6 Messinvarianz und Validität der Lesefreude- und Leseselbstkonzeptskalen über die PIRLS-Österreich-Untersuchungen 2006, 2011 und 2016

Burkhard Gniewosz & Gabriela Gniewosz

Dieser Beitrag untersucht mittels latenter Mehrgruppenmodelle, inwieweit die Lesefreude und das Leseselbstkonzept über die österreichischen PIRLS-Erhebungen der Jahre 2006, 2011 sowie 2016 hinweg in gleicher Weise modelliert werden können. Das stellt eine Voraussetzung für valide Trendvergleiche zwischen den Messzeitpunkten dar. Es wird geprüft, inwieweit sich die wiederholt über die gleichen Items gemessenen Konstrukte in invarianten Messmodellen abbilden lassen. Als Außenkriterien zur ergänzenden Prüfung der Konstruktvalidität werden das Geschlecht sowie die Lesekompetenztestwerte der Schüler/innen herangezogen.
Die Analysen zeigen, dass eine alle drei Messzeitpunkte übergreifende, konstante Modellierung nicht möglich ist. Die Ursache liegt hier v. a. an der eingeschränkten Anzahl der parallel gehaltenen Items über alle drei Messzeitpunkte hinweg. Gestützt auf einen größeren Itempool, wie über die Messzeitpunkte 2011 und 2016, ist es möglich, die Lesefreude und das Leseselbstkonzept messinvariant und über die Außenkriterien zu modellieren. Somit sind valide Trendvergleiche nur über die Messzeitpunkte 2011 und 2016 möglich.

Es ist ein großer Vorteil von internationalen Vergleichsstudien im Bereich des Bildungsmonitorings, dass mittels einer wiederholten Messung derselben Konstrukte Trendanalysen bezüglich der gleichen Population (aber unterschiedlicher Individuen in den jeweiligen Stichproben) möglich sind. Sicherlich erlauben diese Analysen keine Schlussfolgerungen über Veränderungen auf der Individualebene, wie es beispielsweise längsschnittliche Panel-Untersuchungen ermöglichen (Pforr & Schröder, 2015). Dennoch ermöglichen Trendanalysen Schlussfolgerungen auf der Ebene der Bildungssysteme oder Länder, indem sie auf Veränderungen in den untersuchten Konstrukten selbst hinweisen. Eine wesentliche Voraussetzung für eine belastbare Interpretation von Trendanalysen sind verschiedene Arten von Messinvarianz, die über restringierte Messmodelle der zu analysierenden latenten, d. h. nicht direkt beobachtbaren Variablen überprüft werden. Hierfür muss gezeigt werden, dass die Items, die einer latenten Variable zugrunde liegen, zu den verschiedenen Messzeitpunkten in der gleichen Art und Weise das Konstrukt abbilden können. Dies ist eine notwendige, aber nicht hinreichende Bedingung einer validen messzeitpunktübergreifenden Interpretation der gemessenen Variablen.

Für aussagekräftige Trendanalysen muss zudem sichergestellt werden, dass die Variablen zu den unterschiedlichen Messzeitpunkten auch inhaltlich die gleichen Interpretationen zulassen (Konstrukt, Frey & Jude, 2012), was der aktuellen Definition von Validität über die Zeit hinweg entspricht (Messick, 1989). Eine praktikable Möglichkeit, die sich insbesondere in Large-Scale-Studien anbietet, ist, im Sinne der Konstruktvalidität zu zeigen, dass die latenten Konstrukte in gleicher Art und Weise mit verschiedenen Außenkriterien korrelieren. Nach Hartig, Frey & Jude (2012) umfasst die „Konstruktvalidität […] die empirischen Befunde und Argumente, mit denen die Zuverlässigkeit der Interpretation von Testergebnissen im Sinne erklärender Konzepte, die sowohl die Testergebnisse selbst als auch die Zusammenhänge der Testwerte mit anderen Variablen erklären, gestützt wird (Messick, 1995, S. 743)". Eine hierfür geeignete Möglichkeit ist die Prüfung einer angenommenen Zusammenhangsstruktur mehrerer Konstrukte bei gleichzeitiger Kontrolle der Messfehler im Rahmen der Strukturgleichungsmodellierung (ebd.).

