis (Cummins, 1976), the (very) early start of foreign language learning may be detrimental to learners' development in L2 with limited L1 proficiency. Slight advantages for students with a non-majority L1 German background in EFL classes have been reported in the large-scale DESI study in Germany (Hesse, Göbel & Hartig, 2008). For English listening comprehension in Year 4 in Hamburg, no significant differences were found between students who have at least one parent that was born in Germany whereas students with two foreign-born parents scored significantly lower (May, 2007). These results were also confirmed for receptive English proficiency in the Netherlands (de Bot, 2014). In Switzerland, Pfenninger (2014b) draws attention to the potential struggle

learners have to face if their L1 is neither German nor French, which means that for most students English will be their L4. Other studies have found non-L1 students in Year 8 (age 13/14) to achieve significantly lower general language proficiency scores, measured with literacy based C-Tests, than the majority of L1 students (Nikolova & Ivanov, 2010). Even where no significant differences regarding general ELP measured through C-Tests were found, non-L1 German students reported lower grades than their L1 German peers (Jaekel, 2015). As a consequence, for the German state of Baden-Wuerttemberg, Baumert and colleagues argue for moving back early EFL teaching from Year 1 to Year 3 to allow more time for language development in German for students whose L1 is not German (Baumert et al., 2011).

SES and parents' education has been consistently identified as a key predictor of academic success. Studies have shown that lower-SES children enter school with poorer school-readiness skills, which in turn results in overall lower academic achievement and lower cognitive functioning scores (Blair & Raver, 2012). Along the same lines, the PISA studies have highlighted the influence of social variables, e.g. economic and cultural capital, on student attainment across the participating countries (Ehmke & Jude, 2010). Similarly, the impact of SES and other background variables, like the number of books at home, was repeatedly reported in the Trends in International Mathematics and Science Study (TIMSS) and Progress in International Reading Literacy Study (PIRLS) studies (Gustafsson, Hansen & Rosén, 2013). SES has been shown to have a considerably strong impact on L2 achievement in Germany (May, 2007; Nikolova & Ivanov, 2010), a result that coincides with general academic achievement (Klieme, 2010; Muñoz, 2007). These results are generally understood to be a result of early streaming of students for

secondary school after Year 4 of elementary school and the tiered structure of secondary education in Germany (Muñoz, 2007). Lindgren und Muñoz (2013) report that parent's education predicted students' L2 English reading scores.

The cultural capital an individual or family has at their disposal, i.e. the number of books, level of reading enjoyment, as well as museum and concert visits, have been shown to benefit students' academic achievement (Bourdieu & Passeron, 1990). According to PISA, books are a form of objectified cultural capital (Gräsel, Göbel & Stark, 2007). Research on the effects of cultural capital on academic success has, to a large extend, confirmed their positive relationship (see for example Jæger (2011) for a detailed discussion). Whereas L1 has been shown to benefit from cultural capital, in form of books, reading habits or enjoyment of reading (Georg, 2004; Jæger, 2011). Large-scale studies including L1 literacy measures regularly include different measures of cultural capital, SLA rarely considers cultural capital as a variable.

Cognitive abilities have been associated with academic achievement in general and have also been identified as a key variable in L2 acquisition and learning through empirical research (Dallinger, 2015; Genesee, 1976; Jaekel, 2015; Skehan, 1998; Sparks, Patton & Ganschow, 2012). In this context, it is important to distinguish between aptitude and intelligence. Both dimensions are closely related; however, the former is a domain-specific measure of ability, whereas the latter is general in nature and is independent from experience or knowledge (Biedroń, 2011). Beyond language learning itself, intelligence is highly relevant for test-taking (Muñoz, 2008).

4. Research Questions

The present study compares the EFL proficiency of late starters age 8-9 years (LS; the beginning of Year 3 at elementary school) with early starters age 6-7 years (ES; second half of Year 1), who have one and a half years of additional language exposure. The study will compare English reading and listening comprehension scores of both groups at the beginning of Year 5 (t_1; after 2 years/140 hours or 3.5 years/245 hours of English language teaching, respectively) and the beginning of Year 7 (t_2; after 4 years/444 hours or 5.5 years/549 hours, respectively).

Research question 1: Do early starting (ES) students in early English language education outperform late starting (LS) EFL students in language proficiency over time?

Past research (see 3. above) has shown that gender, SES, cultural capital, cognitive abilities and home language may significantly affect foreign language proficiency.

Research question 2: What impact do individual differences (gender, SES, cultural capital, IQ and home language) have on language proficiency?

5. Method and Instruments

5.1. Research Context

5.1.1. Brief background on early foreign language learning in Germany

Germany recently introduced early foreign language education into elementary schools. In most states in Germany, after four years of elementary school students are streamed into secondary schools based on teacher ratings (Ditton & Krüsken, 2006). The early streaming of students in the Germany regularly receives criticism as it causes

social disparities by disadvantaging learners with a lower socio-economic or migration background (Muñoz, 2007). The three-tiered secondary school system consists of *Haupt*- (lower-secondary) and *Realschule* (middle-secondary), which offer a six-year middle-school degree, and *Gymnasium* (upper-secondary) or *Gesamtschule* (comprehensive school), which offer high-school diplomas that provide access to tertiary education.

Most states introduce foreign language learning in Year 3, while six states have opted for a start in Year 1 (Standing Conference of the Ministers of Education, 2013). North-Rhine Westphalia, for example, moved the start of EFL into Year 1 in 2008 after its initial inclusion into the elementary school curriculum in 2003 (MSW - NRW, 2008). Despite the difference in the length, the majority of states in Germany aim for an A1 proficiency at the end of elementary school in Year 4 for reading, writing, listening, speaking and mediation (Standing Conference of the Ministers of Education, 2013), according to the Common European Framework of Reference for Languages (Council of Europe, 2001; Standing Conference of the Ministers of Education, 2013).

5.1.2. Teaching methodology and transition to secondary school

Teaching methodology at this early age (6-7 years in Germany) significantly differs from that used in secondary education and needs to account for students' level of cognitive development as well as other learner characteristics such as L1 proficiency and level of motivation. Particularly in the earliest stages, oracy, i.e. listening and speaking, is emphasized while literacy, i.e. reading and writing, supports, for example, word recognition through regular exposure (MSW - NRW, 2008). Early foreign language

education relies heavily on communicative language learning which aims to provide ample exposure and opportunities to use the L2.

In the process of early foreign language learning, the ultimate success, i.e. sustaining high levels of motivation and continuous development of language proficiency, hinges upon a successful transition from elementary to secondary education. The transition has been identified as an abrupt shift (1) from implicit to explicit teaching and learning that requires metalinguistic knowledge and (2) from an oracy focused curriculum to one that builds on literacy and steep grammatical progression (BIG-Kreis, 2009). Furthermore, a lack of communication between elementary and secondary schools has been identified as a crucial problem (Thürmann, 2009). This is particularly important in harmonizing methodology and content in secondary EFL education.

5.2.Procedure

Student participation was voluntary, written consent was obtained from parents before data collection commenced. Demographic data were collected and language testing was performed between week five and nine of the new school year. Data for the LS cohort were collected in the summers of 2010 and 2012 for Years 5 and 7, respectively while the ES cohort's data were collected in 2012 and 2014. Data collection was conducted during regular school lessons.

Analyses were conducted using SPSS 23 (IBM Corp, 2015b), AMOS 23 (IBM Corp, 2015a) and ConQuest 3.0.1 (Adams, Wu & Wilson, 2012).

5.3.Sample

The sample consists of two cohorts of students (N=5,130) from 31 grammar schools in North-Rhine-Westphalia (NRW), Germany Participating schools are distributed somewhat evenly across North-Rhine Westphalia so that rural and urban areas are considered to similar extents. All 31 schools participated voluntarily in the longitudinal *Ganz In – All Day-Schools for a Brighter Future* study endorsed by the Ministry of Education (MSW - NRW, 2015a). At elementary schools in the state of NRW, English was introduced in Year 3 in 2003; in 2008 it was moved forward into the second half of Year 1. The LS cohort (N=2,632), who started elementary school in 2006 and secondary school in 2010, had received two years (140 hours) of EFL teaching prior to secondary education whereas the ES cohort (N=2,498) had received 3.5 years (245 hours) at that stage. The ES group started school in 2008 and moved on to secondary school in 2012, respectively, which corresponds to a lag of two years between the two sample groups and makes them the first cohort with additional EFL teaching at elementary school (105 additional hours spread over 1.5 years at the age of 6-7 in Years 1 and 2 of elementary school).[1] Only students that took part in the proficiency tests in Year 5 and Year 7 were included in the analyses. Of the initial sample ($N = 6{,}652$; $N_{LS} = 3{,}312$; $N_{ES} = 3{,}340$), 77.1% were retained and assessed longitudinally at the end of Years 5 and 7.

