Z Erziehungswiss (2017) (Suppl 2) 20:259–286
DOI 10.1007/s11618-017-0755-1

ZfE CrossMark

Längsschnittskalierung der Tests zur Überprüfung des Erreichens der Bildungsstandards der Sekundarstufe I im PISA-Längsschnitt 2012/2013

Gabriel Nagy · Nicole Haag · Lüdtke Oliver · Olaf Köller

Online publiziert: 9. Mai 2017

Zusammenfassung Der vorliegende Beitrag widmet sich der Skalierung der in der PISA-Längsschnittstudie 2012/2013 verwendeten Tests zur Überprüfung des Erreichens der Bildungsstandards für den Mittleren Schulabschluss. Vorgestellt werden Analysen zur Übereinstimmung der im Rahmen der Retesterhebung geschätzten Itemparameter mit den in der Ländervergleichsstudie 2012 kalibrierten Parametern sowie die Schätzung individueller Kompetenzniveaus. Darüber hinaus werden Analysen zu den Konsequenzen des in der Retesterhebung verwendeten nicht-balancierten Testdesigns vorgestellt. Es zeigte sich, dass die ermittelten Itemparameter sowohl in gymnasialen als auch in nichtgymnasialen Schulformen eine sehr hohe Übereinstimmung mit den bereits kalibrierten Parametern aufwiesen. Die mittels der Plausible-Value-Technik geschätzten Kompetenzniveaus indizierten sowohl für nichtgymnasiale als auch für gymnasiale Schulformen mit wenigen Ausnahmen Kompetenzzuwächse im Laufe des 10. Schuljahres. Weiterführende Analysen deuteten jedoch drauf hin, dass aufgrund des am zweiten Erhebungszeitpunkt nicht-balancierten Testdesigns mit Verzerrungen bei der Zuwachsschätzung zu rechnen ist. Implikationen der Befunde für die Auswertungen der Leistungszuwächse werden diskutiert.

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

Dr. N. Haag
Mathematik, Institut zur Qualitätsentwicklung im Bildungswesen, Hannoversche Straße 19, 10099 Berlin, Deutschland
E-Mail: nicole.haag@iqb.hu-berlin.de

Prof. Dr. L. Oliver · 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

Schlüsselwörter PISA-Studie · Item Response Theorie · Leistungszuwächse · Nicht-balancierte Testdesigns

Longitudinal IRT Scaling of Tests of the Educational Standards for Lower Secondary Level in the PISA Longitudinal Assessment 2012/2013

Abstract The present article is devoted to the IRT scaling of tests used in the PISA Longitudinal Study 2012/2013 which verify educational standards for lower secondary school certification. It presents analyses investigating the agreement of freely estimated item parameters with the parameters calibrated in the national assessment 2012, and describes the estimation of competence levels. In addition, analyses are presented on the consequences of the unbalanced test design implemented in the retest assessment. Results indicated that the item parameters estimated for non-academic and academic track school types closely matched the pre-calibrated parameters. With few exceptions increases in competence levels during the 10th grade were estimated for both academic and non-academic track school types on basis of ability parameters estimated by means of the plausible value technique. Further analyses suggested, however, that distortions in the growth estimate are to be expected due to the unbalanced test design administered in the second survey period. Implications of these findings for the evaluation of the competence gains are discussed.

Keywords PISA Study · Item Response Theory · Performance gains · Unbalanced test designs

Fachbezogene Bildungsstandards geben die Zielgrößen des Schulunterrichts in Form von Lernergebnissen vor. Konkret legen sie fest, welche Kompetenzen Schülerinnen und Schüler mit Beendigung bestimmter Abschnitte ihrer Schullaufbahn erworben haben sollten (Klieme et al. 2003). Ende 2003 wurden die Bildungsstandards für den Mittleren Schulabschluss für die Fächer Deutsch, Mathematik, Englisch und Französisch verabschiedet. Weitere Standards wurden in den Folgejahren für den Hauptschulabschluss (Deutsch, Mathematik, Englisch und Französisch), den Primarbereich (Deutsch und Mathematik) und die Abiturprüfung (Deutsch, Mathematik, Englisch und Französisch) verabschiedet. Es liegen auch Bildungsstandards für die naturwissenschaftlichen Fächer (Biologie, Physik und Chemie) für den Mittleren Schulabschluss vor.