Unter dieser Perspektive werden in diesem Beitrag die Skalen der Lesefreude und des Leseselbstkonzepts untersucht, wie sie in den PIRLS-Erhebungen 2006, 2011 und 2016 erhoben wurden. Nationale und internationale Forschungsarbeiten verweisen auf die Bedeutung dieser zwei motivationalen Merkmale nicht nur für die Leseleistung (z. B. Froiland & Oros, 2014; Malanchini et al., 2017) von Schülerinnen und Schülern, sondern auch für die weitere akademische Entwicklung (z. B. Simpkins, Fredricks & Eccles, 2012) bis hin zur Berufswahl junger Menschen (z. B. Durik, Vida & Eccles, 2006). Daher sind diese motivationalen Konstrukte regelmäßig Teil der Datenerhebungen in internationalen Vergleichsstudien, wie sie auch die PIRLS-Erhebungen darstellen. Trotz der Bedeutung motivationaler Merkmale für die weitere akademische und nichtakademische Entwicklung sind Lesefreude und Leseselbstkonzept nur selten Gegenstand psychometrischer Analysen. Während in den Berichten zur Veränderung der Leistungsergebnisse über die verschiedenen Erhebungswellen hinweg bereits in der Skalierung auf ein invariantes Messmodell geachtet wird, fehlen in den meisten Fällen derartige Analysen für die eingesetzten motivationalen Skalen.

Methodisches Vorgehen

Dieser Beitrag fokussiert auf Skalen zur Lesefreude und zum akademischen Selbstkonzept im Bereich Lesen bei Schülerinnen und Schülern der 4. Schulstufe. Das Ziel ist, diese Skalen nach dem von Cheung und Rensvold (2002) vorgeschlagenen Verfahren auf Invarianz über die drei Erhebungszeitpunkte zu testen. Hierzu werden die drei Messzeitpunkte als Gruppierungsvariable bzw. als potenzieller Moderator aufgefasst, sodass sich die Invarianz der betrachteten latenten Konstrukte über diese multiplen „Messzeitpunkte" (Gruppe 1: 2006; Gruppe 2: 2011; Gruppe 3: 2016) im Sinne eines Mehrgruppenmodells spezifizieren und testen lässt.

In einem *ersten Schritt* wird die konfigurale Invarianz getestet, d. h. es wird geprüft, ob dieselben Items aus den verschiedenen Erhebungswellen die angenommenen latenten Konstrukte mit einer guten Modellpassungsgüte abbilden. Der *zweite Schritt* bezieht sich auf die Frage nach der sog. metrischen Invarianz. Hierzu werden im Mehrgruppenmodell die Ladungen der Items bzgl. der latenten Faktoren über die Zeit gleichgesetzt. Im *dritten Schritt,* dem Test der skalaren Invarianz, werden zusätzlich die manifesten Intercepts der Items auf Invarianz getestet und hierdurch wird geprüft, ob itemspezifische Niveau- bzw. Mittelwertunterschiede über die Messzeitpunkte existieren. Es schließt sich im *vierten Schritt* der Test der Messgüte bzw. Reliabilität der Einzelitems über die Zeit an, indem die Messfehler der Items gleichgesetzt werden. Verändert sich die Modellpassungsgüte über die spezifizierten Modelle nicht oder nur marginal, kann von einer Invarianz ausgegangen werden. Cheung und Rensvold (2002) schlagen vor, jeweils die Differenz zwischen den Modellpassungsindikatoren der unterschiedlichen Stufen zur Entscheidung heranzuziehen. Unterscheiden sich das strengere (z. B. Schritt 3) und das weniger strenge Modell (z. B. Schritt 2) signifikant voneinander, so ist das Invarianzlevel des weniger strengen Modells anzunehmen, da das strengere Modell zu einer signifikanten Verschlechterung der Modellpassung führen würde. Andernfalls ist das strengere (Invarianz-)Modell anzunehmen. Zur Abschätzung der Unterschiedlichkeit der Modelle verweisen die Autoren auf die Differenz der Comparative-Fit-Indizes (CFI) und schlagen einen Cut-off-Wert von einer Verschlechterung von maximal .01 als akzeptabel vor. Dieses Vorgehen hat sich in der Literatur als Standard etabliert (siehe Putnick & Bornstein, 2016).

Aufbauend auf den invarianten Messmodellen wird ergänzend geprüft, ob sich die latenten Mittelwerte und Varianzen der Lesefreude- und Selbstkonzeptskalen über die betrachteten Messzeitpunkte hinweg unterscheiden. Bezüglich der zeitübergreifenden Mittelwerte und Varianzen ist anzunehmen, dass sich keine bedeutsamen Unterschiede ergeben, da es keine größeren bildungssystemischen Veränderungen im Zeitraum zwischen 2006 und 2016 für die Zielpopulation gab, die einen Kohortenunterschied in den Mittelwerten und Varianzen der Lesefreude- und Selbstkonzeptskalen erwarten ließen. Invariante latente Mittelwerte und Varianzen sind allerdings keine Voraussetzungen für Messinvarianz.

Für die Überprüfung der Konstruktvalidität wird im Anschluss getestet, ob die Zusammenhänge zum einen mit dem Geschlecht und zum anderen mit den Scores der Lesekompetenztests über die Zeit hinweg variieren. Dieses Vorgehen gibt Aufschlüsse über Veränderungen in der Validität der verwendeten Skalen. Basierend auf Befunden zur Lesekompetenz und -motivation sollte hierbei folgendes Muster über die Messzeitpunkte zu replizieren sein: Sowohl die Lesefreude (Steinmayr & Spinath, 2007, 2009) als auch das Selbstkonzept (Lohbeck & Möller, 2017; Susperreguy, Davis-Kean, Duckworth & Chen, 2017) sollten positiv mit den Leistungsindikatoren korrelieren. Zudem ist zu erwarten, dass Mädchen im Vergleich zu Jungen sowohl eine höhere Lesefreude (Baker & Wigfield, 1999; Spinath, Freudenthaler & Neubauer, 2010) als auch ein höheres Selbstkonzept berichten (Wilgenbusch & Merrell, 1999).

Für die Analysen, durchgeführt in Mplus 8 (Muthén & Muthén, 2017), wurde ein Mehrgruppenmodell mit den Messzeitpunkten 2006, 2011 und 2016 als Gruppenfaktor spezifiziert. Das ermöglicht, beide motivationalen Konstrukte gleichzeitig über die drei Messzeitpunkte in einem Modell darzustellen und zu testen. Das invariante Modell wurde schließlich um die Außenkriterien Geschlecht und Lesekompetenzwerte im Rahmen eines (korrelativen) Strukturgleichungsmodells erweitert.

Stichprobe & Instrumente

Den Analysen lagen die österreichischen nationalen Stichproben der Erhebungen PIRLS der Jahre 2006 (N = 5067, 49,4 % weiblich), 2011 (N = 4670, 48,7 % weiblich) und 2016 (N = 4360, 48,5 % weiblich) zugrunde. Es handelt sich hier um repräsentative geklumpte Zufallsstichproben, bezogen auf die Zielpopulation aller österreichischen Schüler/innen im vierten Schuljahr. Die Details bezüglich Sampling-Design und realisierter Quoten sind den technischen Berichten zu entnehmen (Haider & Suchań, 2007; Suchań & Schreiner, 2012; Wallner-Paschon & Itzlinger-Bruneforth, 2016). Für die Analysen wurde das Total Student Weight verwendet.

Die verwendeten Items der Untersuchungen 2011 und 2016 waren parallel. In der Erhebung im Jahr 2006 wurden hingegen weniger und leicht anders formulierte Items verwendet. In Tabelle 1 sind die Itemformulierungen für Lesefreude und Leseselbstkonzept für die parallelen Erhebungen 2011/2016 (rechte Spalte) sowie die Erhebung 2006 (linke Spalte) dargestellt. Alle Items wurden über eine vierstufige Ratingskala in der Selbstauskunft durch die Schüler/innen eingeschätzt (1 – Stimme völlig zu; 2 – Stimme eher zu; 3 – Stimme eher nicht zu; 4 – Stimme überhaupt nicht zu). Die Items wur-

PIRLS 2006	PIRLS 2011 & PIRLS 2016
Lesefreude	
Ich unterhalte mich gern mit anderen Leuten über Bücher.	Ich unterhalte mich gern mit anderen Leuten über das, was ich gelesen habe.
Ich würde mich freuen, wenn mir jemand ein Buch schenken würde.	Ich würde mich freuen, wenn mir jemand ein Buch schenken würde.
Ich finde Lesen langweilig. (R)	Ich finde Lesen langweilig. (R)
Ich lese gern.	Ich lese gern.
Ich lese nur, wenn ich muss. (R)	Ich lese nur, wenn ich muss. (R)
	Ich hätte gern mehr Zeit zum Lesen.
	Ich lerne viel durch das Lesen.
	Ich mag es, Texte zu lesen, die mich zum Nachdenken bringen.
	Ich mag es, wenn ich mich durch ein Buch in andere Welten versetzen kann.
	Für mich ist Lesen Zeitverschwendung. (R)
	Lesen ist eines meiner liebsten Hobbys.
Leseselbstkonzept	
Ich kann nicht so gut lesen wie andere Schüler aus meiner Klasse. (R)	Lesen fällt mir schwerer als vielen Kindern meiner Klasse. (R)
Lesen fällt mir sehr leicht.	Lesen fällt mir leicht.
	Es fällt mir schwer, Geschichten mit schwierigen Wörtern zu lesen. (R)
	Lesen fällt mir schwerer als alle anderen Fächer. (R)
	Normalerweise bin ich gut in Lesen.
Anmerkung: Negativ formulierte Items wurden umkodiert und sind mit (R) gekennzeichnet.	

Tabelle 1: Formulierung der eingesetzten Items (PIRLS 2006 und 2016/2011)

den so rekodiert, dass höhere Werte eine geringere Lesefreude sowie ein niedrigeres Leseselbstkonzept widerspiegeln.