Table 1 Descriptive values, test statistics and effect sizes between cohorts for background variables

N_{total}=5,140	Cohort LS (n_1=2,632)	Cohort ES (n_2=2,498)	Test statistic *(df)*	*p*	Cohen's *d*
Gender (females) (in %)[a]	46.73	47.20	0.11 (1)	.739	-0.009
Age (years) mean (*SD*)[b]	10.37 (5.6)	10.35 (5.3)	-132.68 (5,127)	.182	0.037
Mother tongue (German only) (in %)[a]	75.87	72.80	6.05 (1)	.015	0.097
KFT 5 (raw score, max=25) mean (*SD*)[b]	17.75 (6.20)	17.79 (6.00)	-0.23 (4,815)	.821	-0.007
Books (category from 1 to 5) mean (*SD*)[b]	3.51 (1.13)	3.44 (1.12)	2.36 (5,065)	.042	0.062
Income (category from 1 to 8) mean (*SD*)[b]	5.43 (2.17)	5.58 (2.22)	-2.12 (3,673)	.022	-0.068
English grade in Year 4 (from 1 to 6) mean (*SD*)[b]	1.74 (0.60)	1.73 (0.62)	0.80 (4,963)	.796	0.016

Annotations: Books: 1 '0-10 Books', 2 '11-25 Books', 3 '26-100 Books', 4 '101-200 Books', 5 'more than 200 Books'; Income: 1='less than 10k Euros a year', 2='10k-19,999', 3='20k-29,999', 4='30k-39,999', 5='40k-49,999', 6='50k-59,999',
7='60k-69,999', 8='70k or more'; English grade: 1 = 'very good', 6 = 'insufficient'; [a] χ^2-Tests were used for significance testing.; [b] *t*-Tests were used for significance testing.

Table 1 displays a range of statistics investigating the differences between the two cohorts regarding individual learner characteristics at the beginning of Year 5, i.e. their first year of secondary school. Furthermore, Cohen's *ds* were calculated as indicators for the size of the effects. Early and late starters did not differ significantly with regard to gender (p = .739; see Table 1), age (p = .182), cognitive ability (p = .821), and English grade at the end of Year 4 (p = .796). At t_1 students were on average ten years old and classes consisted of slightly more boys than girls. Significant differences between both cohorts were revealed based on their reported mother tongue (p = .013), cultural capital (p = .018), and parental income (p = .034). In the LS group, fewer students reported to have a non-German L1 background than in the ES group. ES

parents have a slightly higher income (M_{LS} = 5.43 vs. M_{ES} = 5.58, p = .022) while LS households are characterized by a slightly higher cultural "book" capital (M_{LS} = 5.51 vs. M_{ES} = 5.44, p = .042). The effect sizes for these differences were, however, (very) small indicating negligible effects. According to Wolf (1986), the threshold for Cohen´s *d* magnitudes of academic relevance in educational research is .25. In the subsequent statistical analyses, these background variables will be controlled for.

5.4.Instruments

As part of this study receptive English language skills, i.e. listening and reading comprehension, were assessed both at the beginning of Year 5 (t_1) and 7 (t_2). Standardized reading and listening tests were employed that had previously been validated in other large scale studies in Germany (see below). Both listening and reading assessments consisted of item batteries largely based on multiple choice and occasional multi-word or short single-sentence responses.

Year 5 receptive language proficiency was assessed based on the previously validated scales from the Evening study (Engel & Ehlers, 2013). For listening, students answered 28 multiple choice questions on picture recognition (17) and sentence completion (11) in German For reading, 20 multiple choice and four open answer items assessed text understanding (α = .71). In Year 7, the selected scales originated from a state-wide assessment scheme in Brandenburg (*Vergleichsarbeiten der Jahrgangsstufe 6* (VERA 6) [comparative tests in Year 6]; Institute for Educational Quality Improvement [IQB]). For listening, 11 sentence completion items from the Year 5 assessment were used. In addition, 15 multiple choice and 7 open answer items were used to assess

English listening (α = .79). For reading, students answered 11 multiple choice and 15 open answer items (α = .89; (Institut zur Qualitätssicherung im Bildungswesen). Responses were coded dichotomously (correct vs. incorrect). The tests used in Year 5 and Year 7 respectively were identical for both cohorts. To obtain proficiency scores, a simple one-dimensional logistic item-response-model (IRM; (Rasch, 1980) was calculated. In this process, the item parameters were estimated first before the person parameters were fixed. Items were checked for conformity by assessing item characteristic curves, discrimination parameters, mean squared errors (MNSQ), and the respective t-values. The thresholds to determine acceptable item fit were chosen according to common guidelines used in large-scale-assessment studies (Adams & Wu, 2002; Wright & Linacre, 1994). Items that did not conform to the model were excluded prior to the main analyses. A weighted likelihood estimator (Warm, 1989) was employed. To obtain better estimates for the differences within cohorts, both Year 5 and Year 7 sub-samples were scaled simultaneously. The final scores were standardized to a mean of 500 points with a standard deviation (*SD*) of 100.

Cognitive abilities were assessed with the subtest Figural Analogy of the *Kognitiver Fähigkeitstest* [cognitive abilities test] (Heller & Perleth, 2000), as it allows for an estimate of students' general cognitive abilities independent of their L1 language proficiency. In the present sample, the test reached good reliability in line with the norm-sample values (α_{norm} = .94; α_{LS} = .92; α_{ES}=.90).

Demographic variables including cultural capital and home language were based on student´s responses. Regarding cultural capital, the students were asked how many

books were present at their specific household. Five categories were offered: '0-10', '11-25', '26-100', '101-200' or 'more than 200'. Additionally, five pictograms of bookshelves with an according number of books were depicted in the questionnaires to help the children estimate the number of books at home. For home language, participants were asked which language they regarded as their mother tongue. Where data was not available from students because of non-response, parental responses were used instead when available. Income was based on parents' responses. Parents were asked to classify the gross household income per year into the following categories: 'below 10.000€', '10.000 to 19.999€', '20.000 to 29.999€', '30.000 to 39.999€', '40.000 to 49.999€', '50.000 to 59.999€', '60.000 to 69.999€' and '70.000 € or more'. All scales originated from the international TIMSS and PIRLS assessments and were adapted for the use in Germany (Bos, Bonsen, Gröhlich, Guill & Scharenberg, 2009). There was no information on the children's age for 0.9%, on students' L1(s) for 0.1%, on the KFT score for 6.1%, on books for 1.2% and on the income of the household for 28.2%.

6. Results

6.1. Do early starters outperform late starters?

In Year 5, the ES cohort significantly outperformed their LS peers in both reading (M_{ES} = 514.24; M_{LS} = 486.49) and listening comprehension (M_{ES} = 517.24 / M_{LS} = 483.64; Figure 1). The results of the independent t-tests suggest the existence of statistically significant differences between the two cohorts' language proficiency at the beginning of Year 5 for reading ($p \leq .001$; see Table 2) and listening comprehension ($p \leq .001$). The one and a half years of additional EFL or 105 hours translated into a significant advantage of 27.5 points on average in reading and 33.6 point in listening

comprehension, equaling about a third of a standard deviation. The resulting differences amount to a small to medium effect sizes (see Table 2 Cohen*'s ds*).

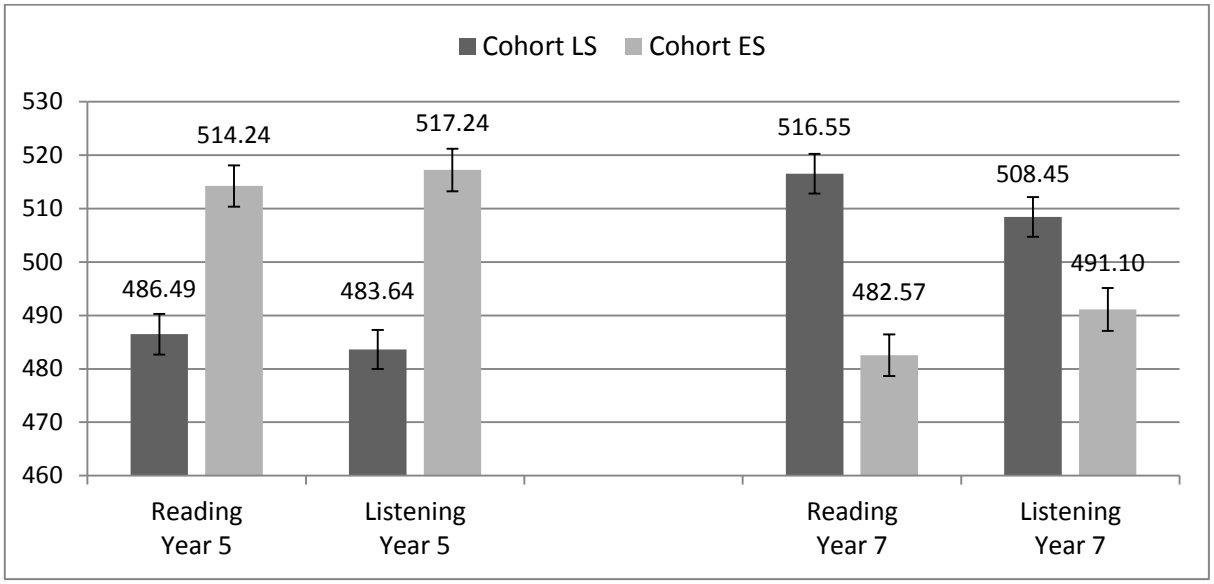


Figure 1 Mean proficiency scores for Reading and Listening in Years 5 and 7 with confidence intervals (95%)

In Year 7, the LS cohort outscored the ES in reading comprehension 516.55 to 482.57 and in listening comprehension 508.45 to 491.10. On average LS scored almost 34 points more in reading and 17.35 in listening comprehension than their ES peers. Independent samples t-test confirmed the significant differences between the two groups for reading ($p \leq .001$; see Table 2) and listening comprehension ($p \leq .001$). In the following two years of secondary education after the initial test, the late starters were able to close the gap and they outperformed their early starting peers.

