Z Erziehungswiss (2017) (Suppl 2) 20:229–258
DOI 10.1007/s11618-017-0749-z

ZfE CrossMark

IRT-Skalierung der Tests im PISA-Längsschnitt 2012/2013: Auswirkungen von Testkontexteffekten auf die Zuwachsschätzung

Gabriel Nagy · Oliver Lüdtke · Olaf Köller · Jörg-Henrik Heine

Online publiziert: 12. Mai 2017

Zusammenfassung Der vorliegende Beitrag beschäftigt sich mit der Item Response Theorie Skalierung der Tests im PISA-Längsschnitt 2012/2013. Vorgestellt werden Analysen zur längsschnittlichen Messinvarianz der Leistungstests in den Domänen Mathematik, Naturwissenschaften und Leseverständnis. Die Ergebnisse zeigen, dass klassische Analyseansätze der Untersuchung der längsschnittlichen Messinvarianz keine Hinweise für bedeutsame Verstöße gegen die Annahme invarianter Itemschwierigkeiten über die Zeitpunkte ergaben. Demgegenüber indizierten erweiterte Analysen unter der Berücksichtigung komplexer Bookleteffekte die Existenz von Testkontexteffekten, die mit dem Effekt der Itemposition kompatibel sind. Die PISA-Tests waren zu beiden Messzeitpunkten von Positionseffekten betroffen, die in der Gruppe der nichtgymnasialen Schulen besonders hoch ausfielen und über die Zeitpunkte zunahmen. Diese Befunde wurden aufgegriffen um Korrekturmaßnahmen abzuleiten, die einer Verzerrung der Ergebnisse durch den selektiven Stichprobendropout und der Zunahme der Positionseffekte entgegenwirken. Die Grenzen der Korrekturmöglichkeiten werden diskutiert.

Prof. Dr. G. Nagy (✉) · Prof. Dr. O. Lüdtke
Pädagogisch-Psychologische Methodenlehre, Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik, Olshausenstraße 62, 24118 Kiel, Deutschland
E-Mail: nagy@ipn.uni-kiel.de

Prof. Dr. O. Lüdtke
E-Mail: oluedtke@ipn.uni-kiel.de

Prof. Dr. O. Köller
Erziehungswissenschaft, Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik, Olshausenstraße 62, 24118 Kiel, Deutschland
E-Mail: koeller@ipn.uni-kiel.de

Dipl.-Psych J.-H. Heine
Zentrum für internationale Bildungsvergleichsstudien, TUM School of Education, Arcisstraße 21, 80333 München, Deutschland

Schlüsselwörter PISA-Studie · Stichprobendropout · Item Response Theorie · Testkontexteffekte

IRT Scaling of the Tests in PISA Longitudinal Assessment 2012/2013: Impact of Test Context Effects on the Growth Estimate

Abstract This paper deals with the item response theory scaling of the tests in PISA longitudinal assessment 2012/2013. It presents analyses testing longitudinal measurement invariance of achievement tests in the domains of mathematics, science and reading comprehension. The analyses showed that conventional approaches of investigating longitudinal measurement invariance revealed no indication of meaningful violations of time-invariant item difficulties. On the other hand, taking into account complex booklet effects, extended analyses indicated the existence of test context effects due to the effect of item positioning. At both measurement points the PISA tests were affected by position effects; these effects being especially noteworthy and increasing over the time points in the group of nonacademic track schools. The findings of the presented study were used to derive corrections in order to counteract distortions of the results by the selective sample dropout as well as the increase in position effects. The limitations of these corrections are discussed.

Keywords PISA study · Sample dropout · Item Response Theory · Text Context Effects

Schulleistungsstudien sollen einen Einblick in die Verteilung der Kompetenzausprägungen einer oder mehrerer (Teil-)Populationen von Schülerinnen und Schülern geben. Um dieses Ziel zu erreichen, werden psychometrische Messmodelle verwendet, die eine unverzerrte und gleichermaßen reliable Schätzung der Populationsverteilung der Kompetenzen erlauben (z. B. Wu 2005). Derartige Verfahren finden zunehmend Eingang in längsschnittlich konzipierten Schulleistungsstudien, die die Erfassung der Kompetenzentwicklung von Schülerinnen und Schülern anstreben.

Längsschnittliche Schulleistungsstudien setzen besondere Anforderungen an die wiederholte Erfassung der schülerseitigen Kompetenzniveaus. Die in diesem Kontext typischerweise verwendeten Messmodelle sind in der Tradition der Item Response Theorie (IRT; Lord 1980) verwurzelt. Ihnen liegt die Annahme zugrunde, dass die Eigenschaften der verwendeten Items (z. B. Itemschwierigkeiten) zeitlich konstant sind, sodass die Änderungen in der Wahrscheinlichkeit korrekter Itemantworten vollständig auf die Veränderung der zugrundeliegenden Personeneigenschaften zurückzuführen sind.

Die vorliegende Arbeit widmet sich der IRT-Skalierung der im PISA-Längsschnitt 2012/2013 eingesetzten Tests, die im Rahmen des PISA-Programms entwickelt wurden. Besonderes Augenmerk wurde auf die Möglichkeit systematischer Veränderungen der Itemschwierigkeiten gelegt, die sich in Abhängigkeit von Merkmalen der Items (z. B. Teilbereich einer Domäne, optimaler Lösungsweg, Form der Problemstellung, usw.) und/oder des Kontextes, in den die Items eines Tests eingebettet sind (z. B. Position im Leistungstest, Schwierigkeit vorangegangener Items,

usw.) ergeben können. Veränderungen der Itemschwierigkeiten in Abhängigkeit von Itemmerkmalen können sich beispielsweise in Folge von Änderungen in den curricularen Schwerpunkten des Unterrichts einstellen (Klieme und Baumert 2001), da Schülerinnen und Schüler in Abhängigkeit von den Inhaltsbereichen der verwendeten Items unterschiedliche Lerngewinne aufweisen könnten. Verstöße gegen die Annahme zeitlich invarianter Itemschwierigkeiten, die in Abhängigkeit des Testkontexts auftreten, können hingegen auf eine Veränderung der Testmotivation der Schülerinnen und Schüler (DeMars 2007) hinweisen, wenn die entsprechenden Effekte in Abhängigkeit von denPositionen der Items in einem Leistungstest auftreten (z. B. Debeer et al. 2014).

1 Messinvarianz in Large-Scale-Studien

Eine zentrale Idee des IRT-Ansatzes ist, dass die Wahrscheinlichkeiten richtiger Itemantworten eine Funktion feststehender Itemeigenschaften (z. B. Schwierigkeiten) und Personenmerkmale (d. h. Kompetenzausprägungen) sind. Der IRT-Ansatz ermöglicht es, die Validität der ermittelten Kompetenzunterschiede zwischen Teilpopulationen (z. B. Schulformen, Geschlechter, usw.) zu untersuchen. Die Validität der Leistungsvergleiche gilt unter anderem dann als gefährdet, wenn sich die Gruppenunterschiede in den beobachteten Itemantworten nicht hinreichend gut durch die Gruppenunterschiede in den zugrundeliegenden Kompetenzen erklären lassen (Holland und Wainer 1993). Formal gesehen bedeutet dies, dass valide Leistungsvergleiche einen hinreichenden Grad an Invarianz der Itemeigenschaften (z. B. Schwierigkeiten) zwischen Gruppen voraussetzen. Gravierende Verstöße gegen diese Invarianzannahme deuten darauf hin, dass das verwendete Testmaterial in den zu vergleichenden Gruppen unterschiedlich „funktioniert“, sodass die ermittelten Leistungsunterschiede auch auf die unterschiedliche „Funktionsweise“ der Testitems zurückgeführt werden könnten. Verstöße gegen die Annahme der Messinvarianz werden in aktuellen Schulleistungsstudien ernst genommen (z. B. Klieme und Baumert 2001). Typischerweise werden vor der Hauptuntersuchung groß angelegte Pilotstudien durchgeführt, in denen das verwendete Testmaterial hinsichtlich seiner Messinvarianz untersucht und optimiert wird (z. B. Pohl und Carstensen 2013).

Das Konzept der Messinvarianz spielt auch im Kontext längsschnittlicher Untersuchungen eine wichtige Rolle, da Änderungen in den beobachteten Itemantworten Lerngewinne und nicht Veränderungen der Messeigenschaften des verwendeten Testmaterials reflektieren sollen (Meade et al. 2005). Somit stellt sich die Frage, inwieweit die Konstanz der Itemeigenschaften über die betrachteten Messzeitpunkte gewährleistet ist. Obwohl die längsschnittliche Messinvarianz von Leistungstests im Rahmen von Pilotstudien optimiert werden kann, ist ein solches Vorgehen nicht üblich, da die Durchführung von Längsschnittstudien sehr zeit-, kosten und ressourcenaufwendig ist. Insofern Verstöße gegen die Messinvarianzannahme festgestellt werden, müssen methodisch-statistische Korrekturmaßnahmen angewendet werden, die einer Verzerrung der Zuwachsmessung entgegenwirken (z. B. Millsap und Kwok 2004).

1.1 Partielle Messinvarianz

Verstößen gegen die Annahme gruppen- und/oder zeitinvarianter Messungen wird in den meisten Studien durch nachträgliche Korrekturen entgegengewirkt. Grob gefasst lassen sich zwei Arten von Maßnahmen identifizieren, die zum einen auf einen Ausschluss von Items mit nicht-invarianten Eigenschaften setzen und zum anderen Verstöße gegen die Invarianzannahme im Messmodell der Kompetenzen berücksichtigen (z. B. Byrne et al. 1989). Im ersten Ansatz werden Items, die gegen die Invarianzannahme verstoßen, identifiziert und anschließend bei der Schätzung der Kompetenzniveaus ausgeschlossen. Im zweiten Ansatz werden die identifizierten nicht-invarianten Items bei der Schätzung der Kompetenzniveaus nicht ausgeschlossen, ihnen werden aber gruppen- und/oder messzeitpunktspezifische Itemparameter zugewiesen. In beiden Ansätzen erfolgt die Identifikation von Gruppenunterschieden bzw. von Zuwachsraten nur anhand der Items mit invarianten Eigenschaften. Die Berücksichtigung von Items, die gegen die Annahme der Messinvarianz im Messmodell verstoßen, führt zu reliableren Kompetenzschätzungen, da die nicht-invariaten Items dennoch Informationen zur Kompetenzschätzung beisteuern (Vandenberg und Lance 2000).

Verstöße gegen die Annahme längsschnittlicher Messinvarianz können aus verschiedenen Gründen auftreten. Abweichungen vom Ideal eines vollständig invarianten Messmodells können zum Beispiel differentiellen Lerngelegenheiten geschuldet sein. In diesem Fall kann eine verstärkte Fokussierung der Unterrichtshalte auf einen spezifischen Teilbereich einer Domäne (z. B. Division in Klassenstufe 3) dazu führen, dass die entsprechenden Items im Vergleich zum restlichen Testmaterial leichter werden und somit die Gesamtmenge aller Items nicht mehr dem Kriterium der Messinvarianz genügt. Ebenso könnte eine mögliche Vernachlässigung bestimmter Teilbereiche in späteren Jahrgangsstufen (z. B. quadratischer Gleichungen in der gymnasialen Oberstufe) dazu führen, dass die entsprechenden Items sogar schwerer werden, was wiederum Verstöße gegen die Annahme zeitlich invarianter Itemschwierigkeiten nach sich zieht. Darüber hinaus sind eine Vielzahl weitere curricularer Einflüsse vorstellbar, die Verstöße gegen die Annahme einer zeitlich invarianten Abfolge von Itemschwierigkeiten nach sich ziehen (z. B. zunehmende Vertrautheit mit bestimmten Darstellungsformen). All diesen Einflussgrößen ist gemeinsam, dass sie mit den Merkmalen der einzelnen Items interagieren (z. B. Teilbereich einer Domäne, optimaler Lösungsweg, Form der Darstellung, usw.) und nicht vom (Test-)Kontext, in dem die einzelnen Items eingebettet sind, abhängen.

Verstöße gegen das Kriterium zeitlich invarianter Itemschwierigkeiten können jedoch auch in Abhängigkeit vom Testkontext eintreten (Brennan 1992). So ist beispielsweise bekannt, dass die Lösungswahrscheinlichkeit eines Items in Folge dessen Position in einem Leistungstest sinkt (z. B. Meyers et al. 2009) und sich in Abhängigkeit der Schwierigkeit der vorangestellten Items ändern kann (z. B. Brennan 1992). Abweichungen vom Invarianzkriterium können sich somit in Folge von Änderungen der konkreten Darbietung der Items ergeben und/oder in Folge von systematischen Änderungen in den individuellen Reaktionen auf eine weitgehend unveränderte Testsituation (z. B. DeMars 2007) auftreten. So könnte eine zeitlich umfangreiche und anstrengende Kompetenztestung über die Messzeitpunkte eine

zunehmend aversive Reaktion auf Seiten der Schülerinnen und Schülern auslösen, die im Verlauf der Testbearbeitung zu einer Reduktion der Testanstrengung und letztendlich eine Herabsetzung der Wahrscheinlichkeit korrekter Itemantworten nach sich zieht (DeMars 2007). Ein solcher Prozess manifestiert sich in einer Zunahme von Testkontexteffekten, die in Abhängigkeit der Itempositionen auftreten.

Im Prinzip lassen sich in beiden Situationen partiell invariante Messmodelle einsetzen, sodass die Zuwächse relativ zu einer Gruppe von Referenzitems erfasst werden, für die das Kriterium der Messinvarianz gilt (Byrne et al. 1989; te Marvelde et al. 2006). Im Fall von Zuwächsen, die in Abhängigkeit von Merkmalen der Items auftreten, kann dies bedeuten, dass die Leistungszuwächse bezüglich einer inhaltlich homogenen Gruppe von Referenzitems definiert werden (z. B. Items zur Division, Subtraktion, usw.). Im Fall zeitspezifischer Testkontexteffekte bietet es sich an Items auszuwählen, die zu allen Messzeitpunkten nicht von derartigen Effekten betroffen sind.