Aufgrund der nach Anzahl und Formulierung unterschiedlichen Items der Messzeitpunkte 2006 und 2011/2016 wurden die weiteren Analysen in zwei Varianten durchgeführt: In der ersten Version wurde der reduzierte Itempool der weitestgehend ähnlichen Items aller drei Messzeitpunkte (2006, 2011 und 2016) in die Analysen aufgenommen. Hierbei ist die Reliabilität für die fünf Items umfassende Lesefreude insgesamt als zufriedenstellend bis gut einzuschätzen (Cronbach's Alpha: 2006: .73; 2011: .72; 2016: .72). Die geringen Reliabilitäten für das Leseselbstkonzept mit nur zwei Items deuten allerdings bereits auf Schwierigkeiten im Messmodell hin (Cronbach's Alpha: 2006: .52; 2011: .55; 2016: .44). In der zweiten Variante wird der auf den 2011 und 2016 größeren Itempool fokussiert, für den sich bereits in den deskriptiven (Vor-)Analysen günstigere psychometrische Eigenschaften zeigen. Für die Lesefreude erbrachte die Skala mit elf Items gute Reliabilitäten (Cronbach's Alpha: 2011: .84; 2016: .83). Die Reliabilität für die fünf Items umfassende Skala des Selbstkonzepts erwies sich als zufriedenstellend bis gut (Cronbach's Alpha: 2011: .74; 2016: .70).

Schließlich werden neben dem Geschlecht (Mädchen = 1; Jungen = 2) der Gesamtscore des Lesekompetenztests sowie die Scores der vier Subskalen „Literarisches Lesen", „Informationslesen", „Interpretieren, Verknüpfen und Bewerten" und „Wiedergeben und einfaches Schlussfolgern" verwendet. Für die Kompetenztestwerte werden die auf einem IRT-Modell für polytome Daten basierenden Plausible Values in die Analysen aufgenommen (Haider & Suchań, 2007; Suchań & Schreiner, 2012; Wallner-Paschon & Itzlinger-Bruneforth, 2016).

Fehlende Werte wurden in allen Modellen über Full-Information-Maximum-Likelihood-Schätzer mit robusten Standardfehlern geschätzt. Einzelne fehlende Werte wurden dabei

nicht ausgeschlossen, sondern basierend auf der gesamten Information im Modell geschätzt. Der Vorteil liegt einerseits in der Maximierung der statistischen Power der zugrundeliegenden Modelle und andererseits in der Vermeidung systematischer Verzerrungen durch den Ausschluss spezifischer Personengruppen mit hohen Raten an fehlenden Werten.

Ergebnisse über Messzeitpunkte 2006, 2011 und 2016

Die Basis für die Invarianztestung über alle drei Messzeitpunkte bilden im ersten gemeinsamen Mehrgruppenmodell die latenten Konstrukte Lesefreude und Leseselbstkonzept. Als manifeste Indikatoren für die Lesefreude wurden für jeden Messzeitpunkt die fünf parallel formulierten Items und für das Selbstkonzept jeweils zwei parallele manifeste Items für die Modellierung der latenten Konstrukte genutzt (siehe Tabelle 1). Da einige Items negativ formuliert waren und in der Folge umkodiert wurden, ist anzunehmen, dass die Antwortmuster bzgl. dieser rekodierten Items einander ähnlicher sind. Bleiben diese aufgrund der Negativformulierung gemeinsamen Varianzanteile auf der Messmodellebene unberücksichtigt, kann dies zu einer Überschätzung der Korrelation der latenten Variablen führen. Um diese Methodenvarianz zu berücksichtigen, wurden die Kovarianzen der Messfehler der umkodierten Items frei geschätzt (siehe correlated uniquenesses, Marsh, Byrne & Craven, 1992). Davon zu unterscheiden ist die Kovarianz der beiden latenten Konstrukte, die ebenfalls frei geschätzt wurde. Diese latente Kovarianz zwischen Lesefreude und Leseselbstkonzept verweist auf das Ausmaß der gemeinsamen Varianzanteile der beiden motivationalen Aspekte auf der Konstruktebene.

Das im ersten Schritt spezifizierte *konfigurale Modell* verweist bereits mit Blick auf die globalen Fit-Indices auf Defizite der Modellpassung (siehe Tabelle 2). Nach Schermelleh-Engel, Moosbrugger & Müller (2003) spricht man von einer guten Modellpassung, wenn RMSEA < .06, SRMR < .06 und CFI > .96. Hier deuten sich bereits Schwierigkeiten in der parallelen Modellierung beider Konstrukte über alle drei Messzeitpunkte hinweg an. Akzeptierte man diese bestenfalls zufriedenstellende Modellpassungsgüte, wäre noch eine *metrische Modellierung* (Schritt 2) der beiden Konstrukte gegeben, da sich der CFI nur minimal verändert. Der dritte Schritt zur *skalaren Messinvarianz* führt jedoch zu einer deutlichen Verschlechterung der Modellpassung und weist darauf hin, dass es – unabhängig von der Ausprägung der latenten Variable – auf Itemebene Mittelwert- bzw. Niveauunterschiede zwischen den drei Erhebungszeitpunkten gibt.

Auch Zusatzanalysen, die weniger strenge Annahmen über die Invarianzstruktur aufstellen, können kein Mindestmaß an Modellpassungsgüte erreichen und sollen an dieser Stelle nicht weiter im Detail berichtet werden. Demnach können weder die Testung der latenten Konstrukte Lesefreude und Leseselbstkonzept in Einzelanalysen, noch die Aufweichung der hier beschriebenen Invarianzkriterien im Sinne einer partiellen Invarianz (siehe kommender Abschnitt) eine Modellpassung erzielen, die eine skalare Messinvarianz rechtfertigen würde.

Zusammenfassend können auf Basis der erzielten Befunde zunächst keine invarianten Messmodelle für die latenten Konstrukte von Lesefreude und Leseselbstkonzept über alle drei Messzeitpunkte angenommen werden. Folglich wurde auf die Prüfung der Konstruktvalidität der hier betrachteten latenten Konstrukte mit den beschriebenen Außenkriterien verzichtet.

Ergebnisse über Messzeitpunkte 2011 und 2016

Auf Basis des größeren Itempools der PIRLS-Erhebungsjahre 2011 und 2016 wurden in einem zweiten Auswertungsblock in Analogie zum ersten Abschnitt die gleichen Analyseschritte unternommen. Die latenten Konstrukte wurden wiederum über – jetzt zwei Erhebungszeitpunkte – parallel modelliert. Im Unterschied zu der vorherigen Modellierung konnten für die Lesefreude elf und für das Leseselbstkonzept fünf Items zur Modellierung der latenten Konstrukte herangezogen werden.

Modell	χ^2	*df*	*p*	RMSEA	SRMR	TLI	CFI	ΔCFI
Konfigural	434,80	30	< .001	.054	.030	.933	.968	
Metrisch	459,77	40	< .001	.047	.033	.948	.967	.001[a)]
Skalar	1335,21	50	< .001	.074	.065	.873	.899	.068[b)]
Messfehler	1325,89	64	< .001	.065	.068	.903	.901	–.002[c)]

Anmerkungen: a) Referenzmodell = Konfigural; b) Referenzmodell = Metrisch; c) Referenzmodell = Skalar; RMSEA: Root Mean Square Error of Approximation; SRMR: Standardized Root Mean Square Residual; TLI: Tucker Lewis Index; CFI: Comparative Fit Index; ΔCFI: Differenzen der CFI-Werte zwischen den Schritten der Invarianzüberprüfung.

Tabelle 2: Ergebnisse der Tests auf Messmodellinvarianz über die Messzeitpunkte 2006, 2011, 2016

Modell	χ^2	*df*	*p*	RMSEA	SRMR	TLI	CFI	ΔCFI
Konfigural	2054,24	176	<.001	.049	.040	.915	.938	
Metrisch	2139,44	190	<.001	.048	.044	.918	.935	.003[b)]
Skalar	2577,73	204	<.001	.051	.049	.907	.921	.014[c)]
Partiell Skalar[a)]	2417,97	203	<.001	.049	.045	.913	.926	.009[c)]
Messfehler	2402,96	219	<.001	.047	.046	.921	.927	–.001[d)]

Anmerkungen: a) Der Intercept des Items „Ich lese nur, wenn ich muss" wurde zu beiden Messzeitpunkten frei geschätzt; b) Referenzmodell = Konfigural; c) Referenzmodell = Metrisch; d) Referenzmodell = partiell Skalar.

Tabelle 3: Ergebnisse der Tests auf Messmodellinvarianz über die Messzeitpunkte 2011, 2016

Die Ergebnisse der Invarianztestungen sind in Tabelle 3 dargestellt. Es zeigt sich, dass die *konfigurale* (Schritt 1) und *metrische Invarianz* (Schritt 2) ohne Schwierigkeiten dargestellt werden können. Der gute globale Modellfit zeigt, dass sowohl die Faktorenstruktur als auch die Faktorenladungen als äquivalent über die Erhebungszeitpunkte von 2011 und 2016 angenommen werden können. Im dritten Schritt, der Testung auf *skalare Invarianz,* wurde das Kriterium für die Invarianz knapp verfehlt. Es liegen also itemspezifische Schwierigkeitsunterschiede (d. h. Unterschiede im Niveau der manifesten Items) zwischen den Erhebungsjahren vor. Eine sich anschließende vertiefende Inspektion der Messmodelle zeigt, dass dies insbesondere auf den Intercept des Items „Ich lese nur, wenn ich muss" zurückzuführen ist. Die Relaxierung der Invarianzrestriktion dieses einen Parameters über beide Messzeitpunkte führt dazu, dass eine *partielle skalare Invarianz* angenommen werden kann. Schmitt, Golubovich und Leong (2011) konnten zeigen, dass eine geringe Anzahl von variierenden Parametern keine verzerrten Schätzungen auf der latenten Ebene nach sich zog. Der letzte Schritt, nämlich die Gleichsetzung der *Messfehler* der manifesten Indikatoren über die Messzeitpunkte hinweg, führte schließlich zu keiner weiteren Verschlechterung der Modellpassungsgüte. Daher kann auch eine Invarianz der Messgüte der einzelnen Items zwischen beiden Messzeitpunkten gezeigt werden. Das ist ein Hinweis darauf, dass die Indikatoren über gleiche Reliabilitäten über die Messzeitpunkte verfügen.