Table 2 Comparing language proficiency of both cohorts in reading and listening comprehension in Year 5 and 7, using independent samples t-tests and *Cohen's* d

N = 5,130	Cohort LS (n_1=2,632)	Cohort ES (n_2=2,498)	Test statistic *(df)*	*p*	Cohen's *d*
Reading Year 5 (mean (SD))[a]	486.49 (99.45)	514.24 (98.61)	-10.03 (5,128)	<.001	-0.280
Reading Year 7 (mean (SD))[a]	516.55 (97.59)	482.57 (99.56)	12.34 (5,128)	<.001	0.345
Listening Year 5 (mean (SD))[a]	483.64 (95.82)	517.24 (101.43)	-12.18 (5,068)	<.001	-0.341
Listening Year 7 (mean (SD))[a]	508.45 (96.86)	491.1 (10.47)	6.23 (5,068)	<.001	0.174

Annotations: [a] *t*-Tests were used for significance testing.

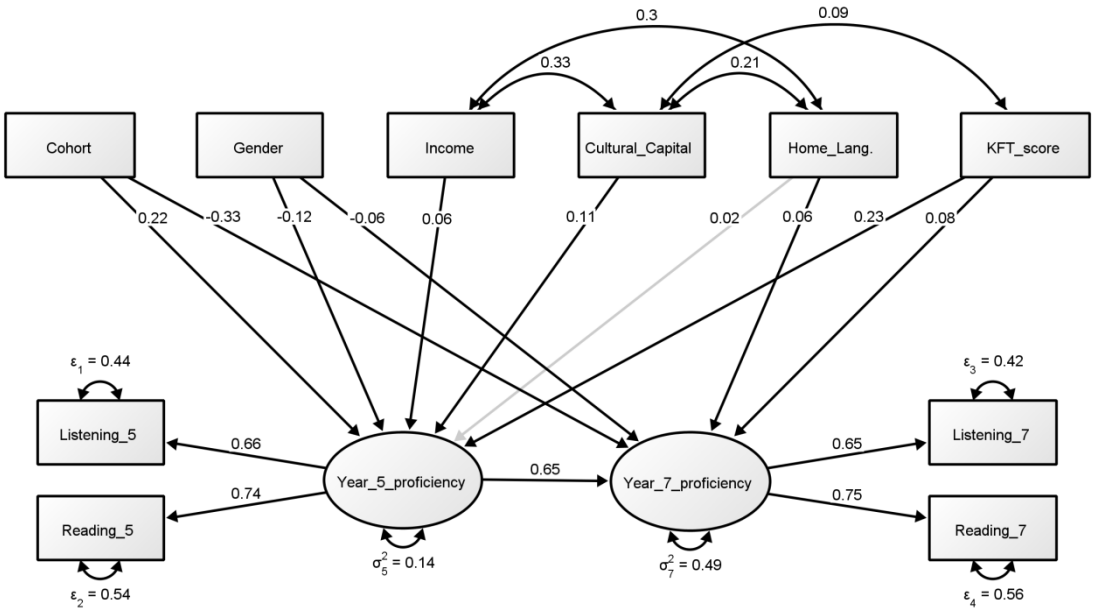
6.2.Which impact do individual differences, e.g. gender, SES, cultural capital, IQ and home language, have on language proficiency?

To assess the complex interrelationships of individual differences and the impact they exert on language proficiency, structural equation modeling (SEM) was employed. SEM provides an analytic framework to observe single or multiple latent variables, their structures and the relationships of those variables, while considering measurement errors. A Maximum Likelihood Estimator was used. To handle missing values, full-information-maximum-likelihood was applied (Lüdtke, Robitzsch, Trautwein & Köller, 2007). The SEM model (Figure 2) includes two latent variables consisting of receptive language performance in Years 5 and 7, respectively. A latent regression between the two latent variables defines the core of the structural model. Extraneous variables were added to assess their relative impact. Covariance between extraneous variables was allowed when appropriate. In addition to the dichotomous cohort variable, five background variables were chosen: gender, income (SES), books (cultural capital), home language (L1) and cognitive abilities. The use of cohort membership as a control variable allows the inspection of the relative importance on the latent variables. By including all

covariates simultaneously, the familiar confounding structures between them can be controlled, in particular when there is high collinearity.

To assess model fit we first considered the χ^2 value of 234.37 which was highly significant (p = .001; df = 25). This was to be expected considering the large sample size. Due to the significant χ^2 value, other fit indices have to be considered (Byrne, 2010; Kim & Bentler, 2006). The comparative fit index (CFI) of .964 indicates that the model is fitting as it lies above the .95 threshold (Hu & Bentler, 1999). A second fit index, the root-mean-square-error-of-approximation (RMSEA) together with the PCLOSE which tests the null hypothesis that RMSEA is less than .05 were also employed. A RMSEA of .040 and a PCLOSE of 1.0 corroborate a good fit of the SEM model.

In Figure 2 the complete model is shown. Single headed arrows indicate factor loadings and standardized regression weights, double headed arrows correlations. σ^2 indicates the variance explained and ε the indicator´s errors. Non-significant effects are displayed by faded out arrows.



Annotations: Cultural capital has been operationalized through the amount of books at home; Home Language is the language students mainly speak at home with their parents and siblings (0= not German, 1=German) ; Gender (0=Boys, 1=Girls)

Figure 2 SEM (standardized path coefficients)

Gender, cognitive abilities, cultural and economic capital as well as cohort contributed to explaining variance of receptive language performance in Year 5 (please see Figure 2 for details). Cohort membership and cognitive abilities show the largest effects on receptive language proficiency, i.e. belonging to the ES cohort and scoring highly on the cognitive abilities test have significant positive impacts. Being female and having higher cultural and economic capital significantly predicted higher test scores. The SEM accounted for 14% of the variance of language proficiency for Year 5. Home language (L1) did not have a significant effect on the receptive language proficiency in Year 5.

Receptive language proficiency in Year 7 received the largest contribution to explain its variance from language proficiency in Year 5's. Cohort, cognitive abilities, gender and home language also significantly predicted the receptive language proficiency in Year 7. The second strongest effect on receptive language proficiency for Year 7 is the cohort variable. In accordance with the descriptive analyses discussed above, the influence from Year 5 was inverted and the LS cohort now scored significantly higher. The remaining variables only contributed little to explain Year 7 receptive language proficiency. Girls and students who scored higher on the cognitive function test still performed slightly better on the language test than their peers. While home language had no impact in Year 5, in Year 7 a marginal effect could be observed. SES measured in Year 5, i.e. both cultural and economic capital, did not retain its predictive power on language proficiency in Year 7. Overall the SEM explained 49% of the receptive language proficiency's variance for Year 7. More variance could be explained for Year 7 due to the Year 5 data serving as a predictor.

7. Discussion

The results presented here are in line with the growing body of research that confirms older learners in the long run to be at an advantage in learning a foreign language over students in early foreign language education with minimal input (Krashen et al., 1979; Larson-Hall, 2008; Muñoz, 2006; Pfenninger, 2014a). Our findings confirm that the combination of an earlier onset and thus more exposure to the foreign language leads to a positive short-term effect for receptive language proficiency in English, as demonstrated for the Year 5 data (also see (Wilden, Porsch & Ritter, 2013) for cross-sectional analyses of Year 5 data). Year 7 data presented inverted results. The LS cohort

not only closed the proficiency gap in both reading and listening comprehension but surpassed their ES peers significantly, outperforming them in both reading and listening comprehension. The LS students thus benefited more from the two years of learning English at the grammar school than the ES students. Manifold explanations are being offered in the research literature which may be applicable to the context of this study. Contextual factors such as the amount of L2 instruction as well as the transition from elementary to secondary school including teaching methodology need to be considered in conjunction with students' learner characteristics to account for these results.

7.1.The impact of age on language proficiency

First, despite the common belief that younger learners are better language learners, research has rather consistently shown that older learners make faster progress in foreign language learning, potentially due to higher levels of cognitive maturity and their ability to learn languages through explicit instruction (Krashen et al., 1979; Muñoz, 2006). The LS cohort in the present study quickly caught up to the ES cohort and build a lead of their own. Munoz (2006) argues that older learners particularly benefit more from the rule-based and grammar reliant language teaching in the introduction phase of secondary education EFL classroom environment than younger learners. Our results thus corroborate recent studies (Muñoz, 2006; Pfenninger, 2014b, 2014a) but also Krashen et al.'s meta-analysis (Krashen et al., 1979). While other studies had much more pronounced age gaps between their cohorts (e.g. (Larson-Hall, 2008; Muñoz, 2006; Pfenninger, 2014b), in the present study this gap is only two years. Differences between both cohorts, for example regarding cognitive maturity, would thus be less conspicuous than in other studies.

Second, both cohorts received only minimal English language input of 90 minutes or less per week over the course of three and a half years for the ES and two years, respectively, for the LS. SLA research has shown that the younger students are, the more L2 language exposure they require to retain it and benefit from this approach in the long term (DeKeyser & Larson-Hall, 2005; Muñoz, 2008). Babies and toddlers are constantly exposed to their L1 by parents, family or preschool environment. While methodology may draw on the playful character of the acquisition process, the minimal amount of instruction may have prevented students from benefitting fully an early start EFL early. Although exposure to English was ultimately doubled, transitioning from Year 4 to 5, even the additional exposure by the same body of EFL teachers did not allow the ES cohort to sustain their EFL proficiency lead, while it was sufficient enough for the LS cohort to surpass their ES peers. The reduction of one hour per week at secondary school may be a crucial mistake – cognitive maturity, explicit instruction and growing L1 literacy make L2 learning more effective at this age (Cummins, 1979; Muñoz, 2008). The additional exposure would indicate that the ES cohort would have a head start which they could build on, possibly the exposure in the early years may have been too sparse to have a longer lasting effect.

Third, a central issue in explaining the outcomes of this study and a crucial factor in ensuring successful early foreign language learning is the transition from elementary to secondary education which has been identified as an Achilles' heel of early EFL teaching (BIG-Kreis, 2009). . As grammar schools feed from different elementary schools, communication may not always be ideal to support students' transition. The transition encompasses several critical issues: (a) an abrupt shift from student-centered, playful

methodology to more teacher directed, academically oriented and faster paced lesson rhythms may have impacted the ES cohort more as they had experienced it for longer in elementary school. (b) A potential mismatch of student-teacher expectations regarding methodology used in class may have caused a decrease in motivation (Courtney et al., 2015; Graham et al., 2016). Particularly if students did not receive input that addressed their L2 proficiency level, e.g. either challenging advanced students and helping others catch up, this may easily cause frustration (c) Just as students vary, so does the competence of elementary school English teachers (Nikolov & Djigunovic, 2006), thus teaching outcomes may vary from elementary school to elementary school. Additionally, EFL teachers at the elementary school level have, to a large extend, not been trained extensively as these programs are relatively new to tertiary education in Germany (Edelenbos et al., 2006). (d) As grammar schools welcome students from several elementary schools, teachers have to be particularly thorough in assessing L2 skills that students have already attained. It is likely that teachers were unsure of what to expect from the 'new' more experienced student cohort and may not yet have adapted their lessons to the overall higher level of language proficiency in an attempt to level the playing field by teaching to the middle. The latter has obvious repercussions for teaching in general and student motivation more specifically. A significant advantage for the LS cohort may have been that they could apply their metacognitive knowledge and thus explicit language learning earlier in the language learning career which might have given them a boost and possibly even added to their motivation as this would have led to faster progression and moments of success feeding into their self-efficacy. The question that remains unanswered is whether teachers were ready to pick students up where

their language proficiency lay or if they taught "business as usual" and potentially disappointed their students' expectations and hurt their motivation? (see for example (Courtney et al., 2015).

Forth, teaching methodology in general at both the elementary and secondary schools holds a crucial role to accommodate the early starting age, maximize the outcome of minimal input and allow for a smooth transition. This includes the availability of adequate textbooks, accounting for the increased language proficiency due to the early start. This is particularly important as EFL teaching secondary schools in German heavily relies on textbooks (MSW - NRW, 2012). While no specific data regarding EFL methodology or textbooks were assessed, a mismatch both in Year 5 and 6 between the EFL methodology employed by teachers and the expectations of student, particularly those of the ES cohort, may have discouraged learners leading to a lack in motivation (Courtney et al., 2015). On the contrary, the LS cohort exhibited original evidence for advantages in institutionalized language learning similar to Munoz' (2008) results.

7.2.Learner characteristics and their impact on language proficiency

In addition to the cohort variable which had a significant impact on receptive language proficiency in SEM, students' background variables also contributed significantly. The present study confirms current SLA research results with girls outscoring boys in both English listening and reading comprehension in Year 5 and 7. The present analyses do not allow further insights, but explanations for better performance by girls have been demonstrated based on, for example, advantages

literacy (Courtney et al., 2015), higher interest in language learning (Helmke, Schrader, Wagner, Nold & Schröder, 2008; Rumlich, 2016), higher levels of self-efficacy (Jaekel, 2015; Mills, Pajares & Herron, 2007) or higher motivation (Ruyffelaert & Hadermann, 2012). Ultimately, societal stereotypes of girls being better at language learning may weigh on boys' self-perception as L2 learners (Kissau, 2006).

Cultural and economic capital has been shown to be a consistent variable predicting success in school (e.g. (Klieme, 2010). The two variables covering SES and cultural and economic capital each showed significant results for Year 5, but their impact faded over time. Grammar schools in Germany preselect their students based on academic achievement, the data is thus biased for SES as there are significant differences with regard to SES and choice of school type in Germany (Ehmke & Jude, 2010). Nevertheless, for Year 5 significant differences for SES within classes affected student performance. Particularly the availability of more books at home can be an indicator of increased L1 vocabulary and levels of literacy which would provide students with an advantage learning an L2 (Cummins, 1979). Among the central ideas of all-day schooling in Germany was the mitigation of SES effects through additional teacher support, the present results may be an indicator for its success as most upper-secondary schools nowadays offer such kind of program in Germany (Holtappels, 2006). Another means of explaining could be that students from lower SES background who have not been able to succeed in the cognitively taxing grammar school environment left the school and were thus excluded from our analyses. In the German school context, particularly at grammar schools, grade retention is still fairly common and may have affected the Year 7 data.

Cognitive abilities measured in Year 5 were the most significant predictor of L2 receptive language proficiency in Year 5 and remained significant for Year 7, albeit at a third of its initial impact. Despite using a figurative test instead of test measuring verbal cognitive abilities, these findings corroborate current L2 research (Dallinger, 2015; Jaekel, 2015; Rumlich, 2016) namely that cognitive abilities are a highly relevant variable in language acquisition and SLA research. The quality of L2 teaching, clarity and student centered teaching as well as overall levels of cognitive abilities in a particular class have been shown to moderate the importance of cognitive abilities for successful language learning (Helmke, Helmke et al., 2008).

Home language had no significant impact on ELP in Year 5 but it gained significance in Year 7, in contrast to findings for secondary (Hesse et al., 2008) and elementary schools in Germany (Steinlen, 2016). Comparable results have been reported in the KESS 10/11 study where the English proficiency divide between L1 German and non-German students expanded over time (Nikolova, 2011).

8. Strengths and Limitations

This study is part of a large-scale research project with two cohorts of students surpassing 5000 participants across 31 schools involving German, English, Mathematics, the natural sciences Biology, Chemistry and Physics, all of which had limited time allotted for assessment. With these time constraints come certain drawbacks. For English and German, for example, it would have been more advantageous to assess verbal cognitive abilities instead of, or along-side the figural cognitive abilities test KFT. Time constraints and the feasibility in school to test a large number of students in

speaking skills are also a central reason for choosing receptive language proficiency. Additional individual difference variables such as motivation, attitudes towards language learning, teacher proficiency, quality of teaching and information regarding the transition from elementary school to secondary school could have provided vital insights into the effect of age of onset, but were either not included in the study or would have gone beyond the scope of this paper. Furthermore, the projects' main scope was the investigation of students' development at grammar schools, as these schools select or stream their students based on their academic achievement in elementary school, reading and listening outcomes at other school tiers might vary significantly (see 5.1. for a brief description of the German school system, e.g. (Klieme, 2010). Additionally, due to the setup of the study and the circumstances in both elementary and grammar schools, it is not possible to attribute the results definitively to either age of onset, Year 1 vs. Year 3, or amount of exposure 245 vs. 140 hours of EFL instruction, as both variables differed between the two cohorts.

The present study was conducted with the first cohort of elementary school students who had been taught English from Year 1 throughout the state of North Rhine-Westphalia, Germany. With the first and only chance to conduct this research at such a scale, but without any adjustment period for teachers, this naturalistic setting was unique and several confounding variables could not be taken into consideration. Teachers and schools, for example, had little to no experience teaching (1) foreign languages to younger students in Year 1 and (2) more experienced EFL students in Year 5 at grammar schools. We would encourage validating the outcomes of this study

through another round of data collections in Year 5 and 7 with another cohort of students.

Recent studies on students' proficiencies regularly take into account the variance that is explained by the clustered or hierarchical data structure, because effects on the levels of the classroom or the school that confound the individual effects violate the assumption of stochastical independence and may lead not only to misspecification of standard errors, but even to the amplification of the observed effects on the individual level (Lüdtke, Robitzsch, Trautwein & Kunter, 2009). That the clustered structure was not controlled for must be seen as a drawback. The IR modelling did not take this into account by adding a mixture component of the computation of the scores because the main research-question of the project did not focus on individual but on mean scores. For the SEM modelling the clustered structure of the data was not considered because the students were often re-assigned at Year 7, so that the respective classes differed between Year 5 and Year 7. Schools have reported several reasons for students changing classes in Year 7 or 6, for example the mandatory introduction of a second foreign language in Year 6, the beginning of Content and Language Integrated Learning (CLIL) streams in Year 7, or newly created focus classes for example geared towards the sciences or humanities. Here multiple-membership or cross classification models would have been necessary, which were deemed overly complex for the current research questions that mostly address structural questions.

8.1.Conclusion

The present study adds to existing research that cautions not to blindly believe the "myth" that the earlier language learning commences in life, the better the outcomes will be. In Germany, the current implementation of beginning EFL in Year 1 with 90 minutes or less of instruction time does not yield the expected results policy makers may have hoped for, at least not in the long run.

Extending EFL into elementary school did not produce the anticipated linguistic outcomes. Admittedly, our study was conducted immediately after the introduction of EFL into Year 1 and some teething troubles, e.g. the lack of qualified English teachers in elementary school or outdated school books in secondary school, may have been resolved by now. However, the amount of exposure in elementary or secondary school has not changed, disregarding established research (Krashen et al., 1979; Larson-Hall, 2008; Muñoz, 2008; Pfenninger, 2014a).

The field of early foreign language learning remains in dire need of empirical research particularly targeting teacher education, the transition from elementary to secondary education and the use of textbooks. For example, the impact elementary school teachers' language proficiency and overall foreign language didactic skills has on students' learning has only received little attention despite its significance (see (de Bot, 2014). Understanding the effect on EFL learning of transitioning from elementary to secondary school requires thorough representative longitudinal studies to optimize language proficiency outcomes long-term. .

Current results are reason for more caution, call for further investigation and, potentially early foreign language learning should be adjusted to take available research more into consideration. We see two potential options for a possible change in early foreign language education:

1. increase the amount of exposure from Year 1 onwards, possibly following immersive or content-based approaches
2. move EFL back into Year 3 or even 5, i.e. secondary education, with more lessons and thus an overall more intensive approach

If a more focused, intensive approach to early foreign language instruction yields better linguistic results, moving EFL back into Year 3 with four instead of two hours might be an option worth pursuing, particularly as level of motivation for L2 learning in elementary school has been shown to be high and attitudes positive. The available research should caution us not to assert that an earlier start into L2 education in elementary school by itself will render more proficient L2 speakers. Ultimately, all involved stakeholders have to ask themselves what they expect from foreign language teaching at elementary and preschool level; and what they can realistically expect from minimal foreign language input of one or two hours per week.

9. Notes

1 Based on 35 weeks of teaching out of 40 of a school year following guidelines by the Department for Education in North-Rhine Westphalia to account for project weeks, school trips, and other missed or canceled lessons (MSW - NRW, 2008)

References

Adams, R., & Wu, M. (2002). PISA 2000 technical report. Paris: OECD.

Adams, R. J., Wu, M. R., & Wilson, M. R. (2012). Conquest. Camberwell: ACER.

Barucki, H., Bliesener, U., Boerner Otfried, Boettger, H., Hoffmann, I.-B., Kierepka, A., …Schlueter, N. (2015). Der Lernstand im Englischunterricht am Ende von Klasse 4: Ergebnisse der BIG-Studie. München: Domino-Verl.

Baumert, J., Artelt, C., Ditton, H., Fend, H., Hasselhorn, M., Macher, I., …Trautwein, U. (2011). Expertenrat „Herkunft und Bildungserfolg“: Empfehlungen für Bildungspolitische Weichenstellungen in der Perspektive auf das Jahr 2020 (BW2020). Retrieved 09.15.2016. Retrieved from http://www.oekostation.de/docs/Expertenbericht_BaWue_online.pdf

Biedroń, A. (2011). Intelligence in Gifted L2 Learners. In M. Pawlak (Ed.), Extending the Boundaries of Research on Second Language Learning and Teaching (129-142). Berlin, Heidelberg: Springer Berlin Heidelberg.

BIG-Kreis. (2009). Fremdsprachenunterricht als Kontinuum: Der Übergang von der Grundschule in die weiterführenden Schulen. Fremdsprachenunterricht in der Grundschule. Munich: Domino Verlag Günther Brinek GmbH.

Birch, B. M. (2015). English L2 reading: Getting to the bottom (Third edition). ESL & applied linguistics professional series. New York: Taylor & Francis.

Birdsong, D. (2006). Age and Second Language Acquisition and Processing: A Selective Overview. Language Learning, 56, 9–49. doi:10.1111/j.1467-9922.2006.00353.x

Blair, C., & Raver, C. C. (2012). Child development in the context of adversity: experiential canalization of brain and behavior. The American Psychologist, 67(4), 309–318. doi:10.1037/a0027493

Börner, O., Engel, G., & Groot-Wilken, B. (Eds.). (2013). Hörverstehen, Leseverstehen, Sprechen: Diagnose und Förderung von sprachlichen Kompetenzen im Englischunterricht der Primarstufe. Münster: Waxmann.

Bos, W., Bonsen, M., Gröhlich, C., Guill, K., & Scharenberg, K. (2009). KESS 7: Skalenhandbuch zur Dokumentation der Erhebungsinstrumente. HANSE - Hamburger Schriften zur Qualität im Bildungswesen: Vol. 4. Münster: Waxmann.

Bourdieu, P., & Passeron, J. C. (1990). Reproduction in education, society and culture. Theory, culture & society. London: Sage.

Byrne, B. M. (2010). Structural equation modeling with AMOS: Basic concepts, applications and programming (2nd ed.). Multivariate Applications Series. New York, London: Routledge Academic.

Carroll, J. B., & Sapon, S. M. (1959). Modern Language Aptitude Test. New York: The Psychological Corporation.

Council of Europe. (2001). Common European Framework of Reference for Languages: Learning, teaching, assessment. Cambridge: Cambridge University Press.

Council of the European Union. (1997). Council Resolution on the early teaching of European Union languages. Retrieved May 18, 2016. Retrieved from http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:31998Y0103(01)&from=EN

Council of the European Union. (2002). Presidency Conclusions: Barcelona European Council. Retrieved May 05, 2016. Retrieved from http://www.consilium.europa.eu/ueDocs/cms_Data/docs/pressData/en/ec/71025.pdf

Courtney, L., Graham, S., Tonkyn, A., & Marinis, T. (2015). Individual differences in early language learning: A study of English learners of French. Applied Linguistics, 1-25. doi:10.1093/applin/amv071

Csapó, B., & Nikolov, M. (2009). The cognitive contribution to the development of proficiency in a foreign language. Learning and Individual Differences, 19(2), 209–218. doi:10.1016/j.lindif.2009.01.002

Cummins, J. (1976). The Influence of Bilingualism on Cognitive Growth: A Synthesis of Research Findings and Explanatory Hypotheses. Working Papers on Bilingualism, No. 9, 1–43.

Cummins, J. (1979). Linguistic Interdependence and the Educational Development of Bilingual Children. Review of Educational Research, 49(2), 222–251. doi:10.3102/00346543049002222

Cummins, J. (1981). Age on arrival and immigrant second language learning in Canada: A Reassessment1. Applied Linguistics, II(2), 132–149. doi:10.1093/applin/II.2.132

Curtain, H. (2000). Time as a factor in early start programmes. In J. Moon & M. Nikolov (Eds.), Research into teaching English to young learners. International perspectives an edited collection of selected papers from two conferences (pp. 87–120). Pecs, Hungary: University Press Pecs.

Dallinger, S. (2015). Die Wirksamkeit bilingualen Sachfachunterrichts. Selektionseffekte, Leistungsentwicklung und die Rolle der Sprachen im deutsch-englischen Geschichtsunterricht (PhD). Pädagogischen Hochschule Ludwigsburg, Ludwigsburg, Germany. Retrieved on 29.09.2016. Retrieved from https://phbl-opus.phlb.de/files/67/Dissertation_Sara+Dallinger.pdf

de Bot, K. (2014). The effectiveness of early foreign language learning in the Netherlands. Studies in Second Language Learning and Teaching, 3, 409–418. doi:10.14746/ssllt.2014.4.3.2

de Graaf, R., & Housen, A. (2009). Investigating effects and effectiveness of L2 instruction. In M. H. Long & C. Doughty (Eds.), Blackwell handbooks in linguistics. The handbook of language teaching (pp. 726–753). Chichester: Wiley-Blackwell.

DeKeyser, R., & Larson-Hall, J. (2005). What does the critical period really mean? In J. F. Kroll & A. M. B. Groot (Eds.), Handbook of bilingualism. Psycholinguistic Approaches (pp. 88–108). Oxford: Oxford Univ. Press.

Ditton, H., & Krüsken, J. (2006). Der Übergang von der Grundschule in die Sekundarstufe I. Zeitschrift für Erziehungswissenschaft, 9(3), 348–372. doi:10.1007/s11618-006-0055-7

Dlugosz, D. (2000). Rethinking the role of teaching a foreign language to young learners. ELT Journal, 54(3), 284–290. doi:10.1093/elt/54.3.284

Edelenbos, P., Johnstone, R., & Kubanek, A. (2006). Die wichtigsten pädagogischen Grundsätze für die fremdsprachliche Früherziehung: Sprachen für die Kinder Europas Forschungsveröffentlichungen, gute Praxis & zentrale Prinzipien. Endbericht der Studie EAC 89/04 (Lot 1). Retrieved May 05, 2016. Retrieved from http://ec.europa.eu/languages/policy/language-policy/documents/young_de.pdf

Education, Audiovisual & Culture Executive Agency. (2012). Key data on teaching languages at school in Europe. Brussels: EURYDICE.

Ehmke, T., & Jude, N. (2010). Soziale Herkunft und Kompetenzerwerb. In E. Klieme (Ed.), PISA 2009. Bilanz nach einem Jahrzehnt (pp. 231–254). Münster: Waxmann.

Ehrman, M. E., & Oxford, R. L. (1995). Cognition Plus: Correlates of Language Learning Success. The Modern Language Journal, 79(1), 67. doi:10.2307/329394

Ellis, N. C. (2011). Implicit and explicit SLA and their interface. In C. Sanz & R. P. Leow (Eds.), Georgetown University round table on languages and linguistics series. Implicit and explicit language learning. Conditions, processes, and knowledge in SLA and bilingualism (pp. 35–45). Washington, D.C: Georgetown University Press.

Ellis, R. (2009). Implicit and explicit learning, knowledge and instruction. In R. Ellis, L. Shawn, & E. Chatherine (Eds.), Second Language Acquisition. Implicit and explicit knowledge in second language learning, testing and teaching (pp. 3–25). Bristol: Multilingual Matters.

Enever, J. (Ed.). (2011). ELLiE: Early Language Learning in Europe: British Council.

Engel, G., & Ehlers, G. (2013). Hören – Zuhören – Verstehen. Möglichkeiten der Analyse, Diagnose und gezielten Förderung des Hörverstehens. In O. Börner, G. Engel, & B. Groot-Wilken (Eds.), Hörverstehen, Leseverstehen, Sprechen. Diagnose und Förderung von sprachlichen Kompetenzen im Englischunterricht der Primarstufe (pp. 44–69). Münster: Waxmann.

Euen, B. (2015). Intelligenz und kognitive Kompetenzen: Das Zusammenspiel von allgemeinen kognitiven Fähigkeiten und Schulleistungen in den Domänen Lesen, Mathematik und Naturwissenschaften am Ende der Grundschulzeit. Norderstedt: Books on Demand.

European Commission. (1995). Teaching and Learning: Towards the Learning Society. Retrieved May 05, 2016. Retrieved from http://europa.eu/documents/comm/white_papers/pdf/com95_590_en.pdf

European Commission. (2011). Language learning at pre-primary school level: Making it efficient and sustainable. A Policy Handbook. Brussels. Retrieved from http://ec.europa.eu/languages/policy/language-policy/documents/early-language-learning-handbook_en.pdf

Flege, J. E., & MacKay, I. R. A. (2011). What accounts for age effects on overall degree of foreign accent? In M. Wrembel, M. Kul, & K. Dziubalska-Kołaczyk (Eds.), Polish Studies in English Language and Literature: v. 31-32. Achievements and perspectives in SLA of speech. New sounds 2010 (pp. 65–82). Frankfurt am Main: Peter Lang.

Gardner, R. C., Tremblay, P. F., & Masgoret, A.-M. (1997). Towards a full model of second language learning: An empirical investigation. Modern Language Journal, 81(3), 344–362. doi:10.1111/j.1540-4781.1997.tb05495.x

Genesee, F. (1976). The role of intelligence in second language learning. Language Learning, 26(2), 267–280. doi:10.1111/j.1467–1770.1976.tb00277.x

Genesee, F., Geva, E., Dressler, C., & Kamil, M. (2006). Cross-linguistic relationships in working memory, phonological processes, and oral language. In D. August & T. Shanahan (Eds.), Developing literacy in second-language learners. Report of the national literacy panel on language minority children and youth (pp. 153–174). Mahwah, N.J., London: L. Erlbaum.

Georg, W. (2004). Cultural capital and social inequality in the life course. European Sociological Review, 20(4), 333–344. doi:10.1093/esr/jch028

Grabe, W., & Stoller, F. L. (2011). Teaching and researching reading (2nd ed.). Applied linguistics in action. Harlow: Longman.

Graham, S., Courtney, L., Tonkyn, A., & Marinis, T. (2016). Motivational trajectories for early language learning across the primary-secondary school transition. British Educational Research Journal, 42(4), 682–702. doi:10.1002/berj.3230

Granena, G. (2014). Language aptitude and long-term achievement in early childhood L2 Learners. Applied Linguistics, 35(4), 483–503. doi:10.1093/applin/amu013

Gräsel, C., Göbel, K., & Stark, R. (2007). Die Entwicklung von Lesekompetenz in der Sekundarstufe: Differentielle Analysen für Schülerinnen und Schüler mit unterschiedlichen Migrationserfahrungen. In O. Böhm-Kasper (Ed.), Kontexte von Bildung. Erweiterte Perspektiven in der Bildungsforschung (pp. 73–92). Münster, New York, München, Berlin: Waxmann.

Grotjahn, R. (2005). Je früher, desto besser? - Neuere Befunde zum Einfluss des Faktors "Alter" auf das Fremdsprachenlernen. In H. Pürschel & T. Tinnefeld (Eds.), Reihe: Fremdsprachen in Lehre und Forschung (FLF): Bd. 38. Moderner Fremdsprachenerwerb zwischen Interkulturalität und Multimedia. Reflexionen und Anregungen aus Wissenschaft und Praxis (pp. 186–202). Bochum: AKS-Verl.

Gustafsson, J.-E., Hansen, K. Y., & Rosén, M. (2013). Effects of home background on student achievement in reading, mathematics, and science at the fourth grade. In M. O. Martin & I. V. S. Mullis (Eds.), TIMSS and PIRLS 2011. Relationships among reading, mathematics, and science achievement at the fourth grade - implications for early learning (pp. 181–287). Boston: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Hartig, J., & Jude, N. (2008). Sprachkompetenzen von Mädchen und Jungen. In E. Klieme (Ed.), Beltz Pädagogik. Unterricht und Kompetenzerwerb in Deutsch und Englisch. Ergebnisse der DESI-Studie (pp. 202–206). Weinheim: Beltz.

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

Helmke, A., Helmke, T., Schrader, F.-W., Wagner, W., Nold, G., & Schröder, K. (2008). Die Videostudie des Englischunterrichts. In E. Klieme (Ed.), Beltz Pädagogik. Unterricht und Kompetenzerwerb in Deutsch und Englisch. Ergebnisse der DESI-Studie (pp. 345–363). Weinheim: Beltz.

Helmke, A., Schrader, F.-W., Wagner, W., Nold, G., & Schröder, K. (2008). Selbstkonzept, Motivation und Englischleistung. In E. Klieme (Ed.), Beltz Pädagogik. Unterricht und Kompetenzerwerb in Deutsch und Englisch. Ergebnisse der DESI-Studie (pp. 244–257). Weinheim: Beltz.

Hesse, H.-G., Göbel, K., & Hartig, J. (2008). Sprachliche Kompetenzen von mehrsprachigen Jugendlichen und Jugendlichen nicht-deutscher Erstsprache. In E. Klieme (Ed.), Beltz Pädagogik. Unterricht und Kompetenzerwerb in Deutsch und Englisch. Ergebnisse der DESI-Studie (pp. 208–230). Weinheim: Beltz.

Holtappels, H. G. (2006). Ganztagsschule: ein Beitrag zur Förderung und Chancengleichheit. In K. Höhmann & H. G. Holtappels (Eds.), Ganztagsschule gestalten. Konzeption, Praxis, Impulse (pp. 10–33). Seelze-Velber: Klett/Kallmeyer.

Housen, A., & Pierrard, M. (2005). Investigating instructed second language acquisition. In A. Housen & M. Pierrard (Eds.), Studies on Language Acquisition: Vol. 25. Investigations in instructed second language acquisition (pp. 1–27). Berlin, New York: Mouton de Gruyter.

Hu, L., & 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

IBM Corp. (2015a). AMOS. Armonk, NY.

IBM Corp. (2015b). SPSS. Armonk, NY.

Institut zur Qualitätssicherung im Bildungswesen. VERA – Ein Überblick. Retrieved September 19, 2016. Retrieved from https://www.iqb.hu-berlin.de/vera

Jæger, M. M. (2011). Does cultural capital really affect academic achievement?: New evidence from combined sibling and panel data. Sociology of Education, 84(4), 281–298. doi:10.1177/0038040711417010

Jaekel, N. (2015). Use and impact of language learning strategies on language proficiency. Investigating the impact of individual difference variables and participation in CLIL streams. Doctoral thesis (Doctoral Thesis). Ruhr-University of Bochum, Bochum. Retrieved on 28.10.2016. Retrieved from https://www.researchgate.net/publication/277014378_Use_and_impact_of_language_learning_strategies_on_language_proficiency_Investigating_the_impact_of_individual_difference_variables_and_participation_in_CLIL_streams

Johnstone, R. (2009). An early start: What are the key conditions for generalized success? In J. Enever, J. Moon, & U. Raman (Eds.), Young learner English language policy and implementation. International perspectives (pp. 31–41). Reading: Garnet Pub.

Kim, K. H., & Bentler, P. M. (2006). Data modeling: Structural equation modeling. In J. L. Green, G. Camilli, & P. B. Elmore (Eds.), Handbook of complementary methods in education research (pp. 161–175). Washington, D.C.: American Educational Research Association; Mahwah.

Kissau, S. (2006). Gender differences in second language motivation: An investigation of micro- and macro-level. Canadian Journal of Applied Linguistics, 9(1), 73–96.

Klieme, E. (Ed.). (2010). PISA 2009: Bilanz nach einem Jahrzehnt. Münster: Waxmann.

Kolb, A., & Mayer, N. (2010). Mehr Kontinuität! Englischkenntnisse aus der Grundschule weiterentwickeln. Der fremdsprachliche Unterricht Englisch, 103(44), 2–6.

Krashen, S. D., Long, M. A., & Scarcella, R. C. (1979). Age, rate and eventual attainment in second language acquisition. TESOL Quarterly, 13(4), 573. doi:10.2307/3586451

Larson-Hall, J. (2008). Weighing the benefits of studying a foreign language at a younger starting age in a minimal input situation. Second Language Research, 24(1), 35–63. doi:10.1177/0267658307082981

Lightbown, P., & Spada, N. M. (2006). How languages are learned (3rd ed.). Oxford handbooks for language teachers. Oxford: Oxford University Press.

Lindgren, E., & Muñoz, C. (2013). The influence of exposure, parents, and linguistic distance on young European learners' foreign language comprehension. International Journal of Multilingualism, 10(1), 105–129. doi:10.1080/14790718.2012.679275

Lüdtke, O., Robitzsch, A., Trautwein, U., & Köller, O. (2007). Umgang mit fehlenden Werten in der psychologischen Forschung. Psychologische Rundschau, 58(2), 103–117. doi:10.1026/0033-3042.58.2.103

Lüdtke, O., Robitzsch, A., Trautwein, U., & Kunter, M. (2009). Assessing the impact of learning environments: How to use student ratings of classroom or school characteristics in multilevel modeling. Contemporary Educational Psychology, 34(2), 120–131. doi:10.1016/j.cedpsych.2008.12.001

Marinova-Todd, S. H., Marshall, D. B., & Snow, C. E. (2000). Three misconceptions about age and L2 learning. TESOL Quarterly, 34(1), 9–34. doi:10.2307/3588095

May, P. (2007). Englisch Hörverstehen am Ende der Grundschulzeit. In W. Bos (Ed.), HANSE - Hamburger Schriften zur Qualität im Bildungswesen: Vol. 2. KESS 4 - Lehr- und Lernbedingungen in Hamburger Grundschulen. Münster: Waxmann.

Mihaljevic Djigunovic, J., & Lopriore, L. (2011). The learner: do individual differences matter? In J. Enever (Ed.), ELLiE. Early Language Learning in Europe (pp. 43–60). British Council.

Mihaljevic Djigunovic, J., Nikolov, M., & Otto, I. (2008). A comparative study of Croatian and Hungarian EFL students. Language Teaching Research, 12(3), 433–452. doi:10.1177/1362168808089926

Mills, N., Pajares, F., & Herron, C. (2007). Self-efficacy of college intermediate French students: Relation to achievement and motivation. Language Learning, 57(3), 417–442. doi:10.1111/j.1467-9922.2007.00421.x

MSW - NRW. (2008). Richtlinien und Lehrpläne für die Grundschule in Nordrhein-Westfalen: Deutsch, Sachunterricht, Mathematik, Englisch, Musik, Kunst, Sport, Evangelische Religionslehre, Katholische Religionslehre. Schule in NRW: Vol. 2012. Frechen: Ritterbach.

MSW - NRW. (2012). Englisch als Kontinuum – von der Grundschule zur weiterführenden Schule: Handreichung für den fortgeführten Englischunterricht in der Sekundarstufe I. Retrieved October 20, 2016. Retrieved from http://www.schulentwicklung.nrw.de/cms/upload/egs/Englisch_als_Kontinuum.pdf

MSW - NRW. (2015a). "Ganz In" - Mit Ganztag mehr Zukunft. Retrieved September 19, 2016. Retrieved from https://www.schulministerium.nrw.de/docs/Schulsystem/Ganztag/Gymnasium/Ganz_In/index.html

MSW - NRW. (2015b). Stundentafel für die Sekundarstufe I - Gymnasium. Retrieved September 19, 2016. Retrieved from https://www.schulministerium.nrw.de/docs/Schulsystem/Schulformen/Gymnasium/Sek-1/Stundentafel.pdf

Muñoz, C. (2006). Age and the rate of foreign language learning. Second Language Acquisition: Vol. 19. Clevedon: Multilingual Matters.

Muñoz, C. (2008). Age-related differences in foreign language learning. Revisiting the empirical evidence. IRAL - International Review of Applied Linguistics in Language Teaching, 46(3), 197–220. doi:10.1515/IRAL.2008.009

Muñoz, C. (2010). On how age affects foreign language learning. In A. Psaltou-Joycey & M. Mattheoudakis (Eds.), Advances in research on language acquisition and teaching. Selected papers (pp. 39–49). Thessaloniki: Greek Applied Linguistics Association.

Muñoz, V. (2007). Report of the Special Rapporteur on the right to education: Mission to Germany. Retrieved September 19, 2016. Retrieved from http://daccess-ods.un.org/access.nsf/Get?Open&DS=A/HRC/4/29/Add.3&Lang=E

Nicholas, H., & Lightbown, P. M. (2008). Defining child second language acquisition, defining roles for L2 instruction. In J. Philp, R. Oliver, & A. Mackey (Eds.), Language Learning & Language Teaching: v. 23. Second language acquisition and the younger learner. Child's play? (pp. 27–51). Amsterdam: John Benjamins.

Nikolov, M., & Djigunovic, J. M. (2006). Recent research on age, second language acquisition, and early foreign language learning. Annual Review of Applied Linguistics, 26, 234–260. doi:10.1017/S0267190506000122

Nikolov, M., & Mihaljević Djigunović, J. (2011). All shades of every color: An overview of early teaching and learning of foreign languages. Annual Review of Applied Linguistics, 31, 95–119. doi:10.1017/S0267190511000183

Nikolova, R. (2011). Englischleistungen und Einstellungen zum Englischunterricht. In U. Vielauf (Ed.), HANSE - Hamburger Schriften zur Qualität im Bildungswesen: Vol. 10. KESS 10/11. Kompetenzen und Einstellungen von Schülerinnen und Schülern an Hamburger Schulen am Ende der Sekundarstufe I und zu Beginn der gymnasialen Oberstufe (pp. 121–158). Münster, München, Berlin [u.a.]: Waxmann.

Nikolova, R., & Ivanov, S. (2010). Englischleistungen. In W. Bos (Ed.), HANSE - Hamburger Schriften zur Qualität im Bildungswesen: Bd. 6. KESS 8. Kompetenzen und Einstellungen von Schülerinnen und Schülern am Ende der Jahrgangsstufe 8 (pp. 49–66). Münster: Waxmann.

Nold, G., & Rossa, H. (2008). Sprechen Englisch [Speaking skills English]. In E. Klieme (Ed.), Beltz Pädagogik. Unterricht und Kompetenzerwerb in Deutsch und Englisch. Ergebnisse der DESI-Studie (pp. 170–179). Weinheim: Beltz.

OECD. (2015). The ABC of gender equality in education: Aptitude, behaviour, confidence: OECD Publishing.

Pfenninger, S. E. (2014a). The literacy factor in the optimal age discussion: A five-year longitudinal study. International Journal of Bilingual Education and Bilingualism, 19(3), 217–234. doi:10.1080/13670050.2014.972334

Pfenninger, S. E. (2014b). The misunderstood variable: Age effects as a function of type of instruction. Studies in Second Language Learning and Teaching, 3, 529–556. doi:10.14746/ssllt.2014.4.3.8

Pfenninger, S. E., & Singleton, D. (2016). Affect trumps age: A person-in-context relational view of age and motivation in SLA. Second Language Research, 32(3), 311–345. doi:10.1177/0267658315624476

Piske, T. (2013). Frühbeginn allein ist nicht genug.: Welchen Einfluss haben Faktoren wie Alter, sprachlicher Input, Geschlecht und Motivation auf die Aussprechentwicklung und die grammatischen Kenntnisse von Zweitsprachenlernern? In C. Bürgel & D. Siepmann (Eds.), Thema Sprache - Wissenschaft für den Unterricht: Vol. 6. Sprachwissenschaft - Fremdsprachendidaktik: neue Impulse (pp. 117–144). Baltmannsweiler: Schneider Verlag Hohengehren.

Rasch, G. (1980). Probabilistic models for some intelligence and attainment tests. Chicago: University of Chicago.

Rauch, D. (2014). Effects of biliteracy on third language reading proficiency, the example of Turkish-German bilinguals. In P. Grommes & A. Hu (Eds.), Hamburg Studies on Linguistic Diversity. Plurilingual Education (pp. 199–218). Amsterdam: John Benjamins Publishing Company.

Rauch, D. P., Naumann, J., & Jude, N. (2012). Metalinguistic awareness mediates effects of full biliteracy on third-language reading proficiency in Turkish-German bilinguals. International Journal of Bilingualism, 16(4), 402–418. doi:10.1177/1367006911425819

Robinson, P. (2003). The cognition hypothesis, task design, and adult task-based language learning. Second Language Studies, 21(2), 45–105.

Robinson, P. (2007). Task complexity, theory of mind, and intentional reasoning: Effects on L2 speech production, interaction, uptake and perceptions of task difficulty. IRAL - International Review of Applied Linguistics in Language Teaching, 45(3). doi:10.1515/IRAL.2007.009

Rost, J. (2004). Lehrbuch Testtheorie - Testkonstruktion (2nd ed.). Psychologie Lehrbuch. Bern: Huber.

Rost, M. (2011). Teaching and researching listening (2nd ed.). Applied linguistics in action. Harlow: Longman.

Rumlich, D. (2016). Evaluating bilingual education in Germany: CLIL students' general English proficiency, EFL self-concept and interest. Frankfurt am Main: Peter Lang.

Ruyffelaert, A., & Hadermann, P. (2012). The impact of age and gender on the learners' motivation and attitudes towards French in secondary education in Flanders. In L. Gómez Chova, A. López Martínez, & I. Candel Torres (Eds.), INTED 2012 publications. International Technology, Education and Development Conference : 6th edition, Valencia, Spain, 5th-7th March, 2012 (pp. 159–165). Valencia, Spain: IATED.

Sälzer, C., Reiss, K., Schiepe-Tiska, A., Prenzel, M., & Heinze, A. (2013). Zwischen Grundlagenwissen und Anwendungsbezug: Mathematische Kompetenz im internationalen Vergleich. In M. Prenzel, C. Sälzer, E. Klieme, & O. Köller (Eds.), PISA 2012. Fortschritte und Herausforderungen in Deutschland (pp. 47–98). Münster: Waxmann.

Saville-Troike, M. (2006). Introducing second language acquisition. Cambridge Introductions to Language and Linguistics. Cambridge, New York: Cambridge University Press.

Schmelter, L. (2010). (K)eine Frage des Alters - Fremdsprachenunterricht auf der Primarstufe. Zeitschrift für Interkulturellen Fremdsprachenunterricht, 15(1), 26–41.

Skehan, P. (1998). A cognitive approach to language learning. Oxford applied linguistics. Oxford: Oxford University Press.

Sparks, R. L., Ganschow, L., & Patton, J. (1995). Prediction of performance in first-year foreign language courses: Connections between native and foreign language learning. Journal of Educational Psychology, 87(4), 638–655. doi:10.1037/0022-0663.87.4.638

Sparks, R. L., Patton, J., & Ganschow, L. (2012). Profiles of more and less successful L2 learners: A cluster analysis study. Learning and Individual Differences, 22(4), 463–472. doi:10.1016/j.lindif.2012.03.009

Sparks, R. L., Patton, J., Ganschow, L., Humbach, N., & Javorsky, J. (2008). Early first-language reading and spelling skills predict later second-language reading and spelling skills. Journal of Educational Psychology, 100(1), 162–174. doi:10.1037/0022-0663.100.1.162

Standing Conference of the Ministers of Education. (2013). Fremdsprachen in der Grundschule – Sachstand und Konzeptionen 2013“. Retrieved September 19, 2016. Retrieved from https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2013/2013_10_17-Fremdsprachen-in-der-Grundschule.pdf

Steinlen, A. (2016). Primary school minority and majority language children in a partial immersion program. Journal of Immersion and Content-Based Language Education, 4(2), 198–224. doi:10.1075/jicb.4.2.03ste

Steinlen, A. K., Hâkansson, G., Housen, A., & Schelletter, C. (2010). Receptive L2 grammar knowledge development in bilingual preschools. In K. Kersten, A. Rohde, C. Schelletter, & A. K. Steinlen (Eds.), Bilingual Preschools: / ed. by Kristin Kersten; Andreas Rohde; Christina Schelletter; Anja K. Steinlen ; Vol. 1. Learning and development (pp. 69–100). Trier: WVT Wissenschaftlicher Verlag Trier.

Storck, J. (2015). Auswirkungen des Übergangs von der Grundschule in die Sekundarstufe I auf das Wohlbefinden und Selbstkonzept von Schülerinnen und Schülern. Reihe Studium und Forschung: H. 24. Kassel: Kassel Univ. Press.

Sunderland, J. (1994). Exploring gender: Questions and implications for English language education. New York, London: Prentice Hall.

Taylor, C., & Lafayette, R. (2010). Academic achievement through FLES: A case for promoting greater access to foreign language study among young learners. The Modern Language Journal, 94(1), 22–42. doi:10.1111/j.1540-4781.2009.00981.x

Thürmann, E. (2009). Anfänge, Übergänge und Perspektiven - Prognosen zu Weiterentwicklung des Englischunterrichts. In G. Engel, B. Groot-Wilken, & E. Thürmann (Eds.), Englisch in der Primarstufe - Chancen und Herausforderungen. Evaluation und Erfahrungen aus der Praxis (pp. 5–22). Berlin: Cornelsen.

Unsworth, S., Persson, L., Prins, T., & Bot, K. de. (2015). An investigation of factors affecting early foreign language learning in the Netherlands. Applied Linguistics, 527–548. doi:10.1093/applin/amt052

van der Slik, F. W. P., van Hout, R. W. N. W., & Schepens, J. J. (2015). The gender gap in second language acquisition: Gender differences in the acquisition of Dutch among immigrants from 88 countries with 49 mother tongues. PloS one, 10(11), e0142056. doi:10.1371/journal.pone.0142056

Vandergrift, L., & Goh, C. C. M. (2012). Teaching and learning second language listening: Metacognition in action. ESL & applied linguistics professional series. New York: Routledge.

Verband Bildung und Erziehung Landesverband NRW. (2006). Synopse: APO-SI aktuell gültige Fassung, 03.05.06 APO-SI Entwurfstext, 24.08.06. Retrieved October 20, 2016. Retrieved from http://www.vbe-nrw.de/downloads/PDF%20Dokumente/APO_SI_syn_pan_06.09.06.pdf

(2012, February 11). Düsseldorf, Germany: MSW - NRW.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. Psychometrika, 54(3), 427–450. doi:10.1007/BF02294627

Wendt, H., Stubbe, T. C., Schwippert, K., & Bos, W. (Eds.). (2015). 10 Jahre international vergleichende Schulleistungsforschung in der Grundschule: Vertiefende Analysen zu IGLU und TIMSS 2001 bis 2011. Münster: Waxmann.

Wilden, E., Porsch, R., & Ritter, M. (2013). Je früher desto besser?: Frühbeginnender Englischunterricht ab Klasse 1 oder 3 und seine Auswirkungen auf das Hör- und Leseverstehen. Zeitschrift für Fremdsprachenforschung, 24(2), 171–201.

Wolf, F. M. (1986). Meta-analysis: Quantitative methods for research synthesis. Quantitative applications in the Social Sciences, 0149-192X: no.07-059. Beverly Hills, London: Sage.

Wright, B. D. & Linacre, J. M. (1994). Reasonable mean-square fit values. Retrieved May 05, 2016. Retrieved from www.rasch.org/rmt/rmt83b.htm