1.2 (Item-)Positionseffekte in Schulleistungstests

In typischen Anwendungen von IRT-Modellen wird die Annahme getroffen, dass der Testkontext, in den ein Item eingebettet ist, keine Auswirkungen auf das individuelle Antwortverhalten ausübt. Dieser Annahme stehen empirische Untersuchungen zu Itemkontexteffekten gegenüber (Brennan 1992; Leary, und Dorans 1985). Ein besonders gut dokumentierter Itemkontexteffekt bezieht sich auf die Auswirkung der Itemposition. Itempositionseffekte führen in der Regel dazu, dass die Lösungswahrscheinlichkeit eines Items mit zunehmender Nähe zum Ende des Tests sinkt (z. B. Meyers et al. 2009). Bisherige Studien haben wiederholt Itempositionseffekte in den PISA-Assessments dokumentiert (Debeer et al. 2014; Hartig und Buchholz 2012; Wu 2010). Diese Effekte werden als Indikatoren einer abnehmenden Persistenz der Testbearbeitung diskutiert (z. B. Debeer et al. 2014).

Bei der Untersuchung von Itempositionseffekten lassen sich zwei Perspektiven unterscheiden. In den meisten Untersuchungen wird dieses Phänomen im Sinne eines festen Effekts interpretiert, dessen Ausprägung nicht zwischen Personengruppen und/oder Messzeitpunkten variiert (z. B. Meyers et al. 2009). In einer solchen Situation können die unerwünschten Auswirkungen von Itempositionseffekten in Längsschnittstudien durch die Konstanthaltung der Itempositionen vermieden werden. Eine zweite Perspektive betrachtet die Ausprägung von Itempositionseffekten als prinzipiell variabel (Debeer und Janssen 2013), sodass sich die Höhe dieser Effekte zwischen Gruppen und/oder Messzeitpunkten unterscheiden kann. Tatsächlich haben einige aktuelle Studien Hinweise dafür erbracht, dass Itempositionseffekte in leistungsschwächeren Schülergruppen stärker ausfallen (Debeer et al. 2014; Hartig und Buchholz 2012). Eine Verstärkung von Itempositionseffekten erscheint auch im Kontext von wiederholten Testdarbietungen plausibel, da sich die wiederholte umfangreiche Testung (der PISA-Test dauert insgesamt zwei Stunden) gerade in low-stakes Situationen negativ auf die Testmotivation auswirken könnte (DeMars 2007).

2 Zielsetzung und Schritte der Längsschnittskalierung

In den nachfolgenden Abschnitten berichten wir detaillierte Analysen, die sich (1) der längsschnittlichen Messinvarianz des im PISA-Längsschnitt eingesetzten Testmaterials, (2) der Robustheit der ermittelten Kompetenzzuwächse gegen Invarianzverstöße, (3) den Auswirkungen des selektiven Stichprobenausfalls auf die Zuwachsschätzungen und (4) der Schätzung von Komptenzniveaus und deren Veränderung widmen. Im ersten Schritt wurde das vorliegende Testmaterial dahingehend untersucht, ob Verstöße gegen die längsschnittliche Messinvarianzannahme vorliegen, die in Abhängigkeit von Merkmalen der Items auftreten und/oder auf differentiell wirkende Testkontexteffekte deuten. Vor dem Hintergrund der großen Bedeutung von Schulformvergleichen wurde zudem untersucht, inwieweit das Ausmaß möglicher Invarianzverstöße zwischen Schulformen variiert. Derartige Unterschiede sind möglich, da sich zum Beispiel die schulformspezifischen Curricula in ihren jahrgangsspezifischen Schwerpunksetzungen voneinander unterscheiden können (z. B. Klieme und Baumert 2001) und die wiederholte Testdurchführung in Schülergruppen mit einem geringeren Kompetenzniveau als besonders aversiv wahrgenommen werden könnte (DeMars 2007). Der zweite Schritt der Auswertungen bezog sich auf die Abschätzung möglicher Konsequenzen, wenn die im ersten Schritt ermittelten Invarianzverstöße nicht berücksichtigt werden. Die Ergebnisse dieser Analysen geben Auskunft über die Robustheit der ermittelten Kompetenzzuwächse gegenüber den vorliegenden Invarianzverstößen. Im dritten Schritt wurden die Einflüsse des selektiven längsschnittlichen Stichprobenausfalls im PISA-Längsschnitt 2012/2013 auf die Zuwachsschätzungen evaluiert. Hier stellte sich die Frage, inwieweit die vollständige Berücksichtigung des Ausfallsprozesses, der maßgeblich von allen in PISA getesteten Kompetenzdimensionen abhängt (Heine et al. 2017), die Ergebnisse der Zuwachsschätzungen verändert. Im vierten Schritt wurden schließlich die schülerseitigen Kompetenzniveaus geschätzt. In Übereinstimmung mit dem in den PISA-Assessments eingesetzten Verfahren haben wir die sogenannte Plausible-Value-Methode (PV; Mislevy et al. 1992) herangezogen, mit deren Hilfe die Abschätzung der Verteilung der Kompetenzniveaus optimiert werde kann.

2.1 Verstöße gegen die Annahme invarianter Itemparameter

Der Aufdeckung von Verstößen gegen das Kriterium der längsschnittlichen Messinvarianz lagen folgende Gedanken zugrunde. Invarianzverstöße, die in Abhängigkeit der Merkmale der einzelnen Items auftreten, sollten sich unabhängig vom Testkontext, in den die Items eingebettet sind, einstellen. Insofern sich nennenswerte Abweichungen ergeben, können diese mit Itemmerkmalen in Verbindung gebracht werden (z. B. Stoffgebiete, Lösungswege, Darstellungsformen usw.), um mögliche curriculare Einflüsse auf die Abweichungen zu identifizieren. Hingegen sollten sich Invarianzverstöße, die in Abhängigkeit des Testkontexts auftreten, die Items unabhängig von ihren spezifischen Merkmalen betreffen. Insofern die Daten derartige Abweichungen vom Invarianzkriterium indizieren, können diese in Abhängigkeit von ausgewählten Merkmalen des Testkontexts betrachtet werden. So können beispielsweise Abnahmen der Testpersistenz identifiziert werden, die sich in positionsspezi-

fischen Parameterabweichungen manifestieren (stärkere Abweichungen zum Ende des Tests hin). Die Untersuchung dieser Phänomene geschah mithilfe IRT-basierter Verfahren, wobei wir uns in Übereinstimmung mit den in PISA verwendeten Messmodellen auf das einparametrische Rasch-Modell (Rasch 1960) beschränkt haben.

2.2 Auswirkungen von Verstößen gegen die Annahme invarianter Itemparameter

Der zweite Schritt der Auswertungen untersuchte, wie sich die im ersten Schritt ermittelten Invarianzverstöße auf die Abschätzung der Kompetenzzuwächse auswirken. Zu diesem Zweck wurden die Ergebnisse unterschiedlicher IRT-Modelle miteinander verglichen, die sich im Grad der angenommenen Messinvarianz voneinander unterschieden. Als Referenzmodell wurde ein zweidimensionales IRT-Modell mit jeweils einer Dimension pro Messzeitpunkt und vollständig invarianten Itemparametern verwendet (z. B. von Davier et al. 2011). Die Ergebnisse dieses Modells wurden mit den Schätzungen eines zweiten Modells verglichen, in dem die im ersten Schritt ermittelten Invarianzverstöße berücksichtigt wurden (d. h. partielle Invarianz unter Berücksichtigung von Invarianzverstößen in Abhängigkeit der Itemmerkmale oder der Itemkontexte). Diese Ergebnisse liefern Hinweise für die Abhängigkeit der Zuwachsschätzungen gegenüber den zuvor ermittelten Verstößen gegen das Invarianzkriterium. Ein Befund, wonach sich die Zuwachsschätzungen deutlich zwischen den konkurrierenden Modellen unterscheiden, indiziert, dass die Ignorierung der Abweichungen vom Idealkriterium der vollständigen Parameterinvarianz zu verzerrten Zuwachsschätzungen führt. Eine solche Situation legt somit nahe, dass die Parameterabweichungen in einem partiell invarianten Messmodell berücksichtigt werden müssen.

2.3 Selektiver Stichprobenausfall und Schätzung von Kompetenzzuwächsen

Die bis zu dieser Stelle skizzierten Auswertungen wurden gesondert für jede der im PISA-Längsschnitt 2012/2013 getesteten Domänen durchgeführt (Mathematik, Leseverständnis und Naturwissenschaften). Diese Analysen basieren somit unter Umständen auf einer unzureichenden Berücksichtigung des selektiven Stichprobenausfalls, der im PISA-Längsschnitt 2012/2013 maßgeblich von allen getesteten Kompetenzdimensionen abhängt (Heine et al. 2017). Um die Auswirkungen der in Abhängigkeit aller Kompetenzdimensionen stattfinden Stichprobenausfalls auf die Abschätzung der Kompetenzentwicklung zu untersuchen, wurden die zuvor beschriebenen Messmodelle um alle zum ersten Messzeitpunkt erhobenen Kompetenzbereiche erweitert. Insofern alle Kompetenzbereiche simultan mit dem Ausfallprozess assoziiert sind, ist zu erwarten, dass ihre Berücksichtigung zu einer Verringerung der Zuwachsschätzung führt, da niedrigere Kompetenzstände zum ersten Messzeitpunkt mit einer höheren Ausfallwahrscheinlichkeit einhergehen (Heine et al. 2017).

2.4 Schätzung der Kompetenzniveaus

Die Schätzung der schülerseitigen Kompetenzniveaus fand im letzten Schritt statt. Zu diesem Zweck wurde ausgehend von den in den vorangegangen Teilschritten erzielten Ergebnissen ein mehrdimensionales Messmodell spezifiziert, das alle im PISA-Längsschnitt getesteten Kompetenzdimensionen zu beiden Messzeitpunkten beinhaltete (3 Kompetenzbereiche × 2 Messzeitpunkte = 6 Dimensionen). Die Skalierung der PISA-Tests geschah auf Grundlage der zuvor rekalibrierten Itemparameter, die mögliche Invarianzverstöße berücksichtigten (partielle Messinvarianz).

Tab. 1 Booklet-Design der PISA-2012–2013 Längsschnittstudien. Grau unterlegte Zellen indizieren nicht berücksichtigte Items

	Position 1	Position 2	Position 3	Position 4
Messzeitpunkt 1				
Booklet 01	M5	N3	M6	N2
Booklet 02	N3	L3	M7	L2
Booklet 03	L3	M6	N1	M3
Booklet 04	M6	M7	L1	M4
Booklet 05	M7	N1	M1	M5
Booklet 06	M1	M2	L2	M6
Booklet 07	M2	N2	M3	M7
Booklet 08	N2	L2	M4	N1
Booklet 09	L2	M3	M5	L1
Booklet 10	M3	M4	N3	M1
Booklet 11	M4	M5	L3	M2
Booklet 12	N1	L1	M2	N3
Booklet 13	L1	M1	N2	L3
Messzeitpunkt 2				
Booklet 01	M2	N2	M3	M7
Booklet 02	M4	M5	L3	M2
Booklet 03	N2	L3	L2	N3
Booklet 06	M7	L2	–	–
Booklet 07	–	–	M2	MR
Booklet 08	–	–	M7	MR
Booklet 09	M2	MR	–	–
Booklet 10	N3	M4	–	–
Booklet 11	–	–	L3	N2
Booklet 12	–	–	M4	L3
Booklet 14	M4	N3	–	–
Booklet 15	–	–	MR	M4
Booklet 16	L2	M7	–	–
Booklet 17	MR	M7	N3	L2
Booklet 18	–	–	N2	M2

M Mathematik (*MR* Itemcluster nur in Retesterhebung verwendet), *N* Naturwissenschaften, *L* Leseverständnis

Die Schätzung der Kompetenzniveaus geschah mithilfe der PV-Methode. Ihr Einsatz liegt in dem Umstand begründet, dass die Schülerinnen und Schüler pro Inhaltsdomäne jeweils nur einen kleinen Teil der eingesetzten Testaufgaben bearbeiten (vgl. Multi-Matrix-Design in Tab. 1). Hinzu kommt, dass der PISA-Längsschnitt 2012/2013 von selektiven Stichprobenausfällen betroffen ist (Heine et al. 2017). Im Gegensatz zu vielen anderen Methoden der Kompetenzschätzung ermöglicht es die PV-Methode, die Kompetenzniveaus auch derjenigen Schülerinnen und Schüler abzuschätzen, die an der zweiten Erhebung im Jahr 2013 nicht teilgenommen haben.

Bei der PV-Methode werden neben den Itemantworten weitere Variablen berücksichtigt, die mit den Kompetenzniveaus und dem Ausfallprozess assoziiert sind (z. B. Kompetenztests in anderen Domänen, familiäre Hintergrundmerkmale, usw.). Eine Reihe von Studien konnte zeigen, dass die PV-Methode eine hervorragende Abbildung der Populationsverteilung der Kompetenzen sowie deren Zusammenhänge mit anderen Merkmalen bietet (z. B. Wu 2005). Im vorliegenden Fall wurden bei der Generierung der PVs neben den Itemantworten ein breiter Kranz von Hintergundvariablen und alle Kompetenzdimensionen berücksichtigt. Konkret haben wir alle Schülerinformationen herangezogen, die auch in der Skalierung der Leistungstests des PISA-Assessments 2012 verwendet wurden (Berücksichtigung von 95 % der Variabilität in allen verfügbaren Schülerinformationen; OECD 2014).

3 Methode

3.1 Datengrundlage der Itemkalibrierung

In die Analysen gingen die im Rahmen von PISA entwickelten Items ein. Das zur ersten Welle verwendete Testdesign entsprach dem in der internationalen PISA-2012-Studie verwendetem Bookletdesign (oberer Teil der Tab. 1). Ein Teil dieser Items wurde den Schülerinnen und Schülern zum zweiten Messzeitpunkt vorgelegt (unterer Teil der Tab. 1). Das Testdesign der Retesterhebung umfasste zudem weitere Items, die nicht dem PISA-Framework entstammten und aus diesem Grund nicht in die Analysen eingingen (offene Felder in Tab. 1). Zudem wurde in der zweiten Erhebung eine Gruppe von Items zum Bereich Mathematik eingesetzt (Cluster MR in Tab. 1), die einem vorangegangenen PISA-Zyklus entnommen wurde (Ramm et al. 2006). Das bedeutet, dass mit Ausnahme des Clusters MR (vgl. Tab. 1), die zum zweiten Messzeitpunkt verwendeten Cluster eine Teilmenge der zum ersten Messzeitpunkt eingesetzten Itemcluster darstellen. Das Bookletdesign ist in Tab. 1 zusammengefasst. In Tab. 1 werden nur die Testbooklets des zweiten Messzeitpunkts dargestellt, die tatsächlich PISA-Items beinhalteten und somit in die Analysen eingingen.

Zur Erstellung der Testhefte wurden die Items zuerst in sogenannten Itemclustern zusammengefasst (M1, M2, usw.), die eine Bearbeitungsdauer von jeweils 30 min beanspruchten. Items wurden jeweils einem einzigen Cluster zugeordnet. Das zum ersten Messzeitpunkt verwendete Bookletdesign wurde anhand eines Youden-Square-Designs konstruiert (Frey et al. 2009). Dieses Konstruktionsprinzip stellte sicher, dass jedes Itemcluster exakt einmal an jeder der vier möglichen Positio-

nen eines Testhefts auftrat und dass jede Paarung zweier Cluster (d. h. gemeinsame Darbietung in einem Booklet unabhängig von deren Reihenfolge im Booklet) exakt einmal umgesetzt wurde. In diesem Youden-Square-Design sind die Itemcluster hinsichtlich der Positionen in den Testheften balanciert, wobei die Position der Items innerhalb der Cluster in allen Booklets unverändert beibehalten wurde. Die Paarung der Itemcluster soll sicherstellen, dass die Kovarianzen zwischen den Kompetenzbereichen abgebildet werden können.

In den zum zweiten Messzeitpunkt verwendeten Bookletdesign wurden 18 Testhefte eingesetzt, wobei eine perfekte Balance der Itempositionen und eine vollständige Paarung aller Itemcluster sich aufgrund des zusätzlich eingefügten Testmaterials (offene Felder in Tab. 1) als nicht praktikabel erwiesen. Nichtsdestotrotz wurde bei der Konstruktion des Bookletdesigns der Retesterhebung darauf geachtet, die mittlere Position der zu einer Domäne gehörenden Itemcluster vergleichbar zur ersten Erhebung zu halten. Die über alle Itemcluster gemittelte Itemclusterposition betrug für Mathematik 2,4 und Naturwissenschaften 2,5. Die Position der Itemcluster zum Bereich Leseverständnis verschob sich jedoch in der zweiten Erhebungswelle leicht in Richtung Testende (2,8). Das bedeutet, dass selbst bei einem zeitlich konstanten Itempositionseffekt mit einer Unterschätzung der Lesekompetenzen zum zweiten Messzeitpunkt sowie der entsprechenden Zuwächse zu rechnen ist.

Ausgehend von der im Beitrag von Heine et al. (2017) beschriebenen Längsschnittstichprobe von $N = 6584$ Schülerinnen und Schülern, die sich auf das Gymnasium (45,2 %) und nichtgymnasiale Schulformen (54,8 %) verteilten, gingen in die Analysen zur Itemkalibrierung jeweils die folgenden Fallzahlen für die drei Domänen ein: $N = 6359$ (Mathematik), $N = 4954$ (Leseverständnis) und $N = 4930$ (Naturwissenschaften). Die Reduktion der Fallzahlen in den letzten beiden Domänen liegt in dem Umstand begründet, dass einige Schülerinnen und Schüler aufgrund der randomisierten Zuweisung der Testhefte zu keinem der beiden Messzeitpunkte Aufgaben zu einem Bereich bearbeitet haben.

3.2 Messmodelle

Die Untersuchung der Messäquivalenz erfolgte im Rahmen von mehrdimensionalen Rasch-Modellen, in denen die Items zum ersten und zweiten Zeitpunkt als Indikatoren separater Kompetenzdimensionen herangezogen wurden (vgl. von Davier et al. 2011). Die entsprechenden Modelle wurden als Mehrguppenanalysen mit Schulform (Gymnasien vs. andere Schulformen) als Gruppierungsvariable aufgesetzt. Mehrkategoriale Items, die auch teilweise gelöst werden konnten, wurden entsprechend dem Partial-Credit-Modell (PCM; Masters 1982) modelliert. Die Analysen zur Itemkalibrierung wurden mit dem Statistikprogramm Mplus 7.2 (Muthén & Muthén, 1998–2012) durchgeführt, wobei das PCM mittels restringierter multinomial logistischer Regressionen, die dem klassischen PCM nach Masters (1982) entsprechen, spezifiziert wurde (vgl. Muraki 1992). Aufgrund der Clusterung der Individualdaten in Schulen wurde die hierarchische Datenstruktur mittels der Analysefunktion „type = complex“ im Rahmen einer robusten Maximum-Likelihood-Schätzung (MLR) berücksichtigt.

Wie in den vorangegangenen Abschnitten erläutert, können Verstöße gegen die Annahme längsschnittlicher (und schulformbezogener) Messinvarianz im Prinzip in Abhängigkeit von Itemmerkmalen (z. B. im Fall jahrgangsspezifischer curricularer Schwerpunktsetzungen) und/oder von Testkontextmerkmalen (z. B. Positionen in den Booklets) auftreten. Im ersten Fall treten Verstöße gegen die Invarianzannahme unabhängig von den vorgelegten Booklets auf. Im zweiten Fall hängen die Invarianzverstöße vom bearbeiteten Booklet ab und können beispielsweise indikativ für quer- und längsschnittliche Invarianzverstöße aufgrund von systematischen Unterschieden in der Persistenz bei der Bearbeitung der Tests sein.[1]

Invarianzverstöße in Abhängigkeit der Itemmerkmale. Die Untersuchung von Invarianzverstößen aufgrund der Inhalte der Items fand ohne explizite Berücksichtigung der Testkonfiguration statt. Ebenso haben wir auf eine Zusammenfassung von Items in inhaltsbezogenen Gruppen (z. B. naturwissenschaftliche Teilfächer, graphische Darstellungen, usw.) verzichtet. Stattdessen haben wir Schwierigkeiten eines jeden Items in jeder Schulform und zu jedem Messzeitpunkt anhand der Daten geschätzt und die Übereinstimmung der empirischen Lösungen über Gruppen und Messzeitpunkte evaluiert. Zu diesem Zweck wurden für jede Inhaltsdomäne zweidimensionale Raschmodelle auf die Daten angepasst. Die Modellschätzung fand unter Berücksichtigung der Gruppierungsvariable Schulform statt, wobei für jede Kombination aus Messzeitpunkt und Schulform spezifische Itemparameter geschätzt wurden. Dichotome Items wurden wie folgt modelliert:

$$P\left(y_{ijtg}=1\right)=\frac{exp\left(\theta_{it}-\beta_{jtg}\right)}{1+exp\left(\theta_{it}-\beta_{jtg}\right)}, \tag{1}$$

wobei y_{ijtg} für die beobachtete Itemantwort des Individuums i (i = 1, 2, ..., N) auf Item j (j = 1, 2, ..., J) zu Messzeitpunkt t (t = 1, 2) in Gruppe g (g = 1, 2) steht, β_{jtg} die Schwierigkeit des Items j zu Messzeitpunkt t in Gruppe g ausweist und θ_{it} die individuelle Ausprägung der Fähigkeitsvariable des Individuums i zu Messzeitpunkt t bezeichnet.

Partial-Credit-Items wurden entsprechend dem PCM mit nicht restringierten Itemschwierigkeiten und minimal restringierten Step-Parametern (Summennormierung; s. unten) modelliert:

$$P\left(y_{ijtg}=m\right)=\frac{exp\left[m\left(\theta_{it}-\beta_{jtg}\right)-\sum_{h=0}^{\mathrm{m}}\delta_{jhtg}\right]}{\sum_{k=0}^{M_j}exp\left[k\left(\theta_{it}-\beta_{jtg}\right)-\sum_{h=0}^{\mathrm{k}}\delta_{jhtg}\right]}. \tag{2}$$

Gl. 2 weist das Modell für die Wahrscheinlichkeit einer Antwort in Kategorie m eines Items j mit M_j Kategorien aus. Die δ-Parameter sind im Sinne sogenannter

[1] An dieser Stelle gilt anzumerken, dass die Clusterposition auch mit der Domänenabfolge konfundiert ist, die sich ebenso im Sinne eines Testkontexteffekts interpretieren lässt (Brennan 1992; Harris 1991).

Step-Parameter zu verstehen, die sich für jedes Item zu 0 aufsummieren, jedoch zu beiden Messzeitpunkten und in beiden Gruppen ansonsten frei geschätzt wurden.

Für das in den Gln. 1 und 2 ausgewiesene Modell wurden alle β-Parameter frei geschätzt. Um deren Identifikation zu gewährleisten, wurden die θ-Variablen in jeder Gruppe und zu jedem Messzeitpunkt auf einen Mittelwert von 0 fixiert. Diese Parametrisierung erlaubt nicht die Erfassung von Leistungszuwächsen, kann jedoch zur Evaluation von differentiellen Verstößen gegen die Messinvarianzannahmen herangezogen werden.

Im vorliegenden Beitrag haben wir uns graphischer Prozeduren bedient (siehe Bejar 1980; Nagy und Neumann 2010). Hierzu haben wir zuerst die frei geschätzten β-Parameter gegeneinander geplottet. Abweichungen von der Annahme der Messinvarianz aufgrund von Inhaltsmerkmalen der Items manifestieren sich in Abweichungen der in einem Streudiagramm dargestellten β-Parameter von einer perfekten linearen Funktion (mit Einheitssteigung). Ein Befundmuster, in dem alle Items einer linearen Funktion entsprechen, impliziert, dass die Daten hinreichend gut durch ein Messmodell mit vollständig invarianten Itemparametern beschrieben werden können (d. h. $\beta_{jtg} = \beta_j$ für alle t und g).

In empirischen Anwendungen ist davon auszugehen, dass selbst bei einer hohen Übereinstimmung der Parameter β_{jtg} zwischen Messzeitpunkten t und/oder Gruppen g, die Abweichungen von einer linearen Funktion statistisch signifikant ausfallen. Aus einer inhaltlichen Perspektive ist jedoch die statistische Signifikanz der Gesamtheit der Abweichungen weniger von Bedeutung, als die Frage nach der praktischen Signifikanz der Abweichungen vom idealtypischen Zusammenhang. Um diese zu evaluieren, wurden zum einen die Übereinstimmung der Itemparameter zwischen Gruppen und Messzeitpunkten graphisch evaluiert (s. oben). Diese Analyse diente der Identifikation von Gruppen von Items mit systematischen (d. h. gemeinsamen) Abweichungen vom idealtypischen Zusammenhang. Zum anderen wurden Statistiken des „differential item functioning“ (DIF; Holland und Wainer 1993) berechnet und gemäß des vom Educational Testing Service (ETS) vorgeschlagenen Systems (Clauser und Mazor 1998) klassifiziert: (A) nicht von DIF betroffen, (B) kaum von DIF betroffen und (C) stark von DIF betroffen.[2]

Invarianzverstöße in Abhängigkeit von Testkontexten. Invarianzverstöße aufgrund von Testkontexteffekten wurden auf Ebene der Itemcluster untersucht. Hierzu wurden Unterschiede in den Schwierigkeiten der Items eines Clusters in Abhängig-

[2] DIF-Statistiken wurden anhand der Ergebnisse des in den Gln. 1 und 2 dargestellten Modells berechnet. Die Übereinstimmung der Itemparameter zwischen Gruppen und innerhalb von Messzeitpunkten wurde berechnet als $DIF_{jt} = \left(\beta_{jt(g=1)} - \overline{\beta}_{t(g=1)}\right) - \left(\beta_{jt(g=2)} - \overline{\beta}_{t(g=2)}\right)$. Die Übereinstimmung der Itemparameter zwischen Messzeitpunkten innerhalb von Gruppen wurde berechnet als $DIF_{jg} = \left(\beta_{j(t=1)g} - \overline{\beta}_{(t=1)g}\right) - \left(\beta_{j(t=2)g} - \overline{\beta}_{(t=2)g}\right)$. In beiden Fällen steht $\overline{\beta}_{tg}$ für die mittlere Itemschwierigkeit in Gruppe g zu Messzeitpunkt t, wobei in die Berechnung dieser Größen unterschiedlich viele Items eingingen, da sich die Zahl gemeinsamer Items zwischen den Analysen unterschied. Items wurden der DIF-Kategorie A (nicht von DIF betroffen) zugeordnet, wenn sich die DIF-Statistiken nicht statistisch signifikant von 0 unterschieden. Items der DIF-Kategorie B (kaum von DIF betroffen) wiesen DIF-Werte auf, die sich zwar signifikant von 0, aber nicht signifikant von |0,4| unterschieden (approximative Mantel-Haenszel-Effektstärke, die signifikant einen Wert von |1,0| übersteigt; vgl. Nagy und Neumann 2010). Alle übrigen Items wurden der DIF-Kategorie C (stark von DIF betroffen) zugeordnet.

keit vom Testbooklet, vom Messzeitpunkt und von der Schulform betrachtet. Die Antwort des Individuums i auf dem Item j, das in Cluster c in Booklet b zum Messzeitpunkt t in der Schulform g dargeboten wurde (y_{ijcbtg}), wurde im Fall dichotomer Items entsprechend einem erweiterten Raschmodell modelliert:

$$P\left(y_{ijcbtg}=1\right)=\frac{exp\left[\theta_{it}-\left(\beta_j+\gamma_{cbtg}\right)\right]}{1+exp\left[\theta_{it}-\left(\beta_j+\gamma_{cbtg}\right)\right]}. \tag{3}$$

In diesem Modell beschreibt der Parameter γ_{cbtg} den Effekt eines Booklets b auf die Schwierigkeiten der Items in Cluster c zu Zeitpunkt t in Gruppe g.[3] Aus Gl. 3 wird ersichtlich, dass die Ausprägungen der Schwierigkeiten der Items eines Clusters als invariant spezifiziert wurden, wobei aber Verschiebungen der mittleren Itemschwierigkeiten zwischen Booklets, Messzeitpunkten und Schulformen zugelassen wurden. Die Invarianz der Itemparameter innerhalb eines Clusters ist eine zentrale Annahme des hier vorgestellten Modells, die vor dessen Anwendung mittels graphischer Verfahren überprüft wurde (s. oben).