Aufbauend auf diesem invarianten Messmodell war zu prüfen, inwieweit sich die latenten Mittelwerte und Varianzen sowie das Korrelationsmuster mit wichtigen Außenkriterien über die Erhebungsjahre 2011 und 2016 hinweg unterscheiden. In Tabelle 4 sind die latenten Mittelwerte und Varianzen beider Konstrukte über die beiden Messzeitpunkte sowie jeweils die Korrelationen mit dem Geschlecht, dem Gesamtscore und den vier Subtestscores der Lesekompetenz dargestellt. Die Ergebnisse entsprechen dem laut Literatur zu erwartenden Zusammenhangsmuster.

Die augenscheinliche Ähnlichkeit in den Parametern über die betrachteten Messzeitpunkte konnte zudem statistisch

	Lesefreude		Leseselbst-konzept	
	2011	2016	2011	2016
Mittelwerte	2,45	2,48	1,67	1,66
Varianzen	0,21	0,20	0,21	0,20
Korrelationen mit				
Gesamttestscore	–.20	–.21	–.34	–.36
Literarisches Lesen	–.20	–.19	–.33	–.34
Informationslesen	–.19	–.22	–.32	–.36
Interpretieren, Verknüpfen und Bewerten	–.20	–.22	–.35	–.37
Wiedergeben und einfaches Schlussfolgern	–.19	–.19	–.31	–.33
Geschlecht	.27	.23	.08	–.04

Anmerkung: Für alle Parameter gilt p < .001.

Tabelle 4: Mittelwerte, Varianzen und Korrelationen über die Messzeitpunkte 2011, 2016

abgesichert werden. Im gleichen Mehrgruppenmodell wurde geprüft, ob sich die Mittelwerte, Varianzen sowie die Korrelationen mit den Außenkriterien zwischen 2011 und 2016 signifikant unterscheiden. Die Ergebnisse der Differenzentests sind in Tabelle 5 dargestellt. Die Mittelwerte und Varianzen von Lesefreude und Leseselbstkonzept sowie die Korrelationen mit den Außenkriterien variieren wie angenommen nicht zwischen den Erhebungen von 2011 und 2016.

Mit Bezug auf die Konstruktvalidität lässt sich für beide Messzeitpunkte feststellen, dass eine niedrig ausgeprägte Lesefreude bzw. ein negativeres Leseselbstkonzept mit einer geringeren Lesekompetenz (sowohl im Gesamtscore als auch in den Teildimensionen) einhergeht. Zudem kann für beide Messzeitpunkte gezeigt werden, dass Jungen im Vergleich zu

	Differenz 2011 vs. 2016	*SE*	Differenz /SE[b)]	*p*
Lesefreude				
Mittelwerte	–0,01	0,01	–0,51	.611
Varianzen	0,01	0,01	0,82	.413
Korrelationen[a)] mit				
Gesamttestscore	0.18	0.78	0.23	.817
Literarisches Lesen	–0.48	1.18	–0.41	.683
Informationslesen	0.90	0.87	1.03	.305
Interpretieren, Verknüpfen und Bewerten	0.27	0.96	0.28	.777
Wiedergeben und einfaches Schlussfolgern	–0.07	0.88	–0.08	.939
Geschlecht	0.01	0.01	1.77	.077
Selbstkonzept				
Mittelwerte	0,01	0,01	0,51	.608
Varianzen	0,02	0,01	1,47	.141
Korrelationen[a)] mit				
Gesamttestscore	0.66	0.88	0.74	.457
Literarisches Lesen	0.21	1.09	0.19	.848
Informationslesen	1.19	1.03	1.16	.247
Interpretieren, Verknüpfen und Bewerten	0.49	0.88	0.55	.580
Wiedergeben und einfaches Schlussfolgern	0.52	1.13	0.47	.642
Geschlecht	0.01	0.01	1.41	.159

Anmerkungen: a) Die Differenztestung erfolgte anhand der Kovarianzen. Daher ergibt sich dieser Differenzparameter bzgl. der Korrelationen nicht aus der numerischen Differenz der dargestellten Korrelationen in Tabelle 4. b) Entspricht einem t-Wert.