Parallel zur Entwicklung der Bildungsstandards wurde im Auftrag der Kultusministerkonferenz (KMK) an der Erstellung von Kompetenztests zur Überprüfung der Erreichung der Bildungsstandards (sog. BISTA-Tests) gearbeitet, wobei die entsprechenden Arbeiten am Institut zur Qualitätsentwicklung im Bildungswesen (IQB) angesiedelt sind. Momentan liegen empirisch erprobte BISTA-Tests für unterschiedliche Fächer und Bildungsabschnitte vor (Pant et al. 2013). Diese Tests bilden die Grundlage der Überprüfung des Erreichens der Bildungsstandards in den Ländern der Bundesrepublik Deutschland. Die entsprechenden ländervergleichend angelegten Studien (IQB-Ländervergleiche, ab 2015 IQB-Bildungstrend) werden in der Se-

kundarstufe I alle drei Jahre durchgeführt und entsprechen damit den Abständen der PISA-Erhebungen. Ab dem Jahr 2009 haben die auf Basis der BISTA-Tests durchgeführten Ländervergleichsstudien die bis dahin mittels der PISA-Tests durchgeführten Vergleichsstudien (sog. PISA-E-Studien) abgelöst.

Die BISTA-Tests zeichnen sich durch einen starken Bezug zum deutschen Schulcurriculum aus. Dies folgt unmittelbar aus der Tatsache, dass Tests wie auch die neuen Curricula parallel zu den Bildungsstandards entwickelt wurden. Im Gegensatz dazu ist die Entsprechung der PISA-Tests zum deutschen Schulcurriculum geringer, da deren Inhalte dem Literacy-Konzept folgen und nicht lehrplanbasiert sind (OECD 2014). Damit steigt das Risiko, dass Lerninhalte, die im Fokus bestimmter Bildungsabschnitte des deutschen Schulsystems stehen, unzureichend mithilfe der PISA-Instrumente abgebildet werden.

Aufgrund der hohen Entsprechung der Inhalte der BISTA-Tests zu den Curricula der 16 Länder ist zu erwarten, dass die entsprechenden Tests sensitiv für die Kompetenzzuwächse in den entsprechenden Bildungsetappen sind (Wagner et al. 2014). Damit einher geht jedoch das Risiko der Verletzung der quer- und längsschnitlichen Messäquivalenz. So ist vorstellbar, dass einzelne Teilbereiche eines Fachs (z. B. Stochastik vs. Funktionen im Bereich Mathematik) in verschiedenen Schulformen in unterschiedlichen Jahrgangsstufen behandelt werden (Klieme und Baumert 2001). In diesem Fall ist zu erwarten, dass sich Schwierigkeiten der eingesetzten Einzelitems in Abhängigkeit der erfassten Teilbereiche zwischen Schulformen unterscheiden, sodass sich Schulformunterschiede unter Umständen nicht über die Gesamtheit der Testinhalte hinweg generalisieren lassen (vgl. Nagy et al. 2017). Ebenso führt eine Situation, in der sich die Schwerpunkte des Unterrichts systematisch zwischen den Messzeitpunkten unterscheiden, zu Veränderungen in den Itemschwierigkeiten, die in Abhängigkeiten der Teilbereiche verschieden hoch ausfallen können. In der Konsequenz lassen sich Kompetenzzuwächse nicht hinreichend über den gesamten Test hinweg generalisieren (vgl. Nagy et al. 2017).[1]

Wie von Nagy et al. (2017) verdeutlicht wurde, sind die zuvor skizzierten Verstöße gegen das Ideal der quer- und längsschnittlichen Messinvarianz, die in Abhängigkeit der Merkmale der verwendeten Items auftreten (z. B. Teilbereiche eines Fachs, Vertrautheit mit Darstellungsformen usw.) nicht die einzigen Störquellen, die sich auf die ermittelten Kompetenzunterschiede auswirken können. Andere Störeffekte sind auf sogenannte Textkontexteffekte (Brennan 1992; Leary und Dorans 1985) zurückzuführen, die zum Ausdruck bringen, dass die Wahrscheinlichkeit einer korrekten Itemlösung vom (Test-)Kontext, in dem ein Item eingebettet ist, abhängt. Eine Situation, in der das Testdesign zwischen zwei Erhebungszeitpunkten verändert wird (z. B. Änderung der Anordnung von Items in einem Test), kann sich somit auf die

[1] Dieser Sichtweise lässt sich entgegnen, dass die BISTA-Tests den Anspruch erheben, das Erreichen der Bildungsstandards über die Gesamtheit der Unterrichtsinhalte eines Fachs zu erheben, sodass gruppen- und/oder zeitspezifische Teildefizite unerheblich für die Interpretation der Gesamttestleistungen und deren Zuwächse sind. Diese Argumentation setzt unserer Meinung jedoch voraus, dass die in den Bildungsstandards vorgenommene Gewichtung der Unterrichtsinhalte mit der Gewichtung der entsprechenden Inhalte in den verwendeten Tests (approximativ) übereinstimmt. Da zum zweiten Messzeitpunkt des PISA-Längsschnitts 2012/2013 nur eine vergleichsweise kleine Teilmenge der BISTA-Items eingesetzt wurde (vgl. Tab. 1), ist davon auszugehen, dass diese Voraussetzung mit hoher Wahrscheinlichkeit nicht erfüllt ist.

Schätzung von Kompetenzzuwächsen auswirken. Hinzukommt, dass die Abschätzung von Gruppenunterschieden in Kompetenzzuwächsen verzerrt ausfallen kann, wenn sich die Höhe von Testkontexteffekten zwischen Gruppen unterscheidet (vgl. Nagy et al. 2017).

Gegenstand des vorliegenden Beitrags ist die Längsschnittskalierung der im Rahmen der PISA-Längsschnittstudie 2012/2013 erhobenen BISTA-Tests (Bereiche: Mathematik, Physik, Chemie und Biologie). Ein Ziel des PISA-Längsschnitts 2012/2013 war es, die Kompetenzzuwächse mittels der BISTA-Tests zu bestimmen. Dies setzte die Verlinkung der zum zweiten Messzeitpunkt des PISA-Längsschnitts 2012/2013 eingesetzten BISTA-Tests mit den parallel zur PISA-Ausgangserhebung erhobenen Kompetenztests der Ländervergleichsstudie 2012 voraus. Vor dem Hintergrund der zuvor angesprochenen Störquellen fokussiert der Beitrag auf Fragen zur längsschnittlichen Messinvarianz der BISTA-Tests. Darüber hinaus wird die Schätzung der Kompetenzniveaus mittels Messmodellen der Item Response Theorie (IRT) vorgestellt. Zudem werden systematische Verzerrungen von Kompetenzzuwächsen, die sich in Folge von Änderungen des eingesetzten Testdesigns ergeben können, untersucht und deren Implikationen für die Auswertung der Kompetenzdaten diskutiert.

1 Die vorliegende Studie

Die Untersuchung der Messeigenschaften des im PISA-Längsschnitt verwendeten BISTA-Tests unterscheidet von der Analyse der PISA-Tests, die von Nagy et al. (2017) vorgestellt wurde. Analog zum Vorgehen beim PISA-Test, wurden mögliche Verstöße gegen die Annahme der Messinvarianz zwischen Schulformen und Messzeitpunkten untersucht (Reise et al. 1993). Allerdings wurde beim BISTA-Test nicht die Kontrolle von Testkontexteffekten auf die ermittelten individuellen Kompetenzausprägungen angestrebt, die im Fall der PISA-Tests insbesondere in Form von Positionseffekten (Meyers et al. 2009) auftraten. Stattdessen wurde der Versuch unternommen, die Einflüsse der Veränderung des Testkontexts auf die ermittelten Kompetenzzuwächse abzuschätzen. Damit war es möglich, mittlere Kompetenzzuwächse zu bestimmen, die (approximativ) um die Effekte einer veränderten Testkonfiguration adjustiert wurden.

Auf eine Adjustierung der individuellen Kompetenzausprägungen um mögliche Testkontexteffekte wurde beim BISTA-Test verzichtet, da hierzu die Kontrolle von Testkontexteffekten zu beiden Messzeitpunkten notwendig ist. Dieser Ansatz setzt voraus, dass zumindest ein Teil der wiederholt getesteten Schülerinnen und Schüler in jeder getesteten Domäne Gruppen von Items (d. h. Itemcluster) bearbeitet haben, die an der ersten Stelle des Tests vorgelegt wurden (vgl. Nagy et al. 2017). Diese Voraussetzung war jedoch aufgrund des zum zweiten Messzeitpunkt verwendeten Testdesigns für drei der insgesamt sieben mittels der BISTA-Tests erhobenen Kompetenzdimensionen nicht erfüllt (s. unten).

Aus diesem Grund wurden die im Rahmen der Ländervergleichsstudie 2012 erzeugten Kompetenzwerte, die in Form von Plausible Values (PVs; Wu 2005) vorlagen, beibehalten (für eine Beschreibung der Skalierung der BISTA-Tests siehe Hecht

et al. 2013). Um die Vergleichbarkeit der im Rahmen des zweiten Messzeitpunkts des PISA-Längsschnitts 2012/2013 erhobenen Kompetenzen mit den Ausgangskompetenzen abschätzen zu können, wurden die neu erhobenen Daten aus verschiedenen Perspektiven evaluiert.

2 Verstöße gegen die Invarianz der Itemparameter in Abhängigkeit von Itemmerkmalen

In einem ersten Schritt wurde die Übereinstimmung der in der Ländervergleichsstudie 2012 kalibrierten Itemparameter mit den zum zweiten Messzeitpunkt des PISA-Längsschnitts in der Gesamtstichprobe und in den verschiedenen Schulformen frei geschätzten Parameter evaluiert. Diese Analysen können Hinweise auf eine Veränderung der Messeigenschaften der Testaufgaben in Abhängigkeit von deren Itemmerkmalen geben (z. B. optimale Lösungswege, Zugehörigkeit zu Teilbereichen eines Fachs). Ergebnisse, die keine Hinweise auf nennenswerte Verstöße gegen die Messinvarianzannahme indizieren, sind eine Voraussetzung für die Vergleichbarkeit der Bedeutung der innerhalb einer Domäne erfassten Kompetenzen zwischen Gruppen und/oder Messzeitpunkten (Meade et al. 2005). Invariante Itemparameter implizieren, dass mit einem Anstieg der zugrundeliegenden Kompetenzdimension um einen bestimmten Wert sich die Lösungswahrscheinlichkeiten[2] aller Items in allen Gruppen und/oder Messzeitpunkten in gleicher Weise ändern (Meredith 1993).

3 Verstöße gegen die Messinvarianz in Abhängigkeit geänderter Testkontexte

In typischen Invarianzstudien wird die Invarianz der absoluten Ausprägungen der Itemparameter als Voraussetzung für die Vergleichbarkeit der Skalenursprünge der latenten Variablen betrachtet (Meredith 1993). Wie von Nagy et al. (2017) demonstriert wurde, können jedoch auch in solchen Situationen die ermittelten Zuwachsschätzungen aufgrund von Testkontexteffekten verzerrt ausfallen. Aus diesem Grund wurden die Auswirkungen von Testkontexteffekten auf die ermittelten Kompetenzveränderungen untersucht. Da eine Adjustierung von Testkontexteffekten auf Ebene der individuellen Messwerte aufgrund der Einschränkungen des verwendeten Testdesigns für den BISTA-Test nicht möglich war (für drei Kompetenzdomänen wurden keine Itemcluster an den ersten Positionen eines Booklets vorgelegt), haben wir uns auf die Untersuchung der Auswirkungen der zum zweiten Messzeitpunkt vorgenommenen Änderungen des Testdesigns auf die mittleren Zuwachsschätzungen beschränkt. Vor dem Hintergrund der aus der Literatur bekannten Befundlage, dass die in Large-Scale-Assessments vorliegenden Testkontexteffekte maßgeblich Positionseffekten geschuldet sind (Leary und Dorans 1985; Meyers et al. 2009; Nagy et al. 2017), sind wir wie folgt vorgegangen.

[2] Im Rahmen des hier verwendeten einparametrischen Raschmodells gilt diese Annahme strenggenommen für die Logits der Lösungswahrscheinlichkeiten.

In einem ersten Schritt haben wir die Variation der Zuwachsschätzungen in Abhängigkeit der zum zweiten Messzeitpunkt bearbeiteten Booklets untersucht. Sofern geschätzte Kompetenzzuwächse von Positionseffekten betroffen sind, sollte beobachtbar sein, dass sie in Booklets, in denen die Items einer Inhaltsdomäne im Mittel näher am Anfang positioniert waren, höher ausfallen als die Kompetenzzuwächse, die aufgrund von Booklets geschätzt wurden, in denen die Items näher am Ende positioniert waren. Diese Erwartung basiert auf der Tatsache, dass das zum ersten Messzeitpunkt verwendete Testdesign hinsichtlich der Itemclusterpositionen balanciert war und die Zuordnung der Schülerinnen und Schüler auf Booklets zu beiden Messzeitpunkten randomisiert vorgenommen wurde. Daraus folgt, dass die Schülerinnen und Schüler, die zum zweiten Messzeitpunkt ein bestimmtes Booklet bearbeitet haben, im Mittel die entsprechenden Items an der mittleren Itemclusterposition bearbeitet haben.

Im zweiten Schritt wurde die Höhe von Kompetenzzuwächsen abgeschätzt, die sich bei einer mittleren Position der zur Erfassung eines Kompetenzbereichs verwendeten Itemcluster zu beiden Messzeitpunkten ergeben würde. Da zum ersten Messzeitpunkt die Darbietung des Testmaterials an der mittleren Position per Design sichergestellt wurde, aber das zum zweiten Messzeitpunkt eingesetzte Design nicht hinsichtlich der Positionen balanciert war, wurden die entsprechenden Zuwächse mittels einer linearen Funktion approximiert. Hinweise für eine verzerrte Zuwachsschätzung liegen dann vor, wenn die über die Booklets des zweiten Messzeitpunkts gemittelten Kompetenzzuwächse von den an der mittleren Position erwarteten Zuwachsschätzungen abweichen.

4 Methode

4.1 Stichprobe und Testheftdesign

Die Auswertungen basierten auf der von Heine et al. (2017) beschriebenen Längsschnittstichprobe von $N = 6584$ Schülerinnen und Schülern (45,2 % Gymnasiastinnen und Gymnasiasten und 54,8 % Schülerinnen und Schüler an nichtgymnasialen Schulformen). Für diese Schülergruppen lagen IRT-skalierte Leistungswerte (jeweils 15 PVs pro Domäne) auf den im Rahmen der Bildungsvergleichsstudie eingesetzten BISTA-Tests vor, wobei sich aufgrund des im Jahr 2012 eingesetzten Testdesigns die Fallzahlen für die einzelnen Domänen voneinander unterschieden. Konkret lagen für $N = 3351$ Schülerinnen und Schüler PVs auf der Globalskala Mathematik vor und für $N = 3342$ Schülerinnen und Schüler lagen PVs auf den naturwissenschaftlichen Kompetenzdimensionen vor. Die geringe Überlappung der beiden Gruppen ist auf das zum ersten Messzeitpunkt verwendete Testdesign zurückzuführen, das lediglich 8 von 70 Booklets mit Items aus allen Inhaltsbereichen umfasste (Hecht et al., 2013).

Von der Ausgangsstichprobe nahmen $N = 4610$ Schülerinnen und Schüler an der Retesterhebung des PISA-Längsschnitts 2012/2013 teil (Heine et al. 2017). Das zur zweiten Welle verwendete Testheftdesign umfasste insgesamt 18 Testhefte, von denen 14 Hefte auch Itemcluster aus den BISTA-Tests beinhalteten. Da für die in

Tab. 1 Itemcluster mit BISTA-Items im Booklet-Design der Retesterhebung der PISA-Längsschnittstudie 2012/2013

	Position 1	Position 2	Position 3	Position 4	Position 5	Position 6
Booklet 04	PH-EG03	CH-FW05	BI-EG05	CH-EG02	BI-FW04	PH-FW04
Booklet 05	BI-FW04	CH-EG02	PH-FW04	CH-FW05	PH-EG03	BI-EG05
Booklet 06	–	–	–	MA-2En	MA-2B	MA-1B1
Booklet 07	MA-5B	MA-4G	MA-4B3	–	–	–
Booklet 08	MA-4B3	MA-5B	MA-2B	–	–	–
Booklet 09	–	–	–	MA-1B1	MA-4G	MA-5B
Booklet 10	–	–	–	PH-EG03	CH-EG02	BI-EG05
Booklet 11	PH-FW04	CH-FW05	BI-FW04	–	–	–
Booklet 12	MA-2B	BI-FW04	PH-FW04	–	–	–
Booklet 13	CH-FW05	PH-EG03	MA-4G	MA-4B3	MA-1B1	MA-2En
Booklet 14	–	–	–	BI-EG05	MA-4B3	CH-EG02
Booklet 15	MA-2B	CH-FW05	PH-EG03	–	–	–
Booklet 16	–	–	–	MA-2En	MA-5B	MA-4G
Booklet 18	MA-2En	MA-2B	MA-1B1	–	–	–

(–) nicht berücksichtigte Items
MA Mathematik, *PH* Physik, *CH* Chemie, *BI* Biologie, *FW* Fachwissen, *EG* Erkenntnisgewinnung

den BISTA-Tests verwendeten Itemcluster jeweils eine Bearbeitungszeit von 20 min veranschlagt wurde, konnten die zur zweiten Welle eingesetzten Testhefte maximal sechs Itemcluster umfassen. Dies traf auf insgesamt drei Booklets zu. Die weiteren 11 Booklets umfassten jeweils 5 Itemcluster (d. h. 2 PISA-Cluster und 3 BISTA-Cluster). Die Zusammenstellung der Booklets gewährleistete eine Bearbeitungszeit von ca. 20 min für jedes BISTA-Itemcluster in jedem Booklet.

Die zur zweiten Welle des PISA-Längsschnitts 2012/2013 eingesetzten BISTA-Cluster sind in Tab. 1 getrennt nach Booklets und Testheftpositionen dargestellt. Wie aus der Tabelle hervorgeht, haben wir zwischen sechs Positionen unterschieden. Des Weiteren wird ersichtlich, dass in der zweiten Welle der PISA-Längsschnitt 2012/2013 unterschiedlich viele Items zur Erfassung der verschiedenen Kompetenzdimensionen verwendet wurden. Für den Bereich Mathematik wurden insgesamt 6 Itemcluster eingesetzt, die sich auf insgesamt 10 Booklets verteilten. Für die naturwissenschaftlichen Fächer Physik, Chemie und Biologie kamen hingegen nur jeweils zwei Booklets zum Einsatz, wobei den Kompetenzbereichen Fachwissen und Erkenntnisgewinnung jeweils ein Itemcluster zugeordnet wurde. Außerdem gilt es bezüglich des Testdesigns zu beachten, dass die Darbietung des Testmaterials nicht ausbalanciert hinsichtlich der Itemclusterpositionen erfolgte. Damit verbunden ist, dass die über Booklets gemittelten Positionen der den verschiedenen Kompetenzdimensionen zugeordneten Itemcluster sich zum Teil deutlich voneinander unterschieden. So lag die mittlere Itemclusterposition für die Bereiche Mathematik (3,52) und Physik-Fachwissen (3,25) relativ nahe an der mittleren Position eines balancierten Designs (3,50). Bei der Erfassung der Dimensionen Physik-Erkenntnisgewinnung (3,00), Chemie-Fachwissen (2,20) und Biologie-Fachwissen (2,75) waren die Itemcluster im Mittel näher am Anfang der Booklets lokalisiert. Im Fall der Dimensionen

Chemie-Erkenntnisgewinnung (4,25) und Biologie-Erkenntnisgewinnung (4,75) waren die Itemcluster stärker in Richtung Testende verschoben.

5 Statistische Analysen

Die Untersuchung der Messeigenschaften der BISTA-Tests geschah mittels IRT-basierter Analysen. In Übereinstimmung mit dem in den Ländervergleichsstudien gewählten Vorgehen haben wir das einparametrische Raschmodell für dichotome Items verwendet (Rasch 1960). Die Untersuchung der Messeigenschaften erfolgte in zwei Schritten, die für jede Kompetenzdimension separat durchgeführt wurden.

Invarianzverstöße in Abhängigkeit der Itemmerkmale. Zur Untersuchung von Verstößen gegen die Invarianz der Itemparameter, die in Abhängigkeit der Itemmerkmale auftreten können (d. h., nicht der Positionierung in den Testbooklets geschuldet waren), wurden die Schwierigkeiten der Items zum zweiten Messzeitpunkt des PISA-Längsschnitts 2012/2013 empirisch in der Gesamtstichprobe sowie in der nach Schulformen aufgeteilten Stichprobe bestimmt. Die entsprechenden Schätzungen wurden mit den Itemparametern der Ländervergleichsstudie 2012 zuerst mittels graphischer Methoden verglichen (Bejar 1980). Im optimalen Fall sollte der Zusammenhang der frei geschätzten Itemparameter mit den Parametern der Ländervergleichsstudie näherungsweise einer perfekten linearen Funktion mit Einheitssteigung folgen. Ein solches Befundmuster impliziert, dass die Daten hinreichend gut durch ein Messmodell mit vollständig invarianten Itemparametern beschrieben werden können.

In empirischen Anwendungen ist davon auszugehen, dass selbst bei einer hohen Übereinstimmung der Itemparameter 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, wurde die Übereinstimmung der Itemparameter zusätzlich mittels Statistiken des *differential item functioning* (DIF; Holland und Wainer 1993) untersucht, die gemäß des vom Educational Testing Service (ETS) vorgeschlagenen Systems (Clauser und Mazor 1998) klassifiziert wurden: (A) nicht von DIF betroffen, (B) kaum von DIF betroffen und (C) stark von DIF betroffen.[3]

[3] DIF-Statistiken wurden anhand der Itemparameter aus der IQB-Ländervergleichsstichprobe und den zum zweiten Messzeitpunkt des PISA-Längsschnitts 2012/2013 frei geschätzten Itemparametern berechnet. Konkret ergab sich die DIF-Statistik für ein Item j als $DIF_j = \left(\beta_{1j} - \acute{\beta}_1\right) - \left(\beta_{2j} - \acute{\beta}_2\right)$, wobei β_{1j} für die Schwierigkeit des Items j in der Ländervergleichsstichprobe 2012 und β_{2j} für die Schwierigkeit von Item j zum zweiten Messzeitpunkt des PISA-Längsschnitts 2012/2013 steht. In beiden Fällen stehen die $\acute{\beta}$-Parameter für die mittleren Schwierigkeiten der betrachteten Items (d. h. die Mittelwerte über Items, die in beiden Erhebungen eingingen). 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.

Da die BISTA-Tests den Anspruch erheben, das Erreichen der Bildungsstandards über die Gesamtheit der Unterrichtsinhalte eines Fachs zu erfassen, können geringere Abweichungen vom Ideal der perfekten Messinvarianz (d. h. exakte Übereinstimmung aller Itemparameter) toleriert werden. Aus diesem Grund erachteten wir die BISTA-Tests einer Inhaltsdomäne dann als hinreichend messinvariant, wenn (1) sich augenscheinlich keine Gruppen von Items identifizieren ließen, die gemeinsam von der linearen Funktion abwichen (graphische Analyse), (2) die Korrelationen der frei geschätzten Itemparameter mit den präkalibrierten Parameter aus dem Ländervergleich hoch war ($r > 0{,}90$), (3) die Items im Mittel absolute DIF-Werte aufwiesen, die unterhalb der Schwelle für stark von DIF betroffene Items lagen ($|DIF| < 0{,}4$) und (4) Items mit hohen DIF-Werten (d. h. $|DIF| > 0{,}4$) die Ausnahme waren.

Auswirkungen von Testkontexteffekten auf die Zuwachsschätzung. Die Untersuchung der Abhängigkeit der Zuwachsschätzungen von Testkontexten erfolgte mit Hilfe eines um Bookleteffekte erweiterten Messmodells. In diesem Modell wurde die Wahrscheinlichkeit einer richtigen Antwort des Individuums i auf dem Item j zum Messzeitpunkt 2 als eine Funktion der vorliegenden Ausgangswerte $\widehat{\theta}_{i1}$ (d. h. PV-Wert des Individuums i zum ersten Messzeitpunkt) und einer latenten Leistungsänderung δ_{i} ausgedrückt:

$$P\left(y_{ij2}=1\right)=\frac{exp\left(\widehat{\theta}_{i1}+\delta_{\mathrm{i}}-\beta_{\mathrm{j2}}\right)}{1+exp\left(\widehat{\theta}_{i1}+\delta_{\mathrm{i}}-\beta_{\mathrm{j2}}\right)}, \tag{1}$$

wobei β_{j2} für die Schwierigkeit des Items j zum zweiten Messzeitpunkt steht. Die Itemschwierigkeiten wurden abhängig von den Ergebnissen des ersten Analyseschritts entweder auf die vorliegenden Werte aus der Ländervergleichsstudie fixiert oder abweichend davon geschätzt.

Das in Gl. 1 dargestellte Modell entspricht einem latenten Differenzwertemodel (von Davier et al. 2011), mit der Ausnahme, dass die Ausgangsleistungen nicht mittels eines Messmodells modelliert werden, sondern aus den Ergebnissen einer vorgeschalteten Skalierung entnommen wurden (jeweils 15 PVs pro Kompetenzdimension, s. unten). Testkontexteffekte wurden durch die Regression des Leistungszuwachses auf das zur zweiten Welle bearbeitete Booklet modelliert:

$$\delta_{\mathrm{i}}=\alpha+\sum_{\mathrm{b}=1}^{\mathrm{B}}\gamma_{\mathrm{b}}+\zeta_{\mathrm{i}}, \tag{2}$$

wobei γ_b den Effekt des Booklets b darstellt, α ein Interceptparameter ist und ζ_i ein individueller Abweichungsterm mit Erwartungswert 0 darstellt. Aus Gründen der Modellidentifikation wurde die Summe der Bookleteffekte auf 0 restringiert ($\sum_{b=1}^{B}\gamma_b=0$), sodass α die mittlere Kompetenzänderung über alle Booklets erfasst.[4]

Die Ergebnisse des in den Gln. 1 und 2 dargestellten Modells quantifizieren die Abhängigkeit der Zuwachsschätzung vom bearbeiteten Booklet. Insofern alle

[4] Das in Gl. 2 dargestellte Modell wurde mittels effektkodierter Bookletindikatoren geschätzt. Die gewählte Kodierung der Booklets stellt sicher, dass sich deren Effekte zu 0 summieren.

γ-Parameter nahe 0 ausfallen, hängt die Zuwachsschätzung nicht vom bearbeiteten Booklet ab, während positive (oder negative) Ausprägungen der γ-Parameter indizieren, dass mit der Bearbeitung eines Booklets b Leistungszuwächse einhergehen, die höher (oder geringer) als die mittleren Zuwächse ausfallen. Die entsprechenden Effekte können als Testkontexteffekte verstanden werden, da jedes Booklet einen spezifischen Kontext darstellt, in dem die Items eingebunden sind. Die Interpretation der so ermittelten Effekte ist aufgrund der zufälligen Zuordnung von Schülerinnen und Schülern auf die zu beiden Messzeitpunkten vorgelegten Booklets gewährleistet. Die γ-Parameter quantifizieren den mittleren Effekt des zum zweiten Messzeitpunkt bearbeiteten Booklets über alle Booklets des ersten Messzeitpunkts.

Zur Untersuchung von Schulformunterschieden in den Bookleteffekten γ_b wurden im zweiten Schritt die Modelle um einen dummykodierten Schulformindikator $g = 0$, 1 (0 = nichtgymnasiale Schulformen, 1 = Gymnasien) erweitert. Das Messmodell in Gl. 1 wurde unverändert beibehalten, wohingegen das Strukturmodell (siehe Gl. 2) wie folgt spezifiziert wurde:

$$\delta_i = \alpha + \gamma_g g_i + \sum_{b=1}^{B} (\gamma_{0b} + \gamma_{1b} g_i) + \zeta_i. \tag{3}$$

In diesem Modell stellt γ_g den Gruppenunterschied im mittleren Leistungszuwachs dar (z. B. Gymnasialeffekt) und α steht für den mittleren Leistungsanstieg in der Referenzgruppe (z. B. nichtgymnasiale Schulformen). Das Modell umfasst ferner Bookleteffekte auf die Zuwachsschätzung (γ_{0b}) sowie die Interaktion der Bookleteffekte mit der Gruppierungsvariable (γ_{1b}), wobei die γ_{0b}-Parameter für die Bookleteffekte in der Referenzgruppe und die γ_{1b}-Parameter für die Unterschiede in den Bookleteffekten in der Vergleichsgruppe relativ zur Referenzgruppe stehen.

Die in den Gln. 2 und 3 dargestellten Bookleteffekte wurden hinsichtlich Positionseffekten evaluiert. Ein Indiz für deren Interpretation als Positionseffekte ist dann gegeben, wenn die Bookletunterschiede in den geschätzten Kompetenzzuwächsen in Übereinstimmung mit der mittleren Position der Itemcluster in den jeweiligen Booklets ausfallen. Dieser Zusammenhang wurde dazu genutzt, die für eine mittlere Itemclusterposition erwarteten Kompetenzzuwächse abzuschätzen. Aufgrund des zum zweiten Messzeitpunkt umgesetzten Testdesigns haben wir hierzu die Bookleteffekte mittels einer linearen Funktion approximiert. Im Fall des in Gl. 2 ausgewiesenen Modells wurde der Bookleteffekt γ_b als eine Funktion der (zentrierten) mittleren Position der Itemclusters des Booklets b, p_b, dargestellt:

$$\gamma_b = a + b\,(p_b - 3.5) + e_b, \tag{4}$$

wobei die Koeffizienten a und b im Sinne eines Regressionsintercepts und eines Regressionsgewichts interpretiert werden können und e_b einen Abweichungsterm darstellt. Die Abweichungsterme summieren sich zu 0 über Bookelts und sind unkorreliert mit den mittleren Itemclusterpositionen der Booklets.

Der Intercept a aus Gl. 4 lässt sich somit als der erwartete Effekt eines hypothetischen Booklets begreifen, in dem die Itemcluster an der mittleren Position vorgelegt

wurden. Er kann verwendet werden, um eine adjustierte Zuwachsschätzung abzuleiten, indem er zum α-Parameter aus Gl. 2 addiert wird (d. h. $\alpha^* = \alpha + a$).

Im Fall des um Schulformeffekte erweiterten Modells (Gl. 3) sind wir analog zur beschriebenen Prozedur vorgegangen. Konkret haben wir den in Gl. 4 beschriebenen Ansatz auf die Bookleteffekte der Referenzgruppe angewandt (d. h. γ_{0b}-Parameter in Gln. 3) und haben das Verfahren für die Bookleteffekte der Vergleichsgruppe wiederholt, die sich als $\gamma_{1b}^* = \gamma_{0b} + \gamma_{1b}$ ergeben. Auf diese Weise wurden schulformspezifische Koeffizienten analog zur Gl. 4 geschätzt, die zur Adjustierung der schulformspezifischen Zuwächