Empirische Sonderpädagogik, 2015, Nr. 3, S. 175-193
ISSN 1869-4845 (Print) · ISSN 1869-4934 (Internet)

Überprüfung von Messinvarianz mittels CFA und DIF-Analysen

Susanne Schwab[1] *& Christoph Helm*[2]

[1] *Universität Bielefeld*
[2] *JKU Linz*

Zusammenfassung

Ziel der vorliegenden Studie ist es, die Bedeutung von Messinvarianz latenter Variablen beim Gruppenvergleich aufzuzeigen und anhand einer Beispielskala die einzelnen Schritte bei der Berechnung von Messinvarianz mit Mplus 7 und R zu veranschaulichen. Messinvarianz wird dabei sowohl vor dem Hintergrund des linearen, konfirmatorischen Faktorenansatzes als auch des nicht linearen, logistischen Item-Response-Theorie-basierten Differential Functioning Ansatzes (DIF) untersucht. Datengrundlage der Studie ist eine Stichprobe von 1037 Schülerinnen und Schülern der vierten und siebten Schulstufe aus drei österreichischen Bundesländern, wovon etwa 11% einen sonderpädagogischen Förderbedarf (SPF) aufweisen. Zur Erfassung von Einsamkeitserleben wurde eine deutsche Kurzversion der Illinois Loneliness and Social Satisfaction Scale (ILSS; Asher, Hymel & Renshaw, 1984; siehe dazu Schwab, 2015b) verwendet. Die angenommene eindimensionale Faktorenstruktur wurde für beide Gruppen (Schülerinnen und Schüler mit und ohne SPF) konfirmatorisch bestätigt. Sowohl die Prüfung der mehrgruppen-konfirmatorischen Faktorenanalysen (CFA) als auch die DIF-Analysen zeigen, dass die Skala für Schülerinnen und Schüler mit und ohne SPF messäquivalent ist. Konfigurale, metrische und skalare Messinvarianz liegen vor. Somit kann diese Skala verwendet werden, um latente Mittelwertvergleiche zwischen diesen beiden Gruppen durchzuführen. Dieser Vergleich belegt, dass Schülerinnen und Schüler mit SPF über eine signifikant höhere Einsamkeit berichten. Zusammenfassend wird auf die hohe Bedeutung von Messinvarianz für vergleichende Studien verwiesen.

Schlüsselwörter: Messinvarianz, Differential Functioning, Sonderpädagogischer Förderbedarf, Einsamkeit

Testing for Measurement Invariance in Students with and without Special Educational Needs – A case example using the Short Form of the Illinois Loneliness and Social Satisfaction Scale

Abstract

This study focuses on the relevance of measurement invariance for group comparisons of latent variables. A case example is used to illustrate the individual steps taken when analysing measurement invariance with Mplus 7 and R. Measurement invariance is examined within the framework of the linear Confirmatory Factor Analysis (CFA), as well as the Differential Item Functioning (DIF) approach stemming from the non-linear logistic item response theory. Participants are 1037 pupils attending 4th and 7th grade in three Austrian federal states. Some 11 % of the student were diagnosed as having special educational needs (SEN). A short German version of the Illinois Loneliness and Social Satisfaction Scale (ILSS; Asher, Hymel & Renshaw, 1984; see Schwab, 2015b) was used as latent construct. The assumed one-dimensional factor structure is confirmed for both groups (pupils with and without SEN). Both the examination of the CFA and the DIF-analyses support measurement equivalence of the scale for pupils with and without SEN. Configural, metric and scalar measurement invariance are confirmed. Consequently, the scale can be used to compare the means between both groups on the latent variable. This comparison shows that pupils with SEN report a significantly higher degree of loneliness. The present study shows the great importance of measurement invariance when groups are compared.

Keywords: measurement invariance, differential item functioning, special educational needs, loneliness

Bedeutung von Messinvarianz für latente Gruppenvergleiche

Zahlreiche Wissenschaftlerinnen und Wissenschaftler der empirischen Sonderpädagogik führen Gruppenvergleiche zwischen Schülerinnen und Schülern mit und ohne sonderpädagogischem Förderbedarf (SPF) durch. Es werden zumeist Gruppenvergleiche in Bezug auf die mittleren Ausprägungen von latenten Variablen, wie beispielsweise dem Selbstkonzept, durchgeführt. Diesbezüglich muss kritisch angemerkt werden, dass die verwendeten Messinstrumente oftmals nicht für Schülerinnen und Schüler mit SPF konzipiert wurden. Dabei können verschiedene Verzerrungen vorliegen. Temme und Hildebrandt (2008) zeigen an einem hypothetischen Beispiel auf, dass bei fehlender Messinvarianz Probanden einer Gruppe Items im Schnitt negativer einschätzen, als Probanden der Vergleichsgruppe, obwohl beide Gruppen über die gleiche Ausprägung der latenten Variable verfügen. Weiter argumentieren sie, dass bei fehlender Messäquivalenz auch die Strukturbeziehungen in einem Pfadmodell, das für beide Gruppen geprüft wird, verzerrt sein können.

In der Testtheorie wird grundsätzlich davon ausgegangen, dass sich der Messwert aus einem wahren Wert, also der tatsächlichen Ausprägung, und einem Messfehler zusammensetzt (z.B. Bühner, 2011). Dieser Messfehler ist ein Zufallsfehler mit dem Erwartungswert von Null. Zusätzlich kann ein Bias unter anderem aufgrund von Motivation oder persönliche Tagesverfassung resultieren. Messinvarianz ist von diesem Bias jedoch insofern zu unterscheiden, als sich diese auf die Vergleichbarkeit von Werten zwischen Gruppen bezieht und untersucht, ob ein latentes Konstrukt durch die verwendeten Items in den Gruppen auch äquivalent erfasst wird. „Gruppenspezifische Eigenschaften (z.B. der kulturelle Hintergrund) können aber die Reaktion der Befragten auf Fragebogenstatements und damit die Messbeziehungen so beeinflussen, dass die beobachteten Indikatoren die ‚wah-

ren' Gruppendifferenzen auf der Konstruktebene, insbesondere Mittelwertdifferenzen und Unterschiede in den strukturellen Beziehungen zwischen den latenten Variablen, nur verzerrt wiedergeben" (Temme & Hildebrandt, 2008, S. 1). Messinvarianz kann daher beispielsweise entstehen, wenn bei den Vergleichsgruppen ein unterschiedliches konzeptionelles Verständnis über ein Konstrukt vorherrscht, was insbesondere bei interkultureller Forschung vorkommt, weil ein und derselbe Begriff in verschiedenen Kulturen unterschiedlich konnotiert oder assoziiert ist (Chen, 2008). Im Rahmen der PISA-Studie 2009 wurden für Leseaufgaben zudem das Itemformat, das Format des vorgelegten, zu lesenden Textes, das kognitive Anspruchslevel der Aufgaben, die Textsorte (Narration, Exposition, Argumentation, ...) sowie der Kontext der Aufgabe (privat, beruflich, schulisch, ...) als nicht messinvariant für Kinder mit und ohne Migrationshintergrund identifiziert, wobei Kinder ohne Migrationshintergrund durch dieses Bias systematisch bevorzugt wurden (Dinis da Costa & Araújo, 2012). Um latente Mittelwerte zwischen verschiedenen Gruppen vergleichen zu können, gilt daher das Vorliegen von Messinvarianz als Voraussetzung (Christ & Schlüter, 2012). Bislang wurde der Messvarianz in der empirischen Sonderpädagogik allerdings kaum Beachtung geschenkt. Teilweise werden zwar Instrumente speziell für die Verwendung bei Schülerinnen und Schülern mit SPF erprobt, wie beim Fragebogen zur Erfassung von Dimensionen der Integration von Schülerinnen und Schülern (FDI) von Haeberlin, Moser, Bless und Klaghofer (1989) (siehe dazu Venetz, Zurbriggen & Eckhart, 2014), dennoch wird zumeist vor latenten Mittelwertvergleichen von Kindern mit und ohne SPF nicht überprüft, ob die Skalen über die Gruppen hinweg invariant messen. Zudem muss bei vielen Instrumenten, wie beispielsweise dem FDI angemerkt werden, dass diese für Kinder mit sprachlichen Schwierigkeiten oder Kinder mit SPF als zu umfangreich und auch hinsichtlich der sprachlichen Formulierungen als zu komplex eingeschätzt werden kann (Venetz et al., 2014). Gerade aber im sonderpädagogischen Forschungsbereich liegt der Gedanke nahe, dass Items bei Schülerinnen und Schülern mit und ohne SPF unterschiedlich funktionieren können. Soweit den Autoren bekannt ist, nahmen Kuhl, Weirich, Haag, Kocaj und Kroth (2013) erstmals eine Messinvarianzprüfung von Kompetenztests in den Fächern Deutsch und Mathematik für Schülerinnen und Schüler mit SPF vor. In Bezug auf Schülerinnen und Schüler mit unterschiedlichen schulischen Fähigkeiten wurde von Steinmetz, Schmidt, Tina-Booh, Schwartz und Wieczorek (2009) bereits bestätigt, dass Messäquivalenz ein Problem darstellen kann. Demzufolge lässt sich auch für Schülerinnen und Schüler mit und ohne SPF vermuten, dass Skalen für diese beiden Gruppen nicht immer gleich gut funktionieren (z.B. Bossaert & Petry, 2013; Nusser, Carstensen & Artelt, 2015). Das bedeutet, dass Schülerinnen und Schüler mit SPF bei gleicher latenter Ausprägung andere Antwortwahrscheinlichkeiten aufweisen. Allerdings gibt es auch Studien, die belegen, dass Instrumente wie die Kurzform des FDI (K-FDI 4-6; Venetz et al., 2014) oder eine Kurzversion der Chedoke-McMaster Attitudes towards Children with Handicaps Skala (Schwab, 2015c) bei Schülerinnen und Schülern mit und ohne SPF Messinvarianz besitzen.

Prüfung von Messinvarianz

Mit der Prüfung der Messinvarianz rücken drei Fragen in den Vordergrund:

(1) Passen die Messmodelle in den einzelnen Gruppen zu den Daten?
(2) Welche Modellparameter unterscheiden sich in den Gruppen? Hier interessiert meist die Ausprägung der latenten Variablen.
(3) Inwieweit ist die Strukturbeziehung zwischen latenten Variablen in den einzelnen Gruppen unterschiedlich?

Der vorliegende Beitrag beschränkt sich auf die ersten beiden Fragen. Zu (1): Im Rahmen der Analyse des Messmodells wird geprüft, ob ein Konstrukt in unterschiedlichen Gruppen bzw. Populationen (z.B. bei Mädchen und Buben, bei Schülerinnen und Schülern mit und ohne SPF) durch die verwendeten Items äquivalent repräsentiert wird. Konkreter: Es wird geprüft, ob die „Messbeziehung zwischen den beobachteten Indikatoren und den ihnen zugrunde liegenden latenten Variablen in den Gruppen gleich sind" (Temme & Hildebrandt, 2008, S 1). Es wird also geprüft, ob ein Messmodell in verschiedenen Subgruppen gültig ist (für eine Übersicht siehe Christ & Schlüter, 2012, S. 49ff.). Je nach Restriktivität der Annahmen über die Gleichheit von Modellparametern zwischen den untersuchten Gruppen werden mehrere Arten von Messinvarianz unterschieden: konfigurale, metrische, skalare, messfehlerbezogene (strikte) und vollständige Messinvarianz, welche hinsichtlich ihrer Restriktivität ansteigen und damit in hierarchischer Beziehung zueinander stehen (Brown, 2006). So ist beispielsweise die konfigurale Messinvarianz Voraussetzung für die strengeren Invarianzen.

Konfigurale Messinvarianz ist die am wenigsten restriktive Form der Messinvarianz und bezieht sich auf eine invariate, also äquivalente Faktorenstruktur. Konkret heißt das, dass in beiden Subgruppen (z.B. Schülerinnen und Schüler mit und ohne SPF) das gleiche Modell mit den gleichen Parametern geschätzt wird, diese aber unterschiedliche Werte annehmen dürfen (=„frei variieren"). Bei Vorliegen dieser Invarianz laden die manifesten Variablen auf einer oder mehreren identen latenten Variablen in beiden Populationen. Dies bedeutet, dass sich sowohl die Anzahl der Faktoren, als auch die Ladungsmuster in den Gruppen nicht signifikant voneinander unterscheiden.

Metrische Invarianz (auch schwache Invarianz genannt) ist gegenüber der konfiguralen Messinvarianz restriktiver, denn zusätzlich zur konfiguralen Invarianz werden nun auch die unstandardisierten Ladungen der manifesten Variablen über die Gruppen hinweg gleichgesetzt. Sowohl die Faktorenstruktur als auch die Faktorenladungen werden nun als äquivalent angenommen. Wird diese Annahme bestätigt, kann davon ausgegangen werden, dass in den untersuchten Subpopulationen die latenten Konstrukte die gleiche inhaltliche Bedeutung haben.

Skalare Invarianz (auch starke Messinvarianz genannt) geht über die konfigurale und metrische Invarianz hinaus und nimmt zusätzlich an, dass die Intercepts der manifesten Variablen über die Gruppen hinweg identisch sind. Das heißt, es müssen für die Stichproben sowohl invariante Faktorenstrukturen, invariante Faktorenladungen als auch invariante Intercepts (Regressionskonstanten) für die manifesten Variablen vorliegen. Bei Nachweis dieser Messinvarianz kann davon ausgegangen werden, dass keine itemspezifischen Schwierigkeitsunterschiede zwischen den Gruppen bestehen. Die Ausprägung in der latenten Variable kann zwischen den Gruppen verglichen werden.

Liegen alle der drei genannten Formen von Messinvarianz vor, so bezeichnet man dies auch als starke faktorielle Invarianz (Sass, 2011).

Strikte Invarianz (strenge Parallelität; Invarianz der Messfehler) bezeichnet die Prüfung der Annahme, dass in allen Subgruppen zusätzlich die Messfehlervarianzen gleich sind. Kann über die skalare Messinvarianz hinaus auch diese Invarianz nachgewiesen werden, so halten die Daten auch einer sogenannten strikten Invarianzprüfung stand. Kann keine strikte Invarianz nachgewiesen werden, so ist das ein Hinweis darauf, dass die Indikatoren über unterschiedliche Reliabilitäten in den Gruppen verfügen (Temme & Hildebrandt, 2008).

Vollständige Invarianz ist die stärkste Form der Invarianzprüfung, da alle Parameter des Modells (z.B. auch Strukturpfade zwischen latenten Variablen eines Strukturgleichungsmodells) über die Gruppen hinweg gleichgesetzt werden.

Im Rahmen empirischer Studien kommt es mit Zunahme der Strenge der Messinvarianz häufig zu einer Ablehnung der Äquivalenzhypothese. In diesen Fällen kann versucht werden, partielle Messinvarianz nachzuweisen: „Ist nur ein kleiner Teil der einem Faktor zugeordneten Indikatoren nichtinvariant, so werden diese z.B. bei weiterführenden Invarianztests oder Tests auf Unterschiede in den Mittelwerten der latenten Variablen zwar weiter einbezogen, durch die Schätzung gruppenspezifischer Parameter (d.h. Faktorladungen, Konstanten und Residualvarianzen) für diese Indikatoren wird aber die fehlende Messinvarianz kontrolliert“ (Temme & Hildebrandt, 2008, S. 19). Die Identifikation der nichtinvarianten Indikatoren erfolgt dabei meist explorativ (Krafft & Litfin, 2002).

In Längsschnittstudien kommt der Messinvarianzprüfung, vor allem wenn die Entwicklung im Hinblick auf eine latente Variable untersucht werden soll, große Bedeutung zu. In diesem Fall ist es notwendig sicherzustellen, dass zu den verschiedenen Messzeitpunkten das gleiche Konstrukt gemessen wird. Dabei geht man wie bei der Mehrgruppenanalyse vor, jedoch mit dem Unterschied, dass die Messzeitpunkte die Gruppen darstellen.

Messäquivalenz vor dem Hintergrund der Item Response Theory (IRT)

Die Multigroup Analysis/Mehrgruppenanalyse (MGA), wie sie hier thematisiert wird, stellt in den Sozialwissenschaften den gebräuchlichsten Ansatz zur Überprüfung von Messinvarianz dar (Temme & Hildebrandt, 2008). Der alternative IRT-Ansatz geht, im Gegensatz zum linearen CFA-Ansatz, von einem nicht linearen, logistischen Zusammenhang zwischen dem Antwortverhalten der Probanden und dem zu messenden latenten Konstrukt aus (für einen Vergleich der beiden Ansätze siehe Meade & Lautenschlager, 2004). Dabei werden je nach IRT-Modell unterschiedlich viele Parameter spezifiziert. Die sparsamste Variante, das Rasch-Modell für dichotome Items, enthält neben dem Parameter für die Personenfähigkeit lediglich einen weiteren Parameter für die Itemschwierigkeit. Eine grafische Invarianzprüfung kann daher erfolgen, indem man die Itemschwierigkeiten für zwei Personen-Subgruppen im Koordinatensystem auf der x- und y-Achse gegenüberstellt. Weichen die Koordinatenpunkte signifikant von der Diagonale ab, ist das ein Hinweis für fehlende Messinvarianz eines Items. In der vorliegenden Studie werden polytome Items mit fünfstufigem Antwortmuster analysiert. Dazu eignet sich das sogenannte Partial Credit Model nach Masters (Heine, 2014; Masters, 1982). Das Modell schätzt neben der Itemschwierigkeit weiter je Item vier Schwellenwerte (= Thresholds) für die fünf Antwortkategorien. Diese Parameter stehen im Fokus der folgenden Itemanalysen. Derartige Messinvarianzprüfungen auf Ebene der Antwortkategorien sind natürlich auch im MGA-Vorgehen möglich und dann sinnvoll, wenn man von einem kategorialen Messniveau der Indikatoritems ausgeht. Eine entsprechende Syntax ist bei Koziol (2010) zu finden.

Fragestellung

Ziel der Studie ist es, zu prüfen, ob ein latenter Mittelwertvergleich für die deutsche Kurzform der Illinois Loneliness and Social Satisfaction Scale (ILSS; Asher et al., 1984; siehe Schwab, 2015b) für Schülerinnen und Schüler mit und ohne SPF zulässig ist, und wenn ja, wie ausgeprägt dieser Unterschied ist. Diesbezüglich ist im ersten Schritt zu analysieren, ob die Skala reliabel ist und ob sich für Schülerinnen und Schüler mit und ohne SPF die verschiedenen Formen der Messinvarianz bestätigen lassen.

Methode

Stichprobe

Im Rahmen des ATIS-SI Projekts (Attitudes Towards Inclusion of Students with disabilities related to Social Inclusion; Schwab, 2015a) wurde eine Befragung bei Schülerinnen und Schülern in drei österreichischen Bundesländern (Steiermark, Niederösterreich, Burgenland) durchgeführt. Für den vorliegenden Artikel werden die Daten des zweiten Messzeitpunktes (Ende des Schuljahres 2013/14) verwendet. An der Erhebung haben insgesamt 1047 Schülerinnen und Schüler teilgenommen, welche aus 61 Schulklassen[1] stammten. Zur Teilnahme eingeladen wurden nur Schulen, welche in der entsprechenden Jahrgangsstufe auch Integrationsklassen führten. Die Schulen wurden telefonisch kontaktiert und gefragt, ob sie bereit wären, an der Studie teilzunehmen und ob eine entsprechende Integrationsklasse vorhanden sei. Die Regelklassen (in denen es keine Schülerinnen und Schüler mit SPF gibt) waren, wenn möglich, Parallelklassen in den gleichen Schulen. Aufgrund fehlender Werte wurden zehn Schülerinnen und Schüler (1%) aus dieser Studie ausgeschlossen. Die Stichprobe besteht daher aus 1037 Schülerinnen und Schülern (512 Jungen, 525 Mädchen) der vierten (39%) und siebten (61%) Schulstufe. 111 Schülerinnen und Schüler weisen einen SPF auf, welcher per Bescheid vom Bezirksschulrat erteilt wurde und sich zumeist (bei 77%) auf den Bereich „Lernen“ bezieht. Die unterschiedlichen Gruppengrößen stellen laut einer Simulationsstudie von Koh und Zumbo (2008) kein Problem für die Messinvarianzprüfung dar. Koh und Zumbo haben im Rahmen ihrer Simulationsstudie die Stichprobengrößen der Subgruppen variiert. Dabei konnte gezeigt werden, dass selbst ein Verhältnis von 200 : 800 den Type I Error in Bezug auf die Hypothese, dass starke und vollständige Messinvarianz vorliegt, nicht beeinflusste. Daher ist nicht davon auszugehen, dass das hier vorliegende Verhältnis von 111 : 926 das Untersuchungsergebnis beeinflusst.

Erhebungsinstrument

Um die von Schülerinnen und Schülern wahrgenommene Einsamkeit im Unterricht zu erfassen, wurde eine deutsche Kurzversion der Illinois Loneliness and Social Satisfaction Scale (ILSS; Asher et al., 1984; siehe dazu Schwab, 2015b) verwendet. Diese sogenannte “Pure-Loneliness-Scale” wurde bereits mehrfach erprobt (siehe z.B. Ladd, Kochenderfer & Coleman, 1996; Qualter et al., 2012; Schwab, 2015b) und umfasst insgesamt vier Items (siehe Tabelle 1), welche auf einer fünfstufigen Ratingskala von (1 = nie bis 5 = immer) zu beantworten sind. Schwab (2015b) überprüfte die Skalenreliabilität für Schülerinnen und Schüler mit und ohne SPF der vierten und der siebten Schulstufe und erhielt durchwegs zufriedenstellende Kennwerte (α = .71-.85). Grundsätzlich ist anzumerken, dass die meisten Probanden die Items eher im Bereich „nie bis selten“ angekreuzt haben. Für die Operationalisierung des SPF wurde auf das Vorhandensein eines SPF-Bescheids zurückgegriffen.

Statistische Analysen

Software. Alle statistischen Analysen werden in den Statistikpaketen Mplus 7 (Muthén & Muthén, 1998-2014) und R (R Core Team, 2014) durchgeführt. Dieses Vorgehen ist zwar redundant, da dieselben Ergebnisse erzielt werden, jedoch stellt das Paper auch einen Service-Beitrag dar, der interes-

[1] Es wird darauf verzichtet, die hierarchische Datenstruktur zu berücksichtigen, da eine Multilevel-Regressionsanalyse ohne Prädiktoren (Nullmodell) mit der z-standardisierten Einsamkeitsskala als Kriterium keine signifikante Varianz auf Klassenebene anzeigte (*ICC* = .001, *n.s.*). Zudem zeigte sich keine signifikante Varianz auf Klassenebene für die vier Nullmodelle der Einzelitems.

sierten Lesern die Syntax der frei zugänglichen Software R liefert (siehe Kästen 5-7 im Anhang), um die hier vorgestellten Ergebnisse nachzurechnen. Die dazu notwendigen Daten können auf Wunsch bei Susanne Schwab angefordert werden.

Confirmatory Factor Analysis (CFA). Als erster Analyseschritt wurde eine konfirmatorische Faktorenanalyse (CFA) der angenommenen Faktorenstruktur (Beziehung zwischen den manifesten Indikatoren und der latenten Variable) für die Gesamtstichprobe durchgeführt. Für die Beurteilung der Güte der CFA werden übliche Fit-Statistiken ($\chi^2/df \leq 3$, $RMSEA \leq .06$, $CFI \geq .95$, $TLI \geq .95$, $RMSEA \leq .08$, $SRMR \leq .08$, Hu & Bentler, 1999; Weiber & Mühlhaus, 2010) als deskriptive Indizien herangezogen. Die Prüfung auf Modellanpassung bzw. mangelnde Modellanpassung wird über die Ergebnisse des χ^2-Tests vorgenommen.

Parameterschätzung. Die am häufigsten verwendete Schätzmethode im Rahmen von konfirmatorischen Faktorenanalysen ist laut Bühner (2011) der Maximum Likelihood-Schätzer. Dieser setzt jedoch eine multivariate Normalverteilung der Items (hier die Messindikatoren der konfirmatorischen Faktorenanalyse) voraus. Mit dem R package „MVN" (Korkmaz, Goksuluk & Zararsiz, 2014) wurden die vier Items der Loneliness-Skala auf multivariate Normalverteilung geprüft. Sowohl die Signifikanztests (Marida: $\chi^2_{skew} = 28.376$, $p = .00$, $z_{kurt} = 82.288$, $p = .00$; Royston: $H = 804.259$, $p = .00$; Henze-Zirkler: $HZ = 352.592$, $p = .00$) als auch der hier nicht berichtete grafische Chi-Square Q-Q Plot zeigen eine deutliche Abweichung der Daten von der multivariaten Normalverteilung an. Aus diesem Grund musste eine Schätzmethode verwendet werden, die gegenüber dieser Voraussetzungsverletzung robust ist. Die Wahl des konkreten Schätzers ist von vielen Aspekten abhängig. Koziol (2010) nennt bspw. die Annahme über das Skalenniveau der Messungen, den Fokus auf Type-I- oder Type-II-Error und die einfachere Interpretation bestimmter Parameterschätzungen als Wahlkriterien. Simulationsstudien kommen bei der Frage nach der Wahl der Schätzmethode im Rahmen der Messinvarianzprüfung (Cheng-Hsien, 2014; Koziol, 2010) zwar zu dem Ergebnis, dass der Robust Weighted Least Square Schätzer (WLSMV-Schätzer) und der Robust Unweighted Least Square Schätzer (ULSMV-Schätzer) unter bestimmten Bedingungen (v.a. bei mehrdimensionalen Modellen und der Annahme kategorialer Messungen) dem Maximum Likelihood Robust-Schätzer (MLR-Schätzer) vorzuziehen sind, da diese zu exakteren Schätzungen der Faktorladungen und ihrer Messfehler kommen und zudem besser zur Bestimmung der Type-I-Error-Rate im Rahmen des χ^2-Modelltests geeignet sind. Demgegenüber besitzt der MLR-Schätzer eine höhere Effizienz und ermöglicht auch die Behandlung von Missings (unter der Missing at Random-Bedingungen). Des Weiteren kommt die Simulationsstudie von Koziol (2010) zu dem Ergebnis, dass der MLR-Schätzer den anderen gegenüber zu bevorzugen ist, wenn der Type-I-Error im Fokus der Analysen steht und von *kontinuierlichen* Messungen ausgegangen wird, was auf die vorliegende Studie zutrifft. Zudem ist der MLR, wie erwähnt, die am häufigsten verwendete Methode, weshalb das Ziel der Vergleichbarkeit der vorliegenden Ergebnisse mit anderen Studien ein weiteres Wahlkriterium darstellt. Aus diesen Gründen wird für die Folgeanalysen der MLR als Schätzmethode verwendet.

Messinvarianz-Prüfung. Für die anschließende Berechnung der Messinvarianz wurde der Step-Up Ansatz verwendet. Darunter ist zu verstehen, dass mit der am wenigsten restriktiven Form der Messinvarianz (konfigurale Messinvarinz) begonnen wird und die Modelle sukzessive restriktiver werden (z.B. Christ & Schlüter, 2012). Ob das jeweils restriktivere, genestete Modell passt, wird mittels χ^2-Differenztest geprüft. Der Unterschied im Modellfit wurde aufgrund des MLR-Schätzers über den Satorra-Bentler skalierten χ^2-Differenztest geprüft (z.B. Christ & Schlüter, 2012). Da der Anpas-

sungstest stichprobensensitiv ist und bei größeren Samples (wie dem vorliegenden) schon bei kleineren Modellverschlechterungen signifikant wird, werden die Modellvergleiche zudem auf Basis der rule of thumb nach Chen (2007; Cheung & Rensvold, 2002) beurteilt. Dabei gilt: Solange der *CFI* nicht um mehr als .02 Einheiten sinkt und der *RMSEA* nicht um mehr als .015 Einheiten steigt, können beide Modelle als, die Datenstruktur gleichgut widerspiegelnd, angesehen werden. Hierbei wird dem sparsameren Modell, das mit weniger Parametern auskommt, der Vorzug gegeben.

Differential Item Functioning (DIF). Die DIF-Analysen wurden auf Basis des Partial Credit Models nach Masters (1982; vgl. auch Andrich, 1978) durchgeführt. Dieses Modell schätzt neben einem Parameter für die Ausprägung des Personenmerkmals weiterhin einen „allgemeinen" Schwierigkeitsparameter pro Item und zusätzlich vier Schwellenwerte (=Thresholds) für die fünf Kategorien je Item. In Mplus wird die DIF-Analyse durch Gleichsetzen der Parameter über die Subgruppen hinweg durchgeführt (ähnlich der Messinvarianzprüfung). In R wird im Paket „pairwise" (Heine, 2014) der „allgemeine" Schwierigkeitsparameter zwischen den Gruppen durch einen graphischen Modelltest auf Invarianz geprüft. Alternativ können auch komplexere, mehrdimensionale IRT-Modelle im Paket „mirt" (Chalmers, 2012) auf Messinvarianz geprüft werden.

Ergebnisse

Itemkennwerte und Ergebnisse der Reliabilitätsanalyse

Die Mittelwerte der Subgruppen sind in Tabelle 1 abgetragen. Zudem zeigt sich bereits auf Basis der manifesten Mittelwerte, dass sich die Schülerinnen und Schüler mit SPF signifikant einsamer fühlen, als diejenigen ohne SPF (t_{1035} = -3.308, $p < .01$, Cohens d = .34). Die Ergebnisse der Reliabilitätsanalyse zeigen sowohl für Schülerinnen und Schüler ohne SPF (α = .80) als auch für jene mit SPF (α = .82) eine zufriedenstellende Reliabilität. Diese Reliabilitäten sind

Tabelle 1: Deskriptive Statistiken zur Pure-Loneliness-Scale

	Gesamt		**Schülerinnen und Schüler ohne SPF**		**Schülerinnen und Schüler mit SPF**		**Korrelationstabelle**			
	n = 1037		**n = 926**		**n = 111**					
	M	SD	M	SD	M	SD	Item 1	Item 2	Item 3	Item 4
Item 1	1.43	0.77	1.42	0.75	1.52	0.94	-			
Item 2	1.34	0.73	1.30	0.66	1.61	1.10	0.71	-		
Item 3	1.36	0.79	1.35	0.76	1.50	1.02	0.61	0.70	-	
Item 4	1.34	0.89	1.31	0.84	1.58	1.20	0.38	0.38	0.36	-
Item 1-4	**5.47**	**2.52**	**5.38**	**2.38**	**6.22**	**3.45**	**0.83**	**0.85**	**0.83**	**0.69**
α [CI]	.80 [.77-.84]		.80 [.76-.84]		.82 [.73-.91]					

Anmerkungen. N = Stichprobengröße, M = Mittelwert, SD = Standardabweichung, α = Cronbach's Alpha, CI = Konfidenzintervall (siehe dazu Fan & Thompson, 2001), alle Korrelationen sind statistisch signifikant $p < .01$. Item 1: In meiner Klasse bin ich einsam. Item 2: In meiner Klasse fühle ich mich alleine. Item 3: In meiner Klasse fühle ich mich ausgegrenzt. Item 4: In meiner Klasse habe ich niemanden, mit dem ich reden kann. Die Annahmen eines tau-äquivalenten Modells wurden in Mplus geprüft und bestätigt (χ^2 = 4.194, df = 4, p = .38), da diese Voraussetzung für das Cronbachs Alpha sind (Graham, 2006).

aufgrund der sich überlappenden Konfidenzintervalle nicht signifikant unterschiedlich voneinander, was auf Messfehlerinvarianz hindeutet. Die Korrelationen zeigen weiter, dass Item 4 „In meiner Klasse habe ich niemanden, mit dem ich reden kann" nicht so hoch mit den anderen drei Items korreliert, wie die übrigen Items der Skala untereinander. Das Item wird (1) aus inhaltlichen Gründen, (2) aus Gründen der Vergleichbarkeit mit anderen Studien, in denen ebenfalls alle vier Items verwendet wurden, sowie (3) der Tatsache, dass das Messmodell mit nur drei Items genau identifiziert wäre (was zu einem saturierten Modell führen und die Interpretation der Modellvergleiche erschweren würde), beibehalten.

Ergebnisse der CFA zur Pure-Loneliness-Scale

Die Berechnung der CFA, getrennt für den gesamten Datensatz sowie für die Daten der Schülerinnen und Schüler mit und ohne SPF, zeigt, dass das angenommene Messmodell in allen Samples gilt, wenn auch der *TLI* und *RMSEA* für das Sample der Schülerinnen und Schüler mit SPF auf eine leichte Abweichung des Modells von den Daten hinweisen (siehe die Modellfits in Abbildung 1). Alle anderen Kennwerte sprechen jedoch für eine gute Modellpassung. Bezüglich der Faktorladungen zeigt sich, wie auch schon in den deskriptiven Analysen, dass das Item 4 am geringsten auf dem latenten Faktor lädt, offenbar aber in der SPF-Gruppe noch am ehesten Ausdruck von Einsamkeit ist. Die hohen Faktorladungen deuten auf Eindimensionalität hin.

In Tabelle 2 sind die Prüfungen der hierarchischen Messinvarianzen über die Gruppen (kein SPF vs. SPF) hinweg vorzufinden. Im Rahmen des konfiguralen Invarianzmodells wurden alle Faktorladungen, Intercepts und Residualvarianzen über beide Subpopulationen hinweg frei geschätzt. Einzig der Mittelwert und die Varianz der latenten Variablen wurden auf 0 bzw. 1 fixiert. Dadurch ergeben sich 4 Freiheitsgrade: Insgesamt liegen 28 verfügbare Informationen vor – je Gruppe vier Mittelwerte, vier Varianzen und sechs Kovarianzen der vier Items. Für jede Gruppe sind vier Faktorladungen, vier Intercepts und vier Residualvarianzen zu schätzen, weshalb 24 der 28 Freiheitsgrade aufgebraucht werden. Die Fit-Statistiken zeigen, dass der Modellfit beim konfiguralen Invarianzmodell ange-

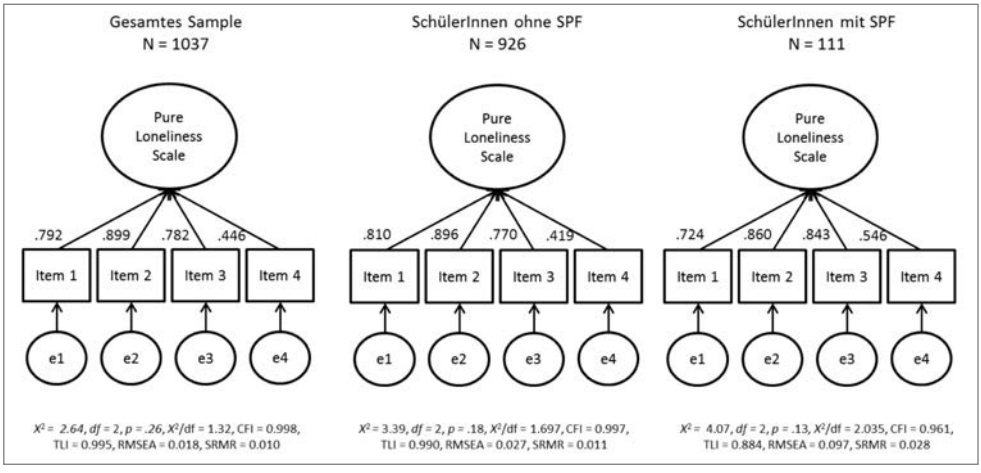

Abbildung 1: CFA für getrennte Samples

Anmerkungen. χ^2 = Chi-Square, *df* = degrees of freedom, *CFI* = comparative fit index, *TLI* = Tucker-Lewis Index, *RMSEA* = root mean square error of approximation.

nommen werden kann. Dies bedeutet, dass der Faktor in beiden Gruppen in ähnlicher Weise konzeptualisiert ist. Auch die Gleichheitsrestriktionen auf den Faktorladungen, die im Rahmen der metrischen Invarianz gesetzt wurden und wodurch drei Freiheitsgrade (die Varianz der latenten Skala in der SPF-Gruppe wurde freigesetzt) gewonnen wurden, verschlechtern die Modellpassung nicht bedeutsam ($\Delta CFI \leq .|.02|$, $\Delta RMSEA \leq .015$). Insofern gilt, dass die Maßeinheit der Skala für Schülerinnen und Schüler mit und ohne SPF identisch ist (metrische Invarianz). Auch nach Gleichsetzung der Intercepts und Freisetzung des Mittelwerts der latenten Skala in der SPF-Gruppe (wodurch weitere drei Freiheitsgrade gewonnen wurden) wurde keine substantielle Verschlechterung der Modellanpassung festgestellt (skalare Messinvarianz). Die *p*-values zeigen an, dass die Erhöhung der Modellabweichung in Relation zu den gewonnen Freiheitsgraden statistisch nicht signifikant ist. Stützt man die Modellevaluation allerdings auf den χ^2-Test, so wird mit einem *p*-value von .03 deutlich, dass das Modell an sich die empirische Kovarianzmatrix nicht widerspiegelt. Da aber die übrigen Fitwerte auf einen akzeptablen Modellfit hinweisen, gehen wir weiterhin von skalarer Messinvarianz aus. Die unterschiedlichen Größen der Subsamples nehmen keinen Einfluss auf die Ergebnisse. Somit kann bereits an dieser Stelle der Prüfung eine vollständige Messinvarianz für die Skala über die Gruppen (kein SPF vs. SPF) hinweg angenommen werden. Insofern sind die Voraussetzungen für einen sinnvollen Vergleich zwischen den latenten Mittelwertunterschieden in dieser Skala gegeben. Die Gruppe der Schülerinnen und Schüler mit SPF verfügt über eine latente mittlere Ausprägung auf der Skala „Pure-Loneliness-Scale“, die um 0.24 Standardabweichungen höher liegt als die latente mittlere Ausprägung der Gruppe der Schülerinnen und Schüler ohne SPF. Dieser Unterschied ist statistisch signifikant (*S.E.* = .086, p = .01). Somit kann festgehalten werden, dass sich Schülerinnen und Schüler mit SPF einsamer fühlen als diejenigen ohne SPF.

Führt man die Messinvarianzprüfung weiter, dann kann für die „Pure-Loneliness-Scale“-Skala weder strikte Invarianz (die Gleichsetzung der Messfehler über die beiden Gruppen hinweg führt zu weiteren vier Freiheitsgraden) noch vollständige Invarianz (die Gleichsetzung des latenten Mittelwerts über die Gruppen hinweg führt zu einem zusätzlichen Freiheitsgrad) nachgewiesen werden, wie Tabelle 2 zeigt. Die Modellverschlechterungen erreichen hier ein signifikantes Ausmaß.

Als nächstes wird untersucht, ob die strikte Invarianz zumindest partiell erreicht werden kann. Tabelle 3 zeigt den Modellvergleich für den Fall, in dem die Messfehlerparameter des ersten und zweiten Items in beiden Gruppen frei geschätzt werden. Die anderen beiden Messfehlerparameter von Item 3 und 4 werden über die beiden Gruppen hinweg gleichgesetzt, weshalb anstatt der in Tabelle 2 zur strikten Invarianz angeführten 14 Freiheitsgrade hier nur 12 vorliegen. Die Teststatistiken des Modellvergleichs zeigen, dass partielle Messinvarianz erreicht werden konnte.

Differential Item Functioning-Analysen vor dem Hintergrund der Item Response Theory

Da die „Pure-Loneliness-Scale“ fünf Antwortkategorien verwendet, existieren vier Schwellenwerte, die Informationen darüber geben, ab welcher Ausprägung auf der latenten Variable eine Person von einer niedrigeren Kategorie in eine höhere überwechselt. Wenn die Items für Schülerinnen und Schüler mit und ohne SPF gleich „funktionieren“ sollen, dann geht man von der Annahme aus, dass diese Schwellenwerte für beide Gruppen auf der latenten Skala gleich lokalisiert (also gleich schwierig) sind (M2). Der Modellvergleich in Tabelle 4 bestätigt diese Annahme bzw. diese Form der Messinvarianz. Die Modellverschlechterung ist nur marginal.

Tabelle 2: Prüfung der Messinvarianz für Schülerinnen und Schüler mit und ohne SPF (N = 926:111)

Modell	χ^2	p	Df	χ^2/df	CFI	RMSEA	RMSEA [CI]	ΔCFI	ΔRMSEA	TRd/Δdf	p
Konfigurale Invarianz	7.41	.12	4	1.85	.993	.041	[.000-.086]	-	-	-	-
Metrische Invarianz	13.40	.06	7	1.91	.987	.042	[.000-.076]	-.006	.001	5.99(3)	.11
Skalare Invarianz	20.10	.03	10	2.01	.979	.044	[.014-.072]	-.008	.002	7.05(3)	.07
Strikte Invarianz	62.12	.00	14	4.44	.901	.081	[.061-.103]	-.078	.037	27.66(4)	.00
"volle" Invarianz	68.79	.00	15	4.59	.890	.083	[.061-.103]	-.011	.002	17.34(1)	.00

Anmerkungen. χ^2 = Chi-Square, *df* = degrees of freedom, *CFI* = comparative fit index, *RMSEA* = root mean square error of approximation, *CI* = Konfidenzintervall (siehe dazu Fan & Thompson, 2001), *p*-values are calculated based the Satorra-Bentler scaled chi-square difference test (Satorra & Bentler, 1999). Im Rahmen der Prüfung der vollständigen Invarianz wurde die Varianz der latenten Variablen in der non-SPF-Gruppe auf 1 fixiert und in der SPF-Gruppe frei geschätzt, da die freie Schätzung dieser Parameter der Voreinstellung im korrespondierenden R-Befehl entspricht.

Tabelle 3: Partielle strikte Invarianz (N = 926:111)

Modell	χ^2	p	Df	χ^2/df	CFI	RMSEA	[CI]	ΔCFI	ΔRMSEA	TRd/Δdf	p
Skalare Invarianz	20.10	.03	10	2.01	.979	.044	[.014-.072]	-	-	-	-
Partielle strikte Invarianz	25.74	.01	12	2.15	.972	.047	[.021-.072]	-.007	.003	4.81(2)	.09

Anmerkungen. χ^2 = Chi-Square, *df* = degrees of freedom, *CFI* = comparative fit index, *RMSEA* = root mean square error of approximation, *CI* = Konfidenzintervall (siehe dazu Fan & Thompson, 2001), *p*-values are calculated based the Satorra-Bentler scaled chi-square difference test (Satorra & Bentler, 1999).

Tabelle 4: Invarianz der Schwellenwerte (Thresholds)

Modell	χ^2	p	Df	χ^2/df	CFI	RMSEA	[CI]	ΔCFI	ΔRMSEA	Value/Δdf	p
M1: frei geschätzte Schwellenwerte	8.80	.07	4	2.20	.999	.048	[.000-.092]	-	-	-	-
M2: fixierte Schwellenwerte	13.26	.65	16	0.83	1.000	.000	[.000-.034]	n.a.	n.a.	7.67(12)	.81

Anmerkungen. χ^2 = Chi-Square, *df* = degrees of freedom, *CFI* = comparative fit index, *RMSEA* = root mean square error of approximation. *CI* = Konfidenzintervall (siehe dazu Fan & Thompson, 2001), *n.a.* = nicht anwendbar. Der *p*-value wurde mit der DIFFTEST-Option für WLSMV-Schätzer berechnet (Muthén & Muthén, 1998-2010). Um M1 zu identifizieren wurde der jeweils erste Schwellenwert pro Item über die Gruppen hinweg gleichgesetzt, wodurch 4 Freiheitsgrade gewonnen werden konnten.

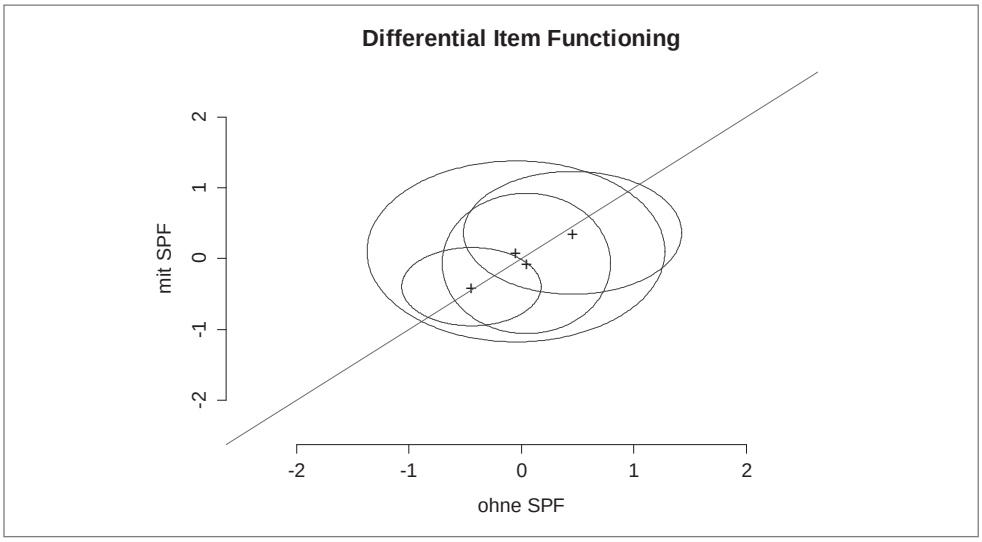

Abbildung 2: DIF-Analysen für die vier „Pure-Loneliness-Scale"-Items

Anmerkungen. Die x und y-Achse bilden die latente Metrik ab, auf der die Itemschwierigkeiten je Subgruppe eingetragen sind

Auch im Hinblick auf die geschätzte „allgemeine" Schwierigkeit der Items zeigt Abbildung 2, dass die Lokation der vier Items auf der latenten Skala für beide Subgruppen annähernd gleich ist. Dies zeigen die Ellipsen an, die das 95%-Konfidenzintervall wiedergeben. Alle Ellipsen überschneiden sich mit der Diagonale. Allerdings ist hier kritisch anzumerken, dass die Größe der Ellipsen von der Stichprobengröße beeinflusst ist. In Summe kann aber auch auf Basis der IRT-basierten DIF-Analysen festgehalten werden, dass die Skalen über zufriedenstellende Messinvarianz verfügen.

Diskussion

Es ist offenkundig, dass die empirische Bildungsforschung auf Basis von Stichproben, die sich aus verschiedenen Schülergruppen (z.B. Gymnasium, Realschule, Hauptschule) mit unterschiedlichen Merkmalen (z.B. Geschlecht, kognitive Unterschiede) zusammensetzen, Aussagen über eben erwähnte Merkmale machen will. Im Forschungsfeld der Inklusions- und Sonderpädagogik rückt das Merkmal „Sonderpädagogischer Förderbedarf" (SPF) ins Zentrum der Analysen. Um diese verschiedenen Varianten von Heterogenität in quantitativen Studien adäquat zu berücksichtigen, ist u.a. sicherzustellen, dass die eingesetzten Messinstrumente für verschiedene Probandengruppen mit unterschiedlichen Merkmalen geeignet sind. Damit ist gemeint, dass Test- oder Fragebogenitems von verschiedenen Schülerinnen und Schülern gleich interpretiert werden müssen bzw. dass nicht eine Subgruppe systematisch bevorzugt, oder benachteiligt wird. Damit ist das Vorliegen von Messinvarianz bzw. Differential Item Functioning angesprochen.

Im Rahmen des vorliegenden Beitrags wurde demonstriert, wie verschieden restriktive Formen der Messinvarianz vor dem Hintergrund der linearen, konfirmatorischen Faktorenanalyse einerseits und der (logistischen) Item-Response-Theory andererseits geprüft werden können. Wenngleich das Vorliegen von Messäquivalenz eine Voraussetzung für die sinnvolle Interpretation von latenten Mittelwertvergleichen ist, so wird in der aktuellen SPF-bezo-

genen Forschung zumeist darauf verzichtet. Dabei laufen Forscher oftmals Gefahr, unzulässige Gruppenvergleiche durchzuführen bzw. verzerrte Ergebnisse zu berichten. Insbesondere bei Schülerinnen und Schülern mit unterschiedlichen kognitiven Voraussetzungen scheint diesem Aspekt eine hohe Bedeutung zuzukommen. Oftmals wird per se davon ausgegangen, dass alle Items einer Skala bei Schülerinnen und Schülern mit und ohne SPF in der gleichen Beziehung zum latenten Konstrukt stehen, ohne dies zu prüfen. Werden Konstrukte in den Gruppen auf unterschiedliche Weise gemessen, so kann dies zu verzerrten Gruppenvergleichen führen.

Im diesem Beitrag konnte an einer Stichprobe von 1037 Schülerinnen und Schülern (111 mit SPF) gezeigt werden, dass die „Pure-Loneliness-Scale“ über zumindest skalare Messinvarianz und partielle strikte Invarianz verfügt, sodass Aussagen über Mittelwertvergleiche auf Basis der latenten Skala „Einsamkeit“ zwischen SPF- und Nicht-SPF-Schülerinnen und Schülern getroffen werden können. Die Ergebnisse verweisen darauf, dass SPF-Schülerinnen und Schüler über höhere Einsamkeit in der Schule berichten. Dies steht im Einklang mit den Ergebnissen bestehender Forschung (z.B. Bossaert, Colpin, Pijl & Petry, 2012; Pijl, Skaalvik & Skaalvik, 2010; Schwab, 2015b). Weitere DIF-Analysen zeigen zudem, dass die Schwellenwerte, sowie die Itemschwierigkeiten für beide Subgruppen nicht signifikant unterschiedlich sind, d.h. dass die Items der Skala für beide Subgruppen „gleich funktionieren“.

Für künftige empirische Vergleiche von Kindern mit und ohne SPF wäre es wünschenswert, wenn vermehrt Skalen diesen kritischen Analysen unterzogen würden, sodass ein Pool an Skalen verfügbar wird, der für Erhebungen in dieser Population geeignet ist.

Literaturverzeichnis

Andrich, D. (1978). A rating scale formulation for ordered response categories. *Psychometrika, 43,* 561-573.

Asher, S. R., Hymel, S. & Renshaw, P.D. (1984). Loneliness in children. *Child Development, 55*, 1456-1464.

Bossaert, G., Colpin, H., Pijl, S. J. & Petry, K. (2012). Loneliness among students with special educational needs in mainstream seventh grade. *Research in Developmental Disabilities, 33,* 1888-1897.

Bossaert, G. & Petry, K. (2013). Factorial validity of the Chedoke-McMaster Attitudes towards Children with Handicaps Scale (CATCH). *Research in Developmental Disabilities, 34*, 1336-1345.

Brown, T. A. (2006). *Confirmatory Factor Analysis*. New York: Guilford.

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

Chalmers, R. P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software, 48*(6), 1-29. Verfügbar unter: http://www.jstatsoft.org/v48I06/ [28.10.2014]

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling, 14,* 464-504.

Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of Personality and Social Psychology, 95,* 1005-1018.

Cheng-Hsien, L. (2014). *The performance of MLR, USLMV, and WLSMV estimation in structural regression model with ordinal variables.* Dissertation. Michigan State University.

Cheung, G. W. & Rensvold, R. B. (2002). Evaluating Goodness-of-Fit Indexes for Testing Measurement Invariance. *Structural Equation Modeling, 9,* 233-255.

Christ, O. & Schlüter, E. (2012). *Strukturgleichungsmodelle mit Mplus – Eine praktische Einführung*. München: Oldenbourg.

Dinis da Costa, P. & Araújo, L. (2012). *Differential item functioning (DIF): What functions differently for immigrant students in PISA 2009 reading items?* Luxemburg: Publications Office of the European Union.

Fan, X. & Thompson, B. (2001). Confidence Intervals about Score Reliability Coefficients, Please. *Educational and Psychological Measurement, 61*(4), 517-531.

Graham, J. M. (2006). Congeneric and (Essentially) Tau-Equivalent Estimates of Score Reliability. *Educational and Psychological measurement, 66*(6), 930-944.

Haeberlin, U., Moser, U., Bless, G. & Klaghofer, R. (1989). *Integration in die Schulklasse. Fragebogen zur Erfassung von Dimensionen der Integration von Schülern FDI 4-6*. Bern: Haupt.

Heine, J.-H. (2014). *pairwise: Rasch Model Parameters by Pairwise Algorithm. R package version 0.2.5*. Verfügbar unter: http://CRAN.R-project.org/package=pairwise [28.10.2014]

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.

Koh, K. & Zumbo, B. (2008). Multi-group confirmatory factor analysis for testing measurement invariance in mixed item format data. *Journal of Modern Applied Statistical Methods, 7(2),* 471-477.

Korkmaz, S., Goksuluk, D. & Zararsiz, G. (2014). MVN: An R Package for Assessing Multivariate Normality. *The R Journal, 6*(2), 151-162.

Koziol, N.A. (2010). *Evaluating Measurement Invariance with Censored Ordinal Data: A Monte Carlo Comparison of Alternative Model Estimators and Scales of Measurement*. Dissertation. University of Nebraska – Lincoln.

Krafft, M. & Litfin, T. (2002). Adoption innovativer Telekommunikationsdienste: Validierung der Rogers-Kriterien bei Vorliegen potenziell heterogener Gruppen. *Zeitschrift für betriebswirtschaftliche Forschung, 54*(2), 64-83.

Kuhl, P., Weirich, S., Haag, N., Kocaj A. & Kroth, A. (2013). *Zur Validität und Messinvarianz bei der Erfassung der Kompetenzen von Schülerinnen und Schüler und Schülern mit sonderpädagogischem Förderbedarf in Large-Scale-Assessments*. Vortrag an der 1. Tagung der GEBF in Kiel von 11. bis 13. März 2013.

Ladd, G. W., Kochenderfer, B. J. & Coleman, C. C. (1996). Friendship quality as a predictor of young children's early school adjustment. *Child Development, 67,* 1103-1118.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149-174.

Meade, A. W. & Lautenschlager, G. J. (2004, April). *Same Question, Different Answers: CFA and Two IRT Approaches to Measurement Invariance*. Symposium presented at the 19th Annual Conference of the Society for Industrial and Organizational Psychology, Chicago, IL. Verfügbar unter: http://www4.ncsu.edu/~awmeade/Links/Papers/CFA%26IRT_MEI%28SIOP04%29.pdf [28.10.2014]

Muthén, B. O. & Muthén, L. K. (1998-2014). Mplus (Version 7) [Computer Software]. Los Angles, CA.

Muthén, L. K. & Muthén, B. O. (1998-2010). *Mplus User's Guide. Sixth Edition*. Los Angeles, CA: Muthén & Muthén.

Nusser, Carstensen, C. H. & Artelt, C. (2015). Befragung von Schülerinnen und Schüler und Schülern mit sonderpädagogischem Förderbedarf Lernen: Ergebnisse zur Messinvarianz. *Empirische Sonderpädagogik, 2,* 99-116.

Pijl, S. J., Skaalvik, E. M. & Skaalvik, S. (2010). Students with Special needs and the Composition of Their Peer Group. *Irish Educational Studies, 29,* 57-70.

Qualter, P., Rotenberg, K. J. Barrett, L., Henzi, P., Barlow, A., Stylianou, M. & Harris, R.A. (2012). Investigating Hypervigilance

for Social Threat of Lonely Children. *Journal of Abnormal Child Psychology, 41,* 325-338.

R Core Team. (2014). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org [23.10.2014]

Sass, D. A. (2011). Testing Measurement Invariance and Comparing Latent Factor Means Within a Confirmatory Factor Analysis Framework. *Journal of Psychoeducational Assessment, 29*(4), 347-363.

Satorra, A. & Bentler, P .M. (1999). *A scaled difference chi-square test statistic for moment structure analysis. Economics Working Papers 412*. Department of Economics and Business, Universitat Pompeu Fabra.

Schwab, S. (2015a). Einstellung zur Integration im Zusammenhang mit sozialer Inklusion – Eine Fragebogenerhebung in österreichischen Integrations- und Regelschulklassen. *Vierteljahresschrift für Heilpädagogik und ihre Nachbargebiete, 84*, 66-67.

Schwab, S. (2015b). Evaluation of a Short-version of the Illinois Loneliness and Social Satisfaction Scale in a Sample of Students with Special Educational Needs – An Empirical Study with Primary and Secondary Students in Austria. *British Journal of Special Education*. DOI: 10.1111/1467-8578.12089

Schwab, S. (2015c). Einflussfaktoren auf die Einstellung von Schülerinnen und Schüler gegenüber Peers mit unterschiedlichen Behinderungen. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 47*(3), 1-11.

Steinmetz, H., Schmidt, P., Tina-Booh, A., Schwartz, S. H. & Wieczorek, S. (2009). Testing Measurement Invariance using Multigroup CFA: Differences between Educational Groups in Human Values Measurement. *Quality and Quantity, 43*, 599-616.

Temme, D. & Hildebrandt, L. (2008). Gruppenvergleiche bei hypothetischen Konstrukten – Die Prüfung der Übereinstimmung von Messmodellen mit der Strukturgleichungsmethodik. In *Schriftenreihe Economic Risk SFB 649 Papers, Discussion Paper 2008-042*. Verfügbar unter: http://edoc.hu-berlin.de/series/sfb-649-papers/2008-42/PDF/42.pdf [28.10.2014]

Venetz, M., Zurbriggen, C. & Eckhart, M. (2014). Entwicklung und erste Validierung einer Kurzversion des „Fragebogens zur Erfassung von Dimensionen der Integration von Schülerinnen und Schüler (FDI 4-6)" von Haeberlin, Moser, Bless und Klaghofer. *Empirische Sonderpädagogik, 2*, 99-113.

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

Dr. Susanne Schwab
Vertretungsprofessur für Erziehungswissenschaft mit dem Schwerpunkt Didaktik und Schulentwicklung im Kontext von Inklusion
Fakultät für Erziehungswissenschaft / AG 5 - Schulpädagogik und Allgemeine Didaktik
Universität Bielefeld
Universitätsstraße 25
Q1-108
33615 Bielefeld
susanne.schwab@uni-bielefeld.de

Dr. Christoph Helm
Vertretungsprofessur für Wirtschaftspädagogik
Otto-Friedrich-Universität Bamberg
Kärntenstraße 7
96052 Bamberg
christoph.helm@jku.at

Anhang

Kasten 1: Ausschnitt aus dem Mplus-Input zur Durchführung einer CFA, getrennt für Schülerinnen und Schüler mit und ohne SPF

```
TITLE: CFA
DATA: FILE IS Einsamkeit.dat;
VARIABLE:NAMES ARE item1 item2 item3 item4 SPF;

USEOBSERVATIONS =  SPF EQ 1;
!hier wird die Stichprobe der Kinder mit SPF verwendet
!für jene ohne SPF müsste man „SPF EQ 0;“ setzen
USEVARIABLES ARE  item1 item2 item3 item4;

ANALYSIS: ESTIMATOR IS WLS;

MODEL: Einsam by  item1 item2 item3 item4; Einsam@1;

OUTPUT:
MODINDICES STANDARDIZED;
```

Anmerkung: ! bezieht sich auf Kommentare, die lediglich der Information dienen

Kasten 2: Ausschnitt aus dem Mplus-Input zur Prüfung der konfiguralen Messinvarianz

```
TITLE: konfigurale Messinvarianz
DATA: FILE IS Einsamkeit.dat;
VARIABLE:NAMES ARE item1 item2 item3 item4 SPF;

Group IS SPF (0=noSPF 1=i);

ANALYSIS: ESTIMATOR IS MLR;

MODEL:
!loadings
Einsam by item1@1 (L1) item2* (L2) item3* (L3) item4* (L4);
!intercepts
[item1*] (I1); [item2*] (I2); [item3*] (I3); [item4*] (I4);
!residuals
item1* (E1); item2* (E2); item3* (E3); item4* (E4);
!factor variance
Einsam*;
!factor mean
[Einsam@0];

MODEL i:
Einsam by item1@1 item2* item3* item4*;
[item1*]; [item2*]; [item3*]; [item4*];
item1-item4*; Einsam*; [Einsam@0];

OUTPUT:
MODINDICES STDYX;
```

Anmerkung: ! bezieht sich auf Kommentare, die lediglich der Information dienen

Kasten 3: Ausschnitt aus dem Mplus-Input zur Prüfung der metrischen Messinvarianz

```
TITLE: metrische Messinvarianz
DATA: FILE IS Einsamkeit.dat;
VARIABLE:NAMES ARE item1 item2 item3 item4 SPF;

Group IS SPF (0 = noSPF 1 = i);

ANALYSIS: ESTIMATOR IS MLR;

MODEL:
!loadings
Einsam by item1* (L1) item2* (L2) item3* (L3) item4* (L4);
!intercepts
[item1*] (I1); [item2*] (I2); [item3*] (I3); [item4*] (I4);
!residuals
item1* (E1); item2* (E2); item3* (E3); item4* (E4);
!factor variance
Einsam@1;
!factor mean
[Einsam@0];

MODEL i:
Einsam by item1* (L1) item2* (L2) item3* (L3) item4* (L4);
[item1*]; [item2*]; [item3*]; [item4*];
item1-item4*; Einsam*; [Einsam@0];

OUTPUT:
MODINDICES STDYX;
```

Anmerkung: ! bezieht sich auf Kommentare, die lediglich der Information dienen

Kasten 4: Ausschnitt aus dem Mplus-Input zur Prüfung der skalaren Messinvarianz

```
TITLE: skalare Messinvarianz
DATA: FILE IS Einsamkeit.dat;
VARIABLE:NAMES ARE item1 item2 item3 item4 SPF;

Group IS SPF (0 = noSPF 1 = i);

ANALYSIS: ESTIMATOR IS MLR;

MODEL:
!loadings
Einsam by item1* (L1) item2* (L2) item3* (L3) item4* (L4);
!intercepts
[item1*] (I1); [item2*] (I2); [item3*] (I3); [item4*] (I4);
!residuals
item1* (E1); item2* (E2); item3* (E3); item4* (E4);
!factor variance
Einsam@1;
!factor mean
[Einsam@0];
```

➥

```
MODEL i:
Einsam by item1* (L1) item2* (L2) item3* (L3) item4* (L4);
[item1*] (I1); [item2*] (I2); [item3*] (I3); [item4*] (I4);
item1-item4*; Einsam*; [Einsam@0];

OUTPUT:
MODINDICES STDYX;
```

Anmerkung: ! bezieht sich auf Kommentare, die lediglich der Information dienen

Kasten 5: R-Script für die Prüfung von Messinvarianz

```
#notwendige Pakete installieren
install.packages(„lavaan“)    # für CFA und SEM
install.packages(„semTools“)  # u.a. für die Messinvarianzprüfung
install.packages(„foreign“)   # um Daten aus SPSS zu laden
install.packages(„pairwise“)  # für die IRT-Analysen
install.packages(„car“)       # für das Umkodieren von Variablen

#notwendige Pakete laden
library(lavaan)
library(semTools)
library(foreign)
library(pairwise)
library(car)

#Ordner bestimmen, in dem die Daten gespeichert sind und in dem auch die folgenden
Output-Daten gespeichert werden
setwd(dir = „H:/ Autor1 Paper/R“)

#Einlesen der Daten
dat <- read.table(„ESP_Messinvarianz_Einsamkeit_fertig.dat“, header=FALSE)

#Faktorenmodell erstellen
model <- ‘ einsam =~ V1 + V2 + V3 + V4 ‘

#Messinvarianzprüfung
measurementInvariance(std.lv = TRUE, model, dat, estimator = „MLR“, group=“V5“,
strict=TRUE)
```

Anmerkung: # bezieht sich auf Kommentare, die lediglich der Information dienen

Kasten 6: R-Script für die Prüfung von partieller Messinvarianz (Fortsetzung Kasten 5)

```
####Partielle Messinvarianz
fit1a<- cfa(model, data = dat, estimator = "MLR", group = "V5", group.equal = c("loa-
dings", "intercepts"))
summary(fit1a, fit.measures=TRUE)
fit1b <- cfa(model, data = dat, estimator = "MLR", group = "V5", group.equal = c("loa-
dings", "intercepts", "residuals"),
          group.partial = c("V1~~V1", "V2~~V2", "einsam~~einsam"))
summary(fit1b, fit.measures=TRUE)
```

Anmerkung: # bezieht sich auf Kommentare, die lediglich der Information dienen

Kasten 7: R-Script für die DIF-Analysen (Fortsetzung Kasten 5)

```
########DIF-Analysen vor dem Hintergrund der IRT ##################
#Der grm-Befehl weiter unten erfordert, dass die Antwortkategorien mit 0 beginnen, daher
ist eine Umkodierung der 4 Items erforderlich
dat1 <- recode(dat$V1,"1=0;2=1;3=2;4=3;5=4")
dat2 <- recode(dat$V2,"1=0;2=1;3=2;4=3;5=4")
dat3 <- recode(dat$V3,"1=0;2=1;3=2;4=3;5=4")
dat4 <- recode(dat$V4,"1=0;2=1;3=2;4=3;5=4")

#Die vier Items werden wieder zu einer Matrix zusammengebaut.
dat_new <- cbind(dat1, dat2, dat3, dat4)

#Befehl für die graphische Messinvarianzprüfung (GMP1)
GMP1 <- grm(daten=dat_new, m=5, splitcrit = dat$V5)
#Plotten des GMP
plot(GMP1, ylab="mit SPF", xlab="ohne SPF", main ="Differential Item Functioning")
```

Anmerkung: # bezieht sich auf Kommentare, die lediglich der Information dienen