Tabelle 5: Differenzparameter bzgl. Mittelwerten, Varianzen und Korrelationen zwischen den Messzeitpunkten 2011 und 2016

Mädchen sowohl über eine geringere Lesefreude als auch ein niedrigeres Leseselbstkonzept berichten.

Schlussfolgerungen

Fasst man die berichteten Ergebnisse zusammen, kommt man zu dem Schluss, dass die geringe Anzahl der Items, die in allen drei Erhebungswellen zur Erfassung der Lesefreude und des Leseselbstkonzepts verwendet wurde, nicht ausreicht, um eine gleiche Messung der Konstrukte über alle drei Zeitpunkte hinweg sicherstellen zu können. Bereits auf der Ebene der unrestringierten bzw. nicht beschränkten Modellierung beider Konstrukte zu allen drei Zeitpunkten (konfigurales Messmodell) zeigte sich eine problematische Passungsgüte. Die auf diesem reduzierten Itempool basierenden Indikatoren repräsentieren also die angenommenen latenten Konstrukte nicht in der gleichen Art und Weise über alle drei PIRLS-Erhebungen. Dies zeigt sich bereits in den schlechteren Reliabilitäten. Daher sollte davon Abstand genommen werden, die Lesefreude und das Leseselbstkonzept über die geringe Anzahl von Items messzeitpunktübergreifend zu modellieren. Eine parallele inhaltliche Interpretation der latenten Variablen wäre, folgt man den hier präsentierten Ergebnissen, nicht möglich.

Wird hingegen eine größere Anzahl von Indikatoren für die Lesefreude und das Leseselbstkonzept verwendet, wie es über die Erhebungswellen von 2011 und 2016 möglich ist, zeigt sich ein positiveres Bild. Alle Stufen der Invarianztestung konnten weitestgehend nachgewiesen (vgl. Tab. 3) werden, sodass eine valide wellenübergreifende parallele Interpretation der latenten Konstrukte möglich ist. Eine solche Interpretation ist allerdings mit Blick auf die erreichte partielle skalare Invarianz vorzunehmen. Ein Item erwies sich in Bezug auf den manifesten Intercept, d. h. den itemspezifischen „Mittelwert", zwischen den Erhebungen 2011 und 2016 als nicht invariant. Allerdings folgt aus dieser partiellen skalaren Invarianz keine schwerwiegende Einschränkung der Interpretation der latenten Mittelwerte und Varianzen bzw. der Kovarianzen auf der Strukturebene (Schmitt, Golubovich & Leong, 2011).

Dieser Beitrag überprüfte ebenfalls die Konstruktvalidität dieser so modellierten latenten Variablen, indem über die Messzeitpunkte von 2011 und 2016 hinweg das Korrelationsmuster zu wichtigen Außenkriterien geprüft und miteinander verglichen wurde. Die korrelativen Beziehungen zwischen Geschlecht bzw. den Lesekompetenztestscores und der Lesefreude sowie dem Leseselbstkonzept unterscheiden sich nicht signifikant zwischen den beiden Erhebungsjahren und stehen im Einklang mit dem in der internationalen Literatur referierten Zusammenhangsmuster (Baker & Wigfield, 1999; Durik et al., 2006; Lohbeck & Möller, 2017; Spinath et al., 2010; Steinmayr & Spinath, 2007, 2009; Susperreguy et al., 2017; Wilgenbusch & Merrell, 1999). Es ist also davon auszugehen, dass auch die inhaltliche Interpretation der beiden als (partiell) invariant modellierten latenten Variablen für beide Messzeitpunkte durchaus vergleichbar ist.

Zusammenfassend kann also in diesem Beitrag gezeigt werden, dass mit der erweiterten Itembatterie zur Lesefreude und zum Leseselbstkonzept, wie sie bei den PIRLS-Erhebungen 2011 und 2016 eingesetzt wurde, beide Konstrukte so mo-

delliert werden können, dass mittels invarianter Messmodelle eine über beide Messzeitpunkte hinweg valide Interpretation von Ausprägungen oder Prädiktionen der latenten Variablen möglich ist. Diese Form der (partiell) invarianten Modellierung ist insbesondere dann zu berücksichtigen, wenn i. S. von Trendanalysen explizit auf messzeitpunkteübergreifende Vergleiche (z. B. zur Beschreibung von Zeitwandel- oder Kohorteneffekten) Bezug genommen wird.

Literatur

Baker, L. & Wigfield, A. (1999). Dimensions of children's motivation for reading and their relations to reading activity and reading achievement. *Reading Research Quarterly, 34* (4), 452–477. doi:10.1598/RRQ.34.4.4

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. doi:10.1207/S15328007SEM0902_5

Durik, A. M., Vida, M. & Eccles, J. S. (2006). Task Values and Ability Beliefs as Predictors of High School Literacy Choices: A Developmental Analysis. *Journal of Educational Psychology, 98* (2), 382–393. doi:10.1037/0022-0663.98.2.382

Froiland, J. M. & Oros, E. (2014). Intrinsic motivation, perceived competence and classroom engagement as longitudinal predictors of adolescent reading achievement. *Educational Psychology, 34* (2), 119–132. doi:10.1080/01443410.2013.822964

Haider, G. & Suchań, B. (Hrsg.). (2007). *PIRLS 2006. Internationaler Vergleich von Schülerleistungen. Technischer Bericht. Lesen in der Grundschule.* Salzburg: ZVB – Österreichisches Projektzentrum für Vergleichende Bildungsforschung.

Hartig, J., Frey, A. & Jude, N. (2012). Validität. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 143–171). Berlin: Springer. doi:10.1007/978-3-642-20072-4_7

Lohbeck, A. & Möller, J. (2017). Social and dimensional comparison effects on math and reading self-concepts of elementary school children. *Learning and Individual Differences, 54*, 73–81. doi:10.1016/j.lindif.2017.01.013

Malanchini, M., Wang, Z., Voronin, I., Schenker, V. J., Plomin, R., Petrill, S. A. et. al. (2017). Reading self-perceived ability, enjoyment and achievement: A genetically informative study of their reciprocal links over time. *Developmental Psychology, 53* (4), 698–712. doi:10.1037/dev0000209

Marsh, H. W., Byrne, B. M. & Craven, R. (1992). Overcoming Problems in Confirmatory Factor Analyses of MTMM Data: The Correlated Uniqueness Model and Factorial Invariance. *Multivariate Behavioral Research, 27* (4), 489–507. doi:10.1207/s15327906mbr2704_1

Messick, S. (1989). Validity. In R. L. Linn (Hrsg.), *Educational measurement* (S. 13–103). Washington, DC: American Council on Education and National Council on Measurement in Education.

Muthén, B. & Muthén, L. (2017). *Mplus (Version 8).* Los Angeles, CA: Muthén & Muthén.

Pforr, K. & Schröder, J. (2015). *Warum Panelstudien?* (GESIS Survey Guidelines). Mannheim, GESIS – Leibniz-Institut für Sozialwissenschaften. doi:10.15465/sdm-sg_007

Putnick, D. L. & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review, 41,* 71–90. doi:10.1016/j.dr.2016.06.004

Schermelleh-Engel, K., Moosbrugger, H. & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online, 8* (2), 23–74.

Schmitt, N., Golubovich, J. & Leong, F. T. (2011). Impact of measurement invariance on construct correlations, mean differences, and relations with external correlates: an illustrative example using Big Five and RIASEC measures. *Assessment, 18* (4), 412–427. doi:10.1177/1073191110373223

Simpkins, S. D., Fredricks, J. A. & Eccles, J. S. (2012). Charting the Eccles' expectancy-value model from mothers' beliefs in childhood to youths' activities in adolescence. *Developmental psychology, 48* (4), 1019–1032. doi:10.1037/a0027468

Spinath, B., Freudenthaler, H. & Neubauer, A. C. (2010). Domain-specific school achievement in boys and girls as predicted by intelligence, personality and motivation. *Personality and Individual Differences, 48* (4), 481–486. doi:10.1016/j.paid.2009.11.028

Steinmayr, R. & Spinath, B. (2007). Predicting School Achievement from Motivation and Personality. *Zeitschrift für Pädagogische Psychologie, 21* (3/4), 207–216. doi:10.1024/1010-0652.21.3.207

Steinmayr, R. & Spinath, B. (2009). The importance of motivation as a predictor of school achievement. *Learning and Individual Differences, 19* (1), 80–90. doi:10.1016/j.lindif.2008.05.004

Suchań, B. & Schreiner, C. (2012). *PIRLS & TIMSS 2011. Die Kompetenzen in Lesen, Mathematik und Naturwissenschaft am Ende der Volksschule. Technischer Bericht.* Salzburg: Bundesinstitut für Bildungsforschung, Innovation und Entwicklung des österreichischen Schulwesens (BIFIE).

Susperreguy, M. I., Davis-Kean, P. E., Duckworth, K. & Chen, M. (2017). Self-Concept Predicts Academic Achievement Across Levels of the Achievement Distribution: Domain Specificity for Math and Reading. *Child Development, 89* (6), 2196–2214. doi:10.1111/cdev.12924

Wallner-Paschon, C. & Itzlinger-Bruneforth, U. (2016). *PIRLS 2016.Technischer Bericht.* Salzburg: Bundesinstitut für Bildungsforschung, Innovation und Entwicklung des österreichischen Schulwesens (BIFIE).

Wilgenbusch, T. & Merrell, K. W. (1999). Gender differences in self-concept among children and adolescents: A meta-analysis of multidimensional studies. *School Psychology Quarterly, 14* (2), 101–120. doi:10.1037/h0089000