Partial-Credit-Items wurden analog zu Gl. 3 modelliert, wobei die abhängige Variable nun die Wahrscheinlichkeit des Vorliegens einer Antwortkategorie m eines Items j mit M_j Kategorien darstellte:

$$P\left(y_{ijcbtg}=m\right)=\frac{exp\left\{m\left[\theta_{it}-\left(\beta_j+\gamma_{cbtg}\right)\right]-\sum_{h=0}^{\mathrm{m}}\delta_{jh}\right\}}{\sum_{k=0}^{M_j}exp\left\{k\left[\theta_{it}-\left(\beta_j+\gamma_{cbtg}\right)\right]-\sum_{h=0}^{\mathrm{k}}\delta_{jh}\right\}}. \tag{4}$$

Die in Gl. 4 aufgeführten Step-Parameter wurden als invariant zwischen Booklets, Messzeitpunkten und Schulformen modelliert. Das bedeutet, dass in diesem Modell die Effekte der Booklets auf die Schwierigkeiten der Partial-Credit-Items beschränkt sind und nicht die Step-Parameter betreffen. Wir haben uns für diese Modellvariante entschieden, da insgesamt nur sehr wenige Items ein Partial-Credit-Format aufwiesen und Effekte auf die Step-Parameter schwierig zu interpretieren sind.

Um die Identifikation des in den Gln. 3 und 4 dargestellten Modells sicherzustellen, wurden alle γ-Parameter (d. h. Bookleteffekte), die sich auf die erste Clusterposition bezogen (z. B. Cluster M5 in Booklet 1 zu Messzeitpunkt 1) auf 0 fixiert. Diese Identifikationsstrategie impliziert die Verankerung der Metrik der θ-Variable in Referenz zur ersten Clusterposition. Diese Verankerung ist insofern sinnvoll, da die so präsentierten Items per Definition nicht von Effekten der jeweiligen Testkonfiguration betroffen sind. Die anhand des in den Gln. 3 und 4 beschriebenen Modells kalibrierten Itemparameter erlauben die Erfassung von Testleistungen und deren Zuwächse in Referenz zur ersten Clusterposition.

[3] Die γ-Parameter wurden mithilfe einer einfachen bzw. multinomialen logistischen Regression der beobachteten Itemantworten auf die dummykodierten Bookletindikatoren bestimmt.

4 Ergebnisse

4.1 Ergebnisse der Itemkalibrierung

Invarianzverstöße in Abhängigkeit von Itemmerkmalen. Der erste Schritt der Untersuchung der Messinvarianz der PISA-Instrumente basierte auf der Schätzung der Itemparameter anhand des in den Gln. 1 und 2 beschriebenen Modells. Die Ergebnisse der Analysen sind summarisch in Tab. 2 zusammengefasst. Die Analysen erbrachten hohe Korrelationen der frei geschätzten Itemparameter zwischen Schulformen innerhalb von Messzeitpunkten (Schulform-DIF in Tab. 2) und zwischen Messzeitpunkten innerhalb von Schulformen (Längsschnitt-DIF in Tab. 2), die

Tab. 2 DIF-Statistiken innerhalb der Messzeitpunkte und zwischen Schulformen (Schulform-DIF), sowie innerhalb von Schulformen zwischen Messzeitpunkten (Längsschnitt-DIF): Korrelation der Itemparameter, mittlerer absoluter DIF (Median in Klammern) und Items nach DIF-Kategorie (ETS-Klassifikation)[A]

	Mathematik		Leseverständnis		Naturwissenschaften	
Schulform-DIF (Erhebung 2012)						
Korrelation	0,97	–	0,98	–	0,96	–
$M_{\|DIF\|}$ ($Mdn_{\|DIF\|}$)	0,33	(0,28)	0,25	(0,22)	0,25	(0,24)
DIF-Kat. A (%)	39	(45,0 %)	26	(60,5 %)	27	(51,9 %)
DIF-Kat. B (%)	35	(42,2 %)	15	(34,9 %)	24	(46,2 %)
DIF-Kat. C (%)	9	(10,8 %)	2	(4,7 %)	1	(1,9 %)
Schulform-DIF (Erhebung 2013)						
Korrelation	0,95	–	0,98	–	0,95	–
$M_{\|DIF\|}$ ($Mdn_{\|DIF\|}$)	0,41	(0,37)	0,26	(0,27)	0,29	(0,24)
DIF-Kat. A (%)	39	(52,0 %)	21	(72,4 %)	23	(67,6 %)
DIF-Kat. B (%)	33	(44,0 %)	7	(24,1 %)	11	(32,4 %)
DIF-Kat. C (%)	3	(4,0 %)	1	(3,5 %)	0	(0,0 %)
Längsschnitt-DIF (Nicht-gymnasiale Schulformen)						
Korrelation	0,98	–	0,99	–	0,98	–
$M_{\|DIF\|}$ ($Mdn_{\|DIF\|}$)	0,22	(0,16)	0,20	(0,17)	0,19	(0,13)
DIF-Kat. A (%)	42	(72,4 %)	19	(65,5 %)	24	(70,6 %)
DIF-Kat. B (%)	16	(27,6 %)	10	(34,5 %)	10	(29,4 %)
DIF-Kat. C (%)	0	(0,0 %)	0	(0,0 %)	0	(0,0 %)
Längsschnit-DIF (Gymnasien)						
Korrelation	0,97	–	0,99	–	0,99	–
$M_{\|DIF\|}$ ($Mdn_{\|DIF\|}$)	0,25	(0,19)	0,25	(0,19)	0,14	(0,11)
DIF-Kat. A (%)	45	(77,6 %)	19	(65,5 %)	31	(91,2 %)
DIF-Kat. B (%)	13	(22,4 %)	10	(34,5 %)	3	(8,8 %)
DIF-Kat. C (%)	0	(0,0 %)	0	(0,0 %)	0	(0,0 %)

DIF-Kat. A nicht von DIF betroffen, *DIF-Kat. B* kaum von DIF betroffen, *DIF-Kat. C* stark von DIF betroffen

einen Wertebereich von $r = 0{,}95$ bis $r = 0{,}99$ umspannten. Die Mittelwerte (Mediane in Klammern) der absoluten DIF-Statistiken bewegten sich insgesamt unterhalb der vom ETS definierten Schwelle für kaum von DIF betroffene Items ($|DIF| < 0{,}4$). Eine Ausnahme von diesem Muster fand sich jedoch in der Erhebung 2013 für den Schulform-DIF im Bereich Mathematik. Hier lag der mittlere absolute DIF nahezu exakt auf der kritischen Schwelle von $|DIF| = 0{,}4$. Insgesamt wurde in allen Vergleichen der größte Anteil der Items als nicht von DIF betroffen klassifiziert, wobei in Abhängigkeit der Kompetenzdomänen und der Vergleiche ein nennenswerter Anteil von Items als gering von DIF betroffen klassifiziert wurde. Dies betraf vorwiegend die querschnittlichen Vergleiche der Mathematikitems (Schulform-DIF). Hier lag der Anteil der gering von DIF betroffenen Items nur knapp unter dem Anteil der nicht von DIF betroffenen Aufgaben. Die in Tab. 2 berichteten Ergebnisse legten somit insgesamt nahe, dass die auf Ebene der Einzelitems vorliegenden Verstöße gegen die quer- und längsschnittliche Invarianzanahme in den Bereichen Lesen und Naturwissenschaften insgesamt so gering ausfielen, dass deren Ausblendung keine praktischen Konsequenzen nach sich zieht. Für den Bereich Mathematik konnten jedoch auf Grundlage der aggregierten Befunde Konsequenzen für querschnittliche Schulformvergleiche nicht unbedingt ausgeschlossen werden.

Die detaillierten Ergebnisse der Modellschätzungen sind in den Abb. 1 bis 3 getrennt für die Bereiche Mathematik, Naturwissenschaften und Leseverständnis ausgewiesen. In den entsprechenden Teilabbildungen sind jeweils bivariate Streudiagramme der Itemschwierigkeiten getrennt nach Itemcluster dargestellt. Die Streudiagramme stellen zum einen die Übereinstimmung der Itemparameter über Schulformen (getrennt nach Erhebungszeitpunkte) und über Erhebungszeitpunkte (getrennt nach Schulformen) dar. Ein hinreichender Grad von Messinvarianz ist dann indiziert, wenn alle Itemschwierigkeiten nahe an einer Linie mit einer Steigung von eins liegen. Die Abb. 1 bis 3 dokumentieren, dass dieses Kriterium (approximativ) für alle Items erfüllt wurde.

Nennenswerte Abweichungen fanden sich nur für ein Item aus dem Bereich Mathematik (Cluster M5), das zum zweiten Messzeitpunkt in beiden Gruppen wider Erwarten schwer ausfiel.[4] Davon abgesehen fanden sich für den Bereich Mathematik keine Hinweise für nennenswerte Abweichungen von der querschnittlichen und längsschnittlichen Invarianzannahme. Der Zusammenhang zwischen den Schwierigkeiten der Items in den Clustern M3 und M5 erscheint zunächst zum zweiten Messzeitpunkt etwas geringer als zum ersten Messzeitpunkt. Diese Abweichung, die maßgeblich für die vergleichsweise hohen absoluten DIF-Werte in dieser Domäne verantwortlich ist (vgl. Tab. 2), lässt sich jedoch auf die geringere Präzision der Parameterschätzungen zurückführen, da diese Items nur von rund 5,6 % der zum zweiten Messzeitpunkt teilnehmenden Schülerinnen und Schülern bearbeitet wurden. Somit lieferten die in Abb. 1 dargestellten Detailergebnisse keine Hinweise für systematische Verstöße gegen die Invarianzannahme auf Ebene der Einzelitems des Bereichs Mathematik.

[4] Dieses Item wurde in den nachfolgenden Skalierungsgängen unter Lockerung der längsschnittlichen Invarianzannahme beibehalten.

Erhebung 2012: Nicht-Gymnasiale Schulformen (X) und Gymnasien (Y)	Erhebung 2013: Nicht-Gymnasiale Schulformen (X) und Gymnasien (Y)	Nicht-Gymnasiale Schulformen: Erhebung 2012 (X) und Erhebung 2013 (Y)	Gymnasien: Erhebung 2012 (X) und Erhebung 2013 (Y)

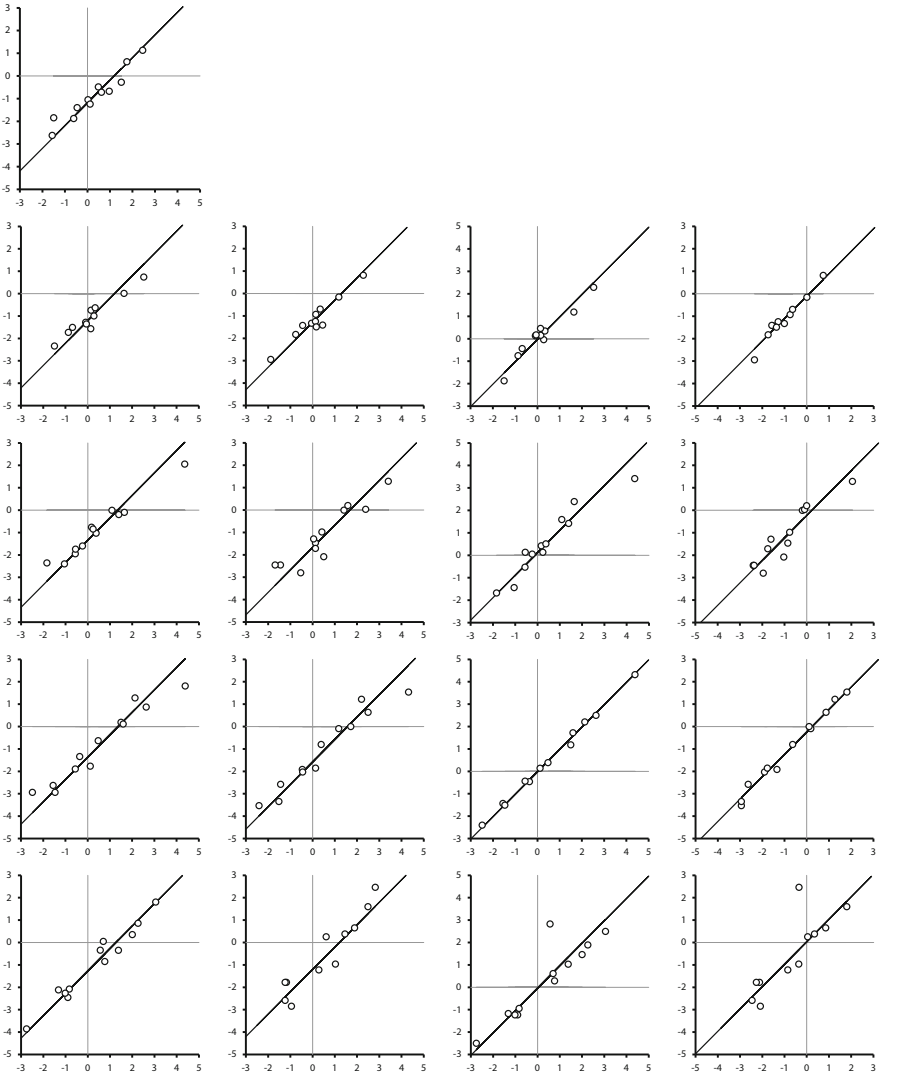

Abb. 1 Bivariate Streudiagramme frei geschätzter Itemparameter für den Bereich Mathematik getrennt nach Itemcluster (*Anmerkungen*: Anordnung der Itemcluster nach Zeilen: M1, M2, M3, M4, M5, M6, M7 und MR [nur in Retesterhebung verwendet])

Erhebung 2012: Nicht-Gymnasiale Schulformen (X) und Gymnasien (Y)	Erhebung 2013: Nicht-Gymnasiale Schulformen (X) und Gymnasien (Y)	Nicht-Gymnasiale Schulformen: Erhebung 2012 (X) und Erhebung 2013 (Y)	Gymnasien: Erhebung 2012 (X) und Erhebung 2013 (Y)

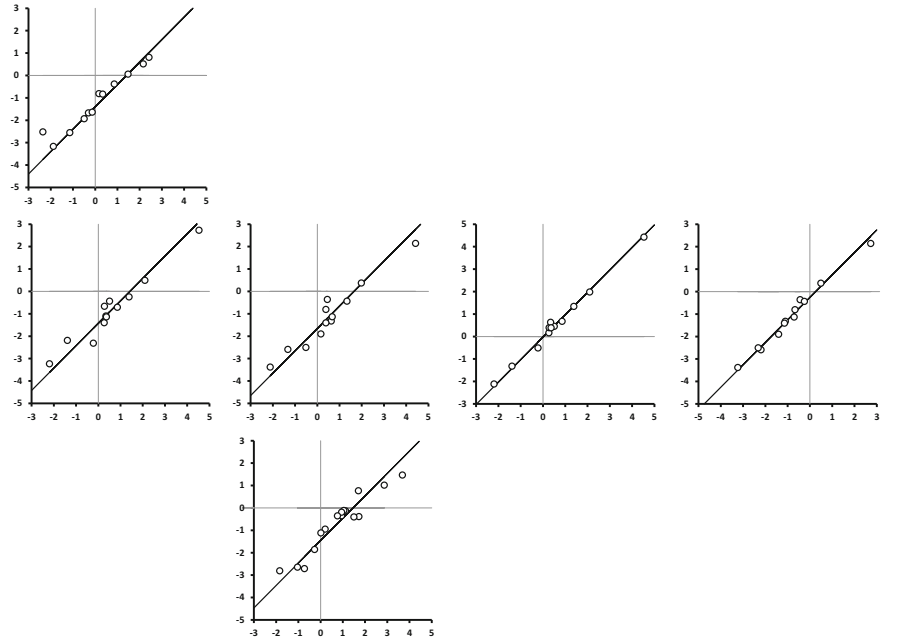

Abb. 1 (Fortsetzung) Bivariate Streudiagramme frei geschätzter Itemparameter für den Bereich Mathematik getrennt nach Itemcluster (*Anmerkungen*: Anordnung der Itemcluster nach Zeilen: M1, M2, M3, M4, M5, M6, M7 und MR [nur in Retesterhebung verwendet])

Für die Bereiche Naturwissenschaften und Leseverständnis ließ sich insgesamt eine sehr hohe Entsprechung der empirischen Lösungen zum Kriterium der absoluten Messinvarianz feststellen (Abb. 2 und 3), die sich mit den auf Grundlage der in Tab. 2 berichteten Ergebnisse erzielten Einschätzungen decken. Die hier beschriebenen Ergebnisse deuten somit in der Gesamtschau darauf hin, dass keine nennenswerten Invarianzverstöße in Abhängigkeit von Itemmerkmalen vorlagen.

Invarianzverstöße in Abhängigkeit von Testkontexten. Die Untersuchung dieser Effekte erfolgte auf Grundlage der zuvor beschriebenen erweiterten Messmodelle (Gln. 3 und 4). Die Ergebnisse der Analysen sind in Abb. 4 graphisch in Form von Streudiagrammen mit der Itemclusterposition auf der x-Achse zusammengefasst. Die offenen weißen Punkte repräsentieren die Bookleteffekte zum ersten Messzeitpunkt. Diese Effekte wurden für Itemcluster, die an der ersten Position präsentiert wurden, auf 0 fixiert. Die grau eingefärbten Punkte stehen für die Summe der Leistungszuwächse auf der ersten Itemclusterposition und der zum zweiten Messzeitpunkt ermittelten Bookleteffekte. Die für jede Itemclusterposition über die Testhefte ge-

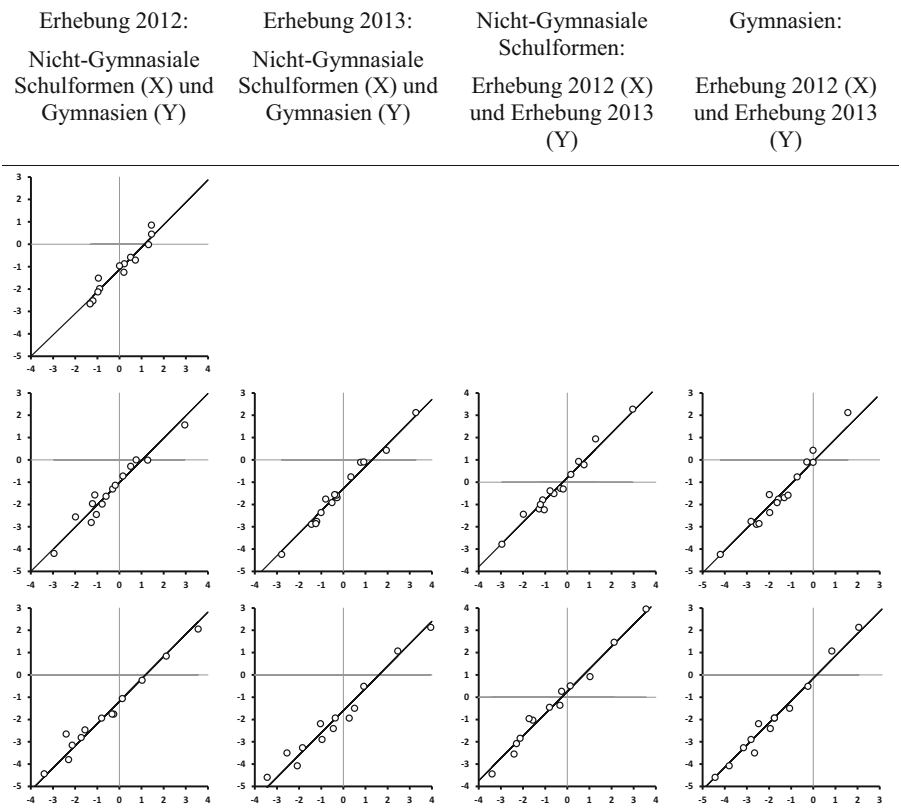

Abb. 2 Bivariate Streudiagramme frei geschätzter Itemparameter für den Bereich Leseverständnis getrennt nach Itemcluster *(Anmerkungen*: Anordnung der Itemcluster nach Zeilen: R1, R2 und R3)

mittelten Effekte sind mit jeweils einer Line markiert (gestrichelt für den ersten Messzeitpunkt und durchgezogen für den zweiten Messzeitpunkt). Der Trend der eingezeichneten Linien liefert ein Indiz dafür, inwieweit die ermittelten Testkontexteffekte im Sinne von Positionseffekten interpretiert werden können.

Insgesamt zeigte das in Abb. 4 dargestellte Befundmuster deutliche Hinweise für die Existenz von Positionseffekten, da die ermittelten Bookleteffekte in eine deutliche Abnahme der Testleistungen über Positionen hinweg mündeten. Die Stärke der Positionseffekte unterschied sich dabei augenscheinlich zwischen Inhaltsdomänen, Messzeitpunkten und Schulformen. Die Positionseffekte waren im Bereich Mathematik am schwächsten und im Leseverständnistest am stärksten ausgeprägt. Schülerinnen und Schüler aus nichtgymnasialen Schulformen waren in allen Domänen stärker von Positionseffekten betroffen. Schließlich nahm die Stärke der Positionseffekte in den Bereichen Naturwissenschaften und Leseverständnis zum zweiten Messzeitpunkt deutlich zu, wobei der Anstieg in den nichtgymnasialen Schulformen besonders stark ausfiel.

Insgesamt legte das vorliegende Befundmuster nahe, dass die Ausblendung der Testkontexteffekte zu einer Verzerrung der querschnittlichen Schulformunterschiede

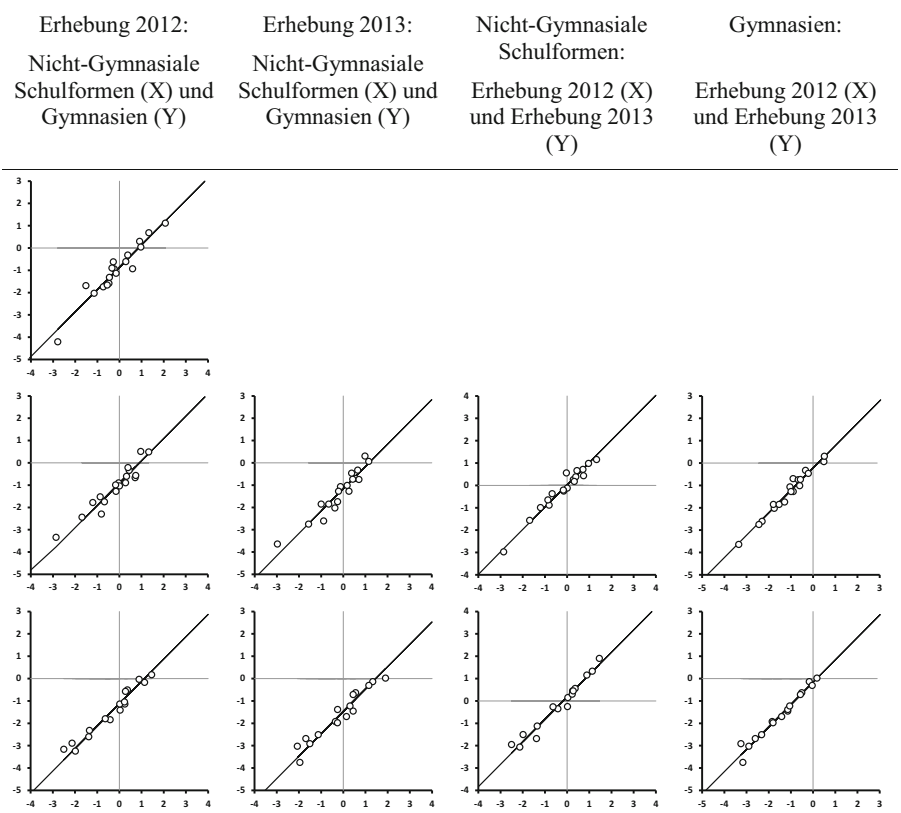

Abb. 3 Bivariate Streudiagramme frei geschätzter Itemparameter für den Bereich Naturwissenschaften getrennt nach Itemcluster (*Anmerkungen*: Anordnung der Itemcluster nach Zeilen: N1, N2 und N3)

(Überschätzung des Leistungsvorsprungs der Gymnasiasten) und der mittleren Änderungsraten (Unterschätzung bei zunehmenden Positionseffekten) führen könnte.

Auswirkungen der Invarianzannahmen und des selektiven Stichprobenausfalls auf die Zuwachsschätzungen. Im Folgenden wird auf die Frage zur Sensitivität der Zuwachsschätzungen gegenüber dem angenommenen Grad an Messinvarianz eingegangen. Konkret wurden die Zuwachsschätzungen, die sich in zwei unterschiedlichen Modellen ergaben, miteinander verglichen. Das erste Modell basierte auf der Annahme vollständig invarianter Itemparameter (zwischen Gruppen und Messzeitpunkte). Im zweiten Modell wurden zeit- und gruppenspezifische Testkontexteffekte zugelassen (Gln. 2 und 3). Die Annahme war, dass die Ausblendung der identifizierten Positionseffekte zu einer Unterschätzung der Zuwachsraten führen würde, wobei dieser Effekt besonders stark in der Gruppe der nichtgymnasialen Schulformen in den Bereichen Leseverständnis und Naturwissenschaften auftreten sollte.

Ein weiterer Effekt, der in diesem Abschnitt dargestellt werden soll, bezieht sich auf die Auswirkungen der Stichprobenselektivität auf die Zuwachsschätzungen. Im

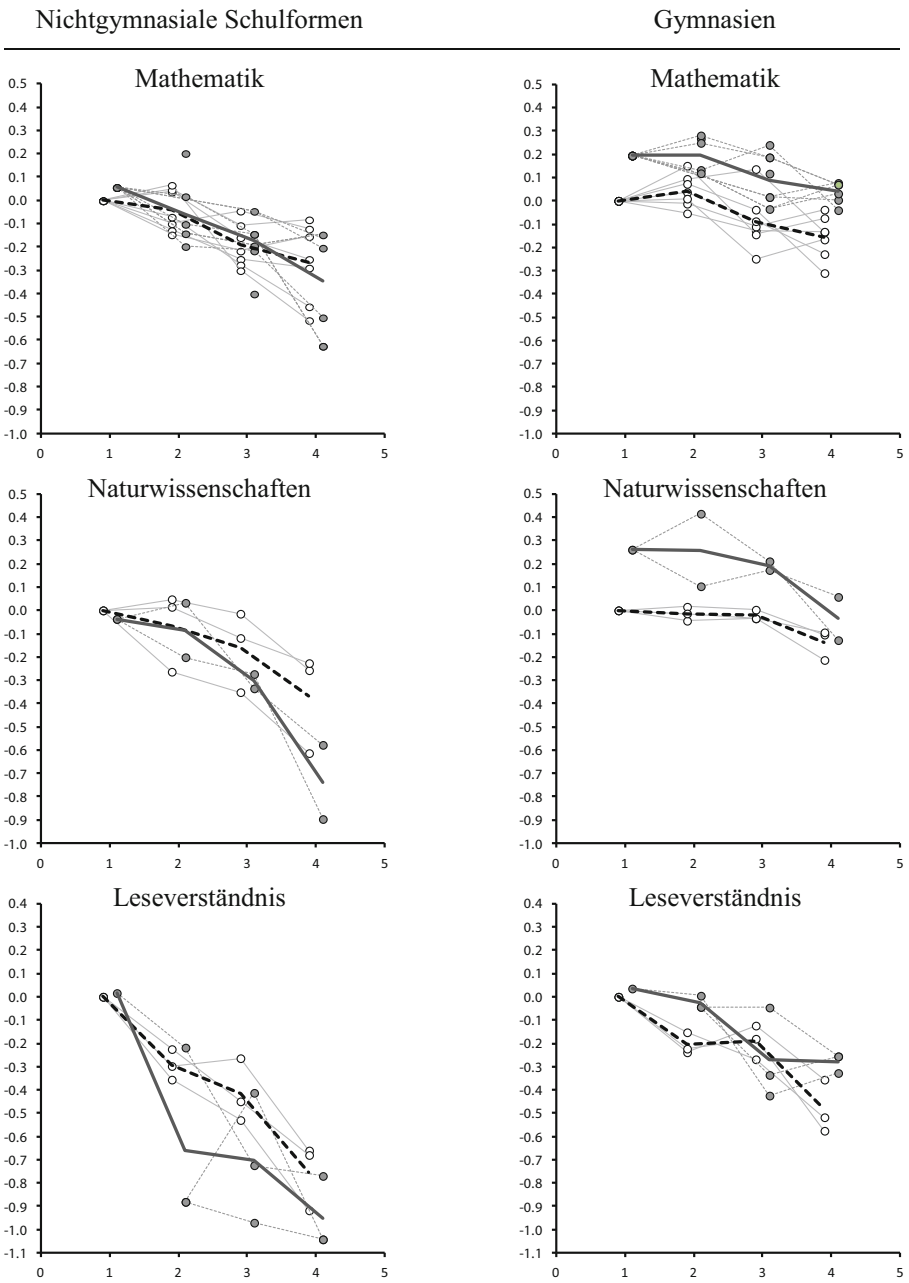

Abb. 4 Testkontexteffekte nach Itemclusterpositionen in den Bereichen Mathematik, Leseverständnis und Naturwissenschaften. Mittlere Logits der Itemcluster nach Itemclusterpositionen (X-Achse). Logits an der ersten Itemclusterposition in der Erhebung 2012 auf 0 normiert. (*Anmerkungen*: *Weiße Kreise* Mittlere Logits nach Itemcluster in der Erhebung 2012, *Schwarze Kreise* Mittlere Logits nach Itemcluster in der Erhebung 2013, *Linien* Mittelwerte der Logits über Itemcluster nach Itemclusterpositionen in Erhebung 2012 (*gebrochene Linien*) und in der Erhebung 2013 (*durchgezogene Linien*))

Tab. 3 Mittlere Zuwachsschätzungen auf Grundlage unterschiedlicher Messmodelle

	Gesamtgruppe		Nicht-gymnasiale Schulformen		Gymnasien	
	Est	(*SE*)	*Est*	(*SE*)	*Est*	(*SE*)
Ohne Berücksichtigung der Ausgangsleistungen						
Mathematik						
Volle Invarianz[A]	0,087	(0,024)**	0,006	(0,036)	0,186	(0,026)**
Testkontexteffekte[A]	0,119	(0,033)**	0,057	(0,045)	0,194	(0,040)**
Leseverständnis						
Volle Invarianz	–0,101	(0,036)**	–0,231	(0,056)**	0,054	(0,038)
Testkontexteffekte	0,025	(0,066)	0,016	(0,087)	0,035	(0,096)
Naturwissenschaften						
Volle Invarianz	0,037	(0,028)	–0,119	(0,041)**	0,225	(0,036)**
Testkontexteffekte	0,098	(0,040)*	–0,037	(0,061)	0,261	(0,048)**
Mit Berücksichtigung der Ausgangsleistungen						
Mathematik						
Volle Invarianz[A]	0,072	(0,018)**	–0,011	(0,023)	0,172	(0,023)**
Testkontexteffekte[A]	0,108	(0,035)**	0,047	(0,041)	0,180	(0,048)**
Leseverständnis						
Volle Invarianz	–0,119	(0,030)**	–0,250	(0,042)**	0,040	(0,040)
Testkontexteffekte	0,017	(0,072)	–0,023	(0,094)	0,066	(0,101)
Naturwissenschaften						
Volle Invarianz	–0,005	(0,023)	–0,181	(0,031)**	0,204	(0,033)**
Testkontexteffekte	0,031	(0,045)	–0,111	(0,060)	0,199	(0,064)**

* $p < 0{,}05$; ** $p < 0{,}01$

vorliegenden Fall sind die zum zweiten Messzeitpunkt teilnehmenden Schülerinnen und Schüler hinsichtlich der Ausgangsleistungen selegiert (vgl. Heine et al. 2017). Des Weiteren berücksichtigen die bis zu dieser Stelle vorgestellten Modelle lediglich die Ausgangsleistungen in einer Domäne. Hinzu kommt, dass die für die Bereiche Naturwissenschaften und Leseverständnis angepassten Modelle nur rund 78 % der Schülerinnen und Schüler in der Gesamtgruppe (N = 6384) berücksichtigen. Vor diesem Hintergrund stellt sich die Frage, inwieweit sich die Berücksichtigung der Ausgangsleistungen in allen Leistungsdomänen auf die Schätzung der Zuwachsraten in der Population auswirkt (z. B. Peugh und Enders 2004). Eine Situation, in der sich die Zuwachsschätzungen nach Berücksichtigung aller Kompetenzbereiche deutlich von den Ergebnissen bei alleiniger Berücksichtigung der fokalen Domäne unterscheiden, indiziert, dass die resultierende Zuwachsschätzung maßgeblich von der Stichprobenselektivität hinsichtlich der anderen Leistungsbereiche abhängt. Um diese Fragestellung zu untersuchen, wurden die jeweiligen Messmodelle im vollständigen Sample unter Berücksichtigung der im Rahmen des PISA-2012-Assessments ermittelten Kompetenzwerte (OECD 2014) erneut geschätzt.[5] Zu diesem Zweck wur-

[5] Die Berücksichtigung der zusätzlichen Variablen wirkte sich nicht auf die Schätzung der Itemparameter und der Testkontexteffekte aus. Aus diesem Grund werden die entsprechenden Befunde hier nicht vorgestellt.

den die domänenspezifischen längsschnittlichen IRT-Modelle (Gln. 3 und 4) um die zum ersten Messzeitpunkt vorliegenden Kompetenzwerte in den nicht korrespondierenden Domänen erweitert. Beispielsweise wurde das Längsschnittmodell für den Bereich Mathematik um die skalierten Kompetenzwerte in den Bereichen Naturwissenschaften und Leseverständnis erweitert, wobei die Korrelationen zwischen den Mathematikkompetenzen und den Ausgangswerten in den Bereichen Lesen und Naturwissenschaften frei geschätzt wurden. In diesem Fall wurde der skalierte Leistungswert im Bereich Mathematik nicht aufgenommen, da die entsprechenden Ausgangsleistungen bereits in der modellierten latente Variable enthalten waren.

Die Ergebnisse dieser Auswertung sind in Tab. 3 zusammengefasst. Die Ergebnisse lassen sich folgt zusammenfassen: Die Berücksichtigung der Testkontexteffekte resultierte in allen Kompetenzbereichen in höheren mittleren Zuwachsschätzungen in der Gesamtgruppe, wobei die Unterschiede in den erzielten Zuwächsen im Bereich Mathematik eher trivial ausfielen. Die Unterschiede traten im Bereich Leseverständnis besonders deutlich hervor. Die Ausblendung der Testkontexteffekte mündete in einer statistisch signifikanten Kompetenzabnahme, die bei Berücksichtigung dieser Effekte jedoch verschwand. Die Aufschlüsselung der Analysen nach Schulformen bestätigte das erwartete Muster, wonach die Unterschiede zwischen den Skalierungen besonders deutlich in der Gruppe der nichtgymnasialen Schulformen in den Bereichen Leseverständnis und Naturwissenschaften zutage traten. Der Grund hierfür ist die stärker ausgeprägte Zunahme des Effekts der Clusterposition in dieser Gruppe. Die Berücksichtigung dieses Effekts durch die Modellierung der Testkontexteffekte wirkte sich nicht nur auf die Zuwachsschätzungen aus, sondern reduzierte auch die querschnittlichen Schulformunterschiede des ersten Messzeitpunkts (Mittelwertdifferenzen Gymnasien – nichtgymnasiale Schulformen: Mathematik: 1,27 vs. 1,23; Naturwissenschaften: 0,96 vs. 0,88; Leseverständnis: 1,10 vs. 0,99; alle Angaben auf der Logit-Metrik).

Ein zweiter Befund ist, dass die Berücksichtigung der Ausgangsniveaus der Kompetenzen in allen Domänen die Zuwachsschätzungen beeinflusste, wobei die Unterschiede in den Bereichen Mathematik und Leseverständnis praktisch gesehen vernachlässigbar waren (Modelle ohne und mit Berücksichtigung aller Ausgangsleitungen in Tab. 3). In diesen Bereichen ließ sich der Effekt des längsschnittlichen Stichprobenausfalls auf die Zuwachsschätzung hinreichend gut durch die Ausgangswerte der jeweiligen Domänen abbilden. Das war im Bereich Naturwissenschaften nicht der Fall, da hier nach Berücksichtigung der anfänglichen Kompetenzniveaus in den Bereichen Mathematik und Leserverständnis eine vergleichsweise starke Reduktion der Leistungsentwicklung festgestellt wurde, die –0,07 Logits in der Gruppe nichtgymnasialer und –0,06 Logits in der Gruppe der Gymnasien betrug (Modelle mit Testkontexteffekten). An dieser Stelle bleibt festzuhalten, dass der in der Gruppe der nichtgymnasialen Schulformen geschätzte Kompetenzrückgang den in Abhängigkeit der Bereiche Mathematik und Leseverständnis stattfinden Stichprobenausfall widerspiegelt.

Tab. 4 Deskriptive Statistiken in der PVs in der Gesamtgruppe und nach Schulformen (Nichtgymnasiale Schulformen und Gymnasien)

	Erhebung 2012			Erhebung 2013			
	M	*SD*	*Rel*	*M*	*SD*	*Rel*	*MD*
Mathematik							
Gesamt	0,55	1,03	0,90	0,64	1,18	0,86	0,09**
Nicht-Gym	–0,00	0,82	0,85	0,02	0,96	0,78	0,02
Gymnasien	1,23	0,83	0,84	1,39	0,98	0,82	0,17**
Leseverständnis							
Gesamt	0,42	0,94	0,85	0,40	1,22	0,81	–0,03
Nicht-Gym	–0,01	0,84	0,81	–0,09	1,18	0,75	–0,07**
Gymnasien	0,95	0,76	0,77	0,99	0,98	0,80	0,03
Naturwissenschaften							
Gesamt	0,40	0,87	0,87	0,40	1,16	0,86	0,00
Nicht-Gym	–0,02	0,79	0,84	–0,18	1,02	0,81	–0,17**
Gymnasien	0,90	0,67	0,78	1,11	0,89	0,78	0,21**

Rel. PV-Reliablität, *MD* Differenz der Mittelwerte auf der Logit-Metrik
* $p < 0{,}05$; ** $p < 0{,}01$

4.2 Ergebnisse der Kompetenzschätzung

Die im vorangegangenen Abschnitt berichteten Ergebnisse zur Messinvarianz der verwendeten Messinstrumente wurden bei der Abschätzung der schülerseitigen Kompetenzniveaus und deren Veränderung berücksichtigt. Dazu wurden die Kompetenzausprägungen unter Berücksichtigung von Testkontexteffekten geschätzt, indem als Referenz für die Kompetenzschätzungen für jeden Messzeitpunkt jeweils die erste Itemclusterposition gewählt wurde. Dieses Verfahren impliziert somit eine Korrektur der geschätzten Kompetenzniveaus und deren Veränderung hinsichtlich der mittleren gruppen- und testzeitpunktspezifischen Textkontexteffekte. Die Abschätzung der Kompetenzniveaus wurde in mehreren Schritten vorgenommen.

Im ersten Schritt wurden anhand der zuvor beschriebenen Schätzungen (Schätzung mit allen Ausgangsleistungen) die booklet-, gruppen- und messzeitpunktspezifischen Item- und Step-Parameter abgeleitet (Gln. 2 und 3). Diese Prozedur mündete für die an der ersten Itemclusterposition präsentierten Items in gruppen- und zeitinvarianten Item- und Step-Parametern, lieferte aber für Items, die an späterer Stelle dargeboten wurden, unterschiedliche Itemparameter, die bis auf wenige Ausnahmen höhere Itemschwierigkeiten an späteren Positionen indizierten (Abb. 4).

Um die so gewonnen Itemparameter angemessen zu berücksichtigen, wurden die vorliegenden Items in einem zweiten Schritt in sogenannte „virtuelle Items“ transformiert. Hierzu wurden alle Items eines Clusters, das nicht zur ersten Itemclusterposition dargeboten wurde, als separate (d. h. virtuelle) Items behandelt. Dies führte zum Beispiel dazu, dass alle Items der ersten Erhebung exakt viermal im Datensatz auftauchten, da sie an vier Positionen dargeboten wurden. Die Itemparameter aller Items wurden um die zuvor bestimmten Schätzungen der Testkontexteffekte adjustiert. Zu diesem Zweck wurde auf die Schwierigkeit eines Items j aus Cluster c, das zum Messzeitpunkt t in Gruppe g in Booklet b dargeboten wurde, der Effekt

des Testkontexts γ_{cbtg} addiert (vgl. Gln. 3 und 4). Vor der Schätzung der Kompetenzniveaus wurden alle nicht erreichten Items in Übereinstimmung mit der in PISA üblichen Prozedur als falsch kodiert. Die finale Abschätzung der Kompetenzniveaus fand schließlich im dritten Schritt statt, wobei die Generierung der PVs getrennt für beide Gruppen (Gymnasien und nicht-gymnasiale Schulformen) vollzogen wurde.

Die PV-Schätzung fand im Rahmen eines sechsdimensionalen Raschmodells (3 Inhaltsdomänen × 2 Messzeitpunkte) statt, wobei das in PISA verwendete Hintergrundmodell zum Einsatz kam (OECD 2014). Im Gegensatz zur ursprünglichen PISA-Studie stellte sich jedoch die Spezifikation von Dummy-Indikatoren für die teilnehmenden Schulen im Hintergrundmodell aufgrund des komplexen Testdesigns als problematisch heraus, sodass von der Verwendung dieser Indikatoren abgesehen wurde.[6] Allerdings berücksichtigte das spezifizierte Hintergrundmodell die gruppenspezifische längsschnittliche Kovarianzstruktur der Kompetenzdimensionen, sowie die gruppen- und zeitspezifischen Zusammenhänge der Kompetenzdimensionen mit den betrachteten Hintergrundvariablen. Die Generierung der PVs fand mithilfe des Programms ConQuest (Wu et al. 2007) statt, wobei eine Monte-Carlo-Integration mit 25.000 Integrationspunkten verwendet wurde.

Ergebnisse der PV-Generierung. In Tab. 4 werden die deskriptiven Statistiken der PVs in der Gesamtstichprobe sowie in der nach Schulformen aufgeteilten Stichprobe berichtet. Die Ergebnisse zeigen, dass die für die Gesamtgruppe ermittelten mittleren Zuwachsschätzungen auf der Logit-Metrik vergleichbar zu den initialen Schätzungen auf Grundlage des Messmodells mit Testkontexteffekten (Tab. 3) ausfielen. Die mittleren Zuwachsschätzungen für die Gymnasien fielen ebenso insgesamt vergleichbar zu den in Tab. 3 berichteten Ergebnissen aus (Modelle mit Testkontexteffekten unter Berücksichtigung aller Ausgangsleistungen). Für die Gruppe der nicht-gymnasialen Schulformen erbrachten die Zuwachsschätzungen im Rahmen der PV-Skalierung für

Tab. 5 Interkorrelationen der PVs nach Schulformen. Untere Dreiecksmatrix: Nichtgymnasiale Schulformen; obere Dreiecksmatrix: Gymnasien

	Erhebung 2012			*Erhebung 2013*		
	Mathe	Leseverst	Naturwiss	Mathe	Leseverst	Naturwiss
Erhebung 2012						
Mathematik	*1,00*	0,68	0,80	**0,88**	0,47	0,66
Leseverständ	0,71	*1,00*	0,71	0,67	**0,60**	0,65
Naturwiss	0,83	0,81	*1,00*	0,74	0,52	**0,72**
Erhebung 2013						
Mathematik	**0,83**	0,68	0,73	*1,00*	0,62	0,74
Leseverständ	0,53	**0,70**	0,60	0,72	*1,00*	0,66
Naturwiss	0,70	0,75	**0,76**	0,81	0,78	*1,00*

Alle $p < 0{,}01$

[6] Konkret konnte keine Konvergenz der zur PV-Generierung eingesetzten IRT-Modelle erzielt werden. Der Grund hierfür liegt aller Wahrscheinlichkeit nach in der geringen Schülerzahl in vielen Einzelschulen. Die längsschnittliche Reduktion der Schülerstichprobe und das zum zweiten Messzeitpunkt eingesetzte Bookletdesign hatten zur Folge, dass in vielen Schulen nur wenige oder gar keine Schüler Iteminformationen zu den zur Retestwelle erhobenen Kompetenzbereichen lieferten.

die Bereiche Leseverständnis und Naturwissenschaften Ergebnisse, die sich von den Ausgangsschätzungen (Tab. 3) unterschieden. Die Veränderungen in den Bereichen Lesen und Naturwissenschaften verringerten sich um ca. 0,05 Logits. Dies führte zu statistisch signifikanten, aber praktisch vernachlässigbaren Kompetenzabnahmen im Bereich Leseverständnis. Der nun im Bereich Naturwissenschaften in der Gruppe nichtgymnasialer Schulformen ermittelte Kompetenzrückgand weist jedoch einen praktisch bedeutsamen Betrag auf.

Es stellt sich für den Bereich Naturwissenschaften die Frage, inwieweit die Kompetenzrückgänge in den nichtgymnasialen Schulformen die reale Entwicklung über ein Schuljahr wiederspiegeln. Im vorliegenden Fall resultiert die Kompetenzabnahme als eine Funktion des Stichprobenausfalls und des Hintergrundmodells (d. h. aufgrund der Ausgangswerte und der Hintergrundvariablen erwartete Leistungsentwicklung für Schülerinnen und Schüler, die dem zweiten Messzeitpunkt ferngeblieben sind) und sollte dementsprechend vorsichtig interpretiert werden.

Neben den Zuwächsen in den Mittelwerten ist die Veränderung der Streuung ein weiterer zentraler Aspekt der Kompetenzentwicklung. Die in Tab. 4 dargestellten Ergebnisse dokumentieren einen Anstieg der Streuung in allen Kompetenzdomänen. Die Zunahmen waren im Bereich Mathematik vergleichsweise gering ausgeprägt. Hier bewegte sich das Verhältnis der Standardabweichungen in beiden Gruppe auf einem vergleichbaren Niveau (nichtgymnasiale Schulformen: 1,16; Gymnasien: 1,18). Streuungszuwächse, die auf einem höheren Niveau angesiedelt waren, fanden sich im Bereich der Naturwissenschaften (nichtgymnasiale Schulformen: 1,29; Gymnasien: 1,33). Der stärkste Anstieg in der Variabilität von 1,40 wurde im Bereich Leseverständnis in der Gruppe nichtgymnasialer Schulformen festgestellt (Gymnasien: 1,28). Interessanterweise war in den Kompetenzdomänen, in denen die Zunahme der Leistungsheterogenität stärker ausgeprägt war (Leseverständnis und Naturwissenschaften), auch eine stärkere Zunahme der Testkontexteffekte zu verzeichnen (Abb. 4). Dies lässt vermuten, dass die in den Bereichen Leseverständnis und Naturwissenschaften vorliegenden Varianzzunahmen in Teilen auf einen Anstieg individueller Unterschiede in der Ausprägung der Testkontextkontexteffekte zurückzuführen sein könnte.

Die aufgrund der PVs ermittelten Interkorrelationen der Kompetenzdomänen sind in Tab. 5 für die in der nach Schulformen aufgeteilten Stichprobe berichtet. Die in dieser Studie ermittelten Zusammenhänge dokumentieren erneut den starken Zusammenhang der Kompetenzbereiche. Die Zusammenhänge fielen aber aufgrund der nach Schulform aufgeteilten Stichprobe etwas geringer aus als die in anderen PISA-Berichtsbänden dargestellten Zusammenhänge (z. B. OECD 2014). In beiden Gruppen ergaben sich zu beiden Messzeitpunkten qualitativ vergleichbare Korrelationsmuster, wobei die Kompetenzen in den Bereichen Mathematik und Naturwissenschaften sowie Leseverständnis und Naturwissenschaften besonders hoch zusammenhingen. Ebenso fiel das Muster der Reteststabilitäten (d. h. Korrelationen einer Kompetenzdimension über die Zeit) vergleichbar aus. Mathematik erzielte die höchste und Leseverständnis die geringste Stabilität. Die zum zweiten Messzeitpunkt erfassten domänenspezifischen PVs korrelierten mit den jeweils entsprechenden Ausgangsleistungen am höchsten. Dieses Muster stellte sich aber nicht für die zeitversetzten Korrelationen der anfänglichen Kompetenzstände ein. Die PVs im

Leseverständnis korrelierten höher (bzw. gleich hoch) mit den zum zweiten Messzeitpunkt erhobenen Kompetenzen in den Bereichen Mathematik und Naturwissenschaften. Ebenso bewegten sich die zeitversetzten Korrelationen der PVs des Bereichs Naturwissenschaften mit Mathematik ungefähr auf dem Niveau der entsprechenden Reteststabilität.

Das in Tab. 5 berichtete Korrelationsmuster unterstreicht die anhand der Zuwächse der Varianzen und der mittleren Testkontexteffekte indizierte qualitative Veränderung der Retestmessungen in den Bereichen Leseverständnis und Naturwissenschaften. Die vergleichsweise geringe Stabilität sowie das undifferenzierte Muster der zeitversetzten Korrelationen der Ausgangsmessungen sprechen für die Vermutung, dass die zum zweiten Messzeitpunkt eingesetzten Maße auch Varianzkomponenten, wie zum Beispiel individuelle Unterschiede in Positionseffekten, umfassen, die nicht auf die zugrundeliegenden Kompetenzen zurückgeführt werden können.

5 Zusammenfassung und Diskussion

Der vorliegende Beitrag verfolgte das Ziel eine zeit- und schulforminvariante Metrik der in der PISA-Längsschnittstudie der Jahre 2012–2013 erhobenen schülerseitigen Kompetenzausprägungen zu etablieren. Besonderes Augenmerk wurde auf mögliche Verstöße der Annahme vollständiger Messinvarianz (Reise et al. 1993) gelegt, die sich in Abhängigkeit der Iteminhalte (z. B. Klieme und Baumert 2001) und/oder differentiell wirkenden Testkontexteffekte (Brennan 1992) ergeben könnten. Die vorliegenden Ergebnisse lieferten Hinweise für Abweichungen vom Ideal eines vollständig messinvarianten Messmodells, die in erster Linie auf die Einflüsse von Testkontexteffekten zurückgeführt werden konnten. Die hier identifizierten Testkontexteffekte spiegelten hauptsächlich Positionseffekte wider, deren Stärke sich zwischen Schulformen und Messzeitpunkten unterschied. Schülerinnen und Schüler an nichtgymnasialen Schulformen waren in allen Leistungsdomänen stärker von Testkontexteffekten betroffen und die Höhe der entsprechenden Effekte stieg insbesondere in den Bereichen Leseverständnis und Naturwissenschaften zur zweiten Erhebungswelle hin an.

Das zur Generierung der finalen Schätzungen der Kompetenzniveaus verwendete IRT-Modell berücksichtigte die in Abhängigkeit der Schulform und des Messzeitpunkts vorliegenden Testkontexteffekte. Die generierten PVs basieren dabei auf der Annahme, dass die zur Verankerung der Metrik verwendeten Itemcluster (d. h. Itemcluster, die an der ersten Position der Testhefte eingesetzt wurden) relativ frei von Testkontexteffekten sind. Bei der Verwendung der hier generierten PVs müssen jedoch die Grenzen der vorgenommenen Korrektur bedacht werden. Insbesondere sind hier zwei Punkte hervorzuheben, die sich erstens auf die zur Verankerung der Leistungsmetrik gewählten Itemcluster und zweitens auf die berücksichtigten Determinanten der Textkontexteffekte (Schulform) beziehen.

Die Annahme, dass die zur ersten Clusterposition präsentierten Items nicht von Textkontexteffekten betroffen sind, dürfte nur näherungsweise zutreffen. Erstens können sich Positionseffekte bereits während der Bearbeitung des ersten Itemclusters einstellen (Meyers et al. 2009). Zweitens ist davon auszugehen, dass die zum

zweiten Messzeitpunkt erfassten Kompetenzstände von einem Effekt der wiederholten Testdurchführung betroffen sind (z. B. DeMars 2007). Derartige Effekte sind ein Kennzeichen nahezu aller Längsschnittstudien, wobei deren Identifikation aufgrund fehlender Vergleichsgruppen nahezu unmöglich ist. Testwiederholungseffekte können zu artifiziellen Überschätzungen der Kompetenzentwicklung führen, wenn das identische Testmaterial wiederholt getestet eingesetzt wird (Kulik et al. 1984). Im PISA-Längsschnitt traten derartige Wiederholungen aufgrund des eingesetzten Multi-Matrix-Designs nur selten auf, sodass dieser Faktor keine besondere Rolle spielen dürfte.[7] Testwiederholungseffekte können jedoch auch zu einer Reduktion der Testleistungen führen, wenn die wiederholte Bearbeitung eines Tests zu einer Reduktion der Testmotivation führt (DeMars 2007). Unsere Ergebnisse lieferten Hinweise dafür, dass die Testmotivation über den Verlauf des Tests in der Messwiederholung stärker abnahm. Sie schließen jedoch nicht die Möglichkeit aus, dass die Testmotivation der Schülerinnen und Schüler bereits zu Beginn der zweiten Testung geringer ausgeprägt war.

Die Reskalierung der Leistungsdaten berücksichtigte neben messzeitpunktspezifischen Ausprägungen von Testkontexteffekten auch Schulformunterschiede in diesen Störfaktoren. Diese Erweiterung wurde vorgenommen, um die Validität von Schulformvergleichen zu optimieren. Insgesamt gilt aber zu berücksichtigen, dass Testkontexteffekte nicht nur zwischen Schulformen, sondern auch zwischen Schulen und Individuen variieren können (Debeer et al. 2014). Für die erzeugten PVs bedeutet dies, dass ein Teil der Variation in den Kompetenzschätzungen auch auf Unterschiede in der Höhe der Testkontexteffekte zurückzuführen sein könnte. Wir vermuten, dass insbesondere die in den nichtgymnasialen Schulformen zum zweiten Messzeitpunkt erhobenen Testleistungen in den Bereichen Leseverständnis und Naturwissenschaften von diesem Effekt betroffen sind. Die mittleren Testkontexteffekte und die längsschnittlichen Varianzzunahmen fielen hier besonders hoch aus. Die Modellierung individueller Unterschiede in Positionseffekten (Debeer und Janssen 2013), die auch artifiziellen Varianzzuwächsen entgegenwirken kann, erscheint als eine interessante Alternative zum hier verwendeten Ansatz. Dieser Ansatz erwies sich in der vorliegenden Studie jedoch als nicht praktikabel.[8]

[7] Wir sind dieser Fragestellung nachgegangen, indem wir die Lösungswahrscheinlichkeiten der zum zweiten Messzeitpunkt eingesetzten Items zwischen Schülerinnen und Schülern verglichen haben, die die entsprechenden Items bereits zum ersten Messzeitpunkt bearbeitet haben und solchen, die die Items zum ersten Mal vorgelegt bekamen. Es konnte kein Vorteil einer wiederholten Bearbeitung spezifischer Items nachgewiesen werden.

[8] Wir haben uns aus verschiedenen Gründen gegen diesen Ansatz entschieden, da er erstens auf nur eine Form von Textkontexteffekten (Positionseffekten) beschränkt ist, zweitens Kenntnisse über die funktionale Form von Positionseffekten voraussetzt, drittens eine ausreichende Zahl von Permutationen von Itemclustern über Positionen benötigt und viertens zu einer Verdoppelung der Zahl der Dimensionen in den finalen Skalierungsmodellen führt. Letzteres hätte zur Folge, dass die PV-Generierung mittels eines 12-dimensionalen Messmodells erfolgen müsste, das kaum mit den gängigen Softwarepaketen und Computern im Rahmen einer Maximum-Likelihood-Schätzung handzuhaben ist.

5.1 Implikationen für die Nutzung der PVs im PISA-Längsschnitt

Die Ergebnisse der hier berichteten Längsschnittskalierung legen insgesamt nahe, dass die Robustheit der Analyseergebnisse zwischen Kompetenzdomänen und Schulformen variieren dürfte. Hinsichtlich der Kompetenzdomänen erwies sich die Messung im Bereich Mathematik als besonders verlässlich (hohe Reliabilität und Stabilität sowie schwach ausgeprägte Positionseffekte und Varianzzunahmen). Die berichteten Detailanalysen sprechen insgesamt dafür, dass die für diese Domäne erreichte Messpräzision detaillierte Auswertungen in allen Schulformen ermöglichen (z. B. mittlere Zuwächse und Variabilität von Zuwächsen).

Für die Bereiche Naturwissenschaften und Leseverständnis zeichnete sich ein Bild ab, wonach sich die Veränderungsmessung in der Gruppe der Gymnasiasten als relativ robust herausstellte. Die mittleren Zuwachsschätzungen, die mit einem leichten Anstieg der Leistungsvariabilität einhergehen, entsprachen insgesamt den Erwartungen. Demgegenüber erscheinen in der Gruppe der nichtgymnasialen Schulformen einige Aspekte der Zuwachsschätzungen als problembehaftet. Für den Bereich Leseverständnis lässt der besonders deutliche Anstieg der Positionseffekte vermuten, dass der vergleichsweise starke Anstieg der Leistungsvariabilität unter Umständen auf die Zunahme individueller Unterschiede in Positionseffekten zurückzuführen sein könnte. Eine derartige Situation würde implizieren, dass die Variabilität der im Jahr 2013 erfassten Kompetenzstände mit individuellen Unterschieden in Positionseffekten konfundiert ist.

Im Bereich Naturwissenschaften erscheint die in der Gruppe der nichtgymnasialen Schulformen festgestellte mittlere Leistungsabnahme als besonders problematisch, da sich diese aus der Extrapolation der Zusammenhänge der Hintergrunddaten ableitete. Insbesondere führten die starken Zusammenhänge der zum ersten Messzeitpunkt erhobenen Kompetenzen in den Bereichen Lesen und Mathematik mit den zum zweiten Messzeitpunkt gemessenen naturwissenschaftlichen Kompetenzen (vgl. Tab. 5) dazu, dass denjenigen Schülerinnen und Schülern, die der zweiten Erhebung ferngeblieben sind und zum ersten Messzeitpunkt niedrige Kompetenzniveaus in den Bereichen Mathematik und Lesen aufwiesen, zum zweiten Messzeitpunkt Kompetenzniveaus zugeschrieben wurden, die im Mittel unterhalb der zum ersten Messzeitpunkt geschätzten Niveaus lagen. Inwieweit diese Extrapolation der Daten im Sinne eines statistischen Artefakts zu interpretieren ist, oder aber eine erwartungstreue Schätzung der tatsächlichen Kompetenzentwicklung an nichtgymnasialen Schulformen darstellt, ist eine offene Frage, die mit den vorliegenden Daten nicht abschließend geklärt werde kann. [9]

[9] Insgesamt stellt sich auch die Frage, inwieweit die Abschätzung der Kompetenzzuwächse ohne Berücksichtigung von Kovariaten eine validere Darstellung der Leistungsentwicklung liefert. Da diese Frage nicht eindeutig beantwortet werden kann, haben wir uns für die Verwendung aller vorliegenden Daten entschieden, da deren Ausblendung ebenso in Frage gestellt werden kann. Es steht außer Frage, dass die Hinzunahme weiterer Variablen die Reliabilität von Leistungsunterschieden erhöht (Mislevy et al. 1992).

6 Resümee

Mit der Modellierung der Veränderungen der Messeigenschaften der wiederholt eingesetzten Leistungstests haben wir in diesem Beitrag den Fokus der Analysen um die Rolle von Testkontexteffekten und deren Veränderung erweitert. Unsere Befunde deuten klar drauf hin, dass die Abweichungen vom Optimalkriterium vollständig invarianter Messmodelle auf die Zunahme (unerwünschter) Testkontexteffekte (Brennan 1992) – und hier insbesondere Positionseffekte (Leary und Dorans 1985) – zurückzuführen sind. Die Befundlage indiziert insgesamt, dass sich die wiederholt getesteten Schülerinnen und Schüler durch eine Abnahme der Persistenz der Testbearbeitung (Debeer et al. 2014) auszeichnen, wobei die entsprechenden Abnahmen in der Gruppe der Schülerinnen und Schüler an nichtgymnasialen Schulformen besonders hoch ausgeprägt ist. Die im Rahmen des aktuellen Beitrags vorgeschlagenen Skalierungsmodelle ermöglichen eine (partielle) Korrektur dieser unerwünschten Effekte und haben das Potential, den im Rahmen längsschnittlicher Trendschätzungen auftretenden Verzerrungen entgegenzuwirken.

Literatur

Bejar, I. I. (1980). A procedure for investigating the unidimensionality of achievement tests based on item parameter estimates. *Journal of Educational Measurement*, *17*, 283–296.

Brennan, R. L. (1992). The context of context effects. *Applied Measurement in Education*, *5*, 225–264.

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*, 456–466.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differential item functioning test items. *Educational Measurement: Issues and Practice*, *17*, 31–44.

Debeer, D., & Janssen, R. (2013). Modeling item position effects within an IRT framework. *Journal of Educational Measurement*, *50*, 164–185.

Debeer, D., Buchholz, J., Hartig, J., & Janssen, R. (2014). Student, school, and country differences in sustained test-taking effort in the 2009 PISA reading assessment. *Journal of Educational and Behavioral Statistics*, *39*, 502–523.

DeMars, C. E. (2007). Changes in rapid-guessing behavior over a series of assessments. *Educational Assessment*, *12*, 23–45.

Frey, A., Hartig, J., & Rupp, A. (2009). Booklet designs in large-scale assessments of student achievement: theory and practice. *Educational Measurement: Issues and Practice*, *28*, 39–53.

Goldstein, H. (2009). Handling attrition and non-response in longitudinal data. *Longitudinal and Life Course Studies*, *1*, 63–72.

Graham, J. W. (2003). Adding missing-data-relevant variables to FIML-based structural equation models. *Structural Equation Modeling*, *10*, 80–100.

Graham, J. W., Cumsille, P. E., & Elek-Fisk, E. (2003). Methods for handling missing data. In J. A. Schinka & W. F. Velicer (Hrsg.), *Handbook of Psychology* (Bd. 2, S. 87–114). New York: Wiley.

Harris, D. (1991). Effects of passage and item scrambling on equating relationships. *Applied Psychological Measurement*, *15*, 247–256.

Hartig, J., & Buchholz, J. (2012). A multilevel item response model for item position effects and individual persistence. *Psychological Test and Assessment Modeling*, *54*, 418–431.

Hawkes, D., & Plewis, I. (2006). Modelling non-response in the national child development study. *Journal of the Royal Statistical Society A*, *169*, 479–492.

Heine, J. H., Nagy, G., Meinck, S., Zühlke, O., Mang, J. (2017). Empirische Grundlage, Stichprobenausfall und Adjustierung im PISA-Längsschnitt 2012–2013. *Zeitschrift für Erziehungswissenschaft.* doi:10.1007/s11618-017-0756-0.

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. London: Routledge.

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*, 385–402.

Kulik, J. A., Kulik, C. L. C., & Bangert, R. L. (1984). Effects of practice on aptitude and achievement test scores. *American Educational Research Journal*, *21*, 435–447.
Leary, L. F., & Dorans, N. J. (1985). Implications for altering the context in which test items appear: a historical perspective on an immediate concern. *Review of Educational Research*, *55*, 387–413.
Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. London: Routledge.
Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174.
Meade, A. W., Lautenschlager, G. J., & Hecht, J. E. (2005). Establishing measurement equivalence and invariance in longitudinal data with item response theory. *International Journal of Testing*, *5*, 279–300.
Meyers, J. L., Miller, G. E., & Way, W. D. (2009). Item position and item difficulty change in an IRT-based common item equating design. *Applied Measurement in Education*, *22*, 38–60.
Millsap, R. E., & Kwok, O.-M. (2004). Evaluating the impact of partial factorial invariance on selection in two populations. *Psychological Methods*, *9*, 93–115.
Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, *29*, 133–161.
Muraki, E. (1992). A generalized partial credit model: application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159–176.
Muthén, L. K., & Muthén, B. O. (2012). *Mplus user's guide* (7. Aufl.). Los Angeles: Muthén & Muthén.
Nagy, G., & Neumann, M. (2010). Psychometrische Aspekte des Tests zu den voruniversitären Mathematikleistungen in TOSCA-2002 und TOSCA-2006: Unterrichtsvalidität, Rasch-Homogenität und Messäquivalenz. In U. Trautwein, M. Neumann, G. Nagy, O. Lüdtke & K. Maaz (Hrsg.), *Schulleistungen von Abiturienten: Die neu geordnete gymnasiale Oberstufe auf dem Prüfstand* (S. 281–306). Wiesbaden: VS.
Organization for Economic Cooperation and Development (OECD) (2009). *PISA data analysis manual*. Paris: OECD Publishing.
Organization for Economic Cooperation and Development (OECD) (2014). *PISA 2012 technical report*. Paris: OECD Publishing.
Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: a review of reporting practices and suggestions for improvement. *Review of educational research*, *74*, 525–556.
Pohl, S., & Carstensen, C. H. (2013). Scaling of competence tests in the national education panel study – many questions, some answers, and further challenges. *Journal for Educational Research Online*, *5*, 189–216.
Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Kopenhagen: Paedagogiske Institut.
Ramm, G., Prenzel, M., Baumert, J., Blum, W., Lehmann, R., Leutner, D., Neubrand, M., Pekrun, H.-G. R., Rost, J., & Schiefele, U. (Hrsg.). (2006). *PISA 2003: Dokumentation der Erhebungsinstrumente*. Münster: Waxmann.
Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance. *Psychological Bulletin*, *114*, 552–566.
te Marvelde, J. M., Glas, C. A., Van Landeghem, G., & Van Damme, J. (2006). Application of multidimensional item response theory models to longitudinal data. *Educational and Psychological Measurement*, *66*, 5–34.
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*, 4–70.
von Davier, M., Xu, X., & Carstensen, C. H. (2011). Measuring growth in a longitudinal large-scale assessment with a general latent variable model. *Psychometrika*, *76*, 318–336.
Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation*, *31*, 114–128.
Wu, M. (2010). Measurement, sampling, and equating errors in large-scale assessments. *Educational Measurement: Issues and Practice*, *29*, 15–27.
Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). *ACERConQuest, Version 2.0: generalized item response modelling software*. Camberwell: Australian council for Educational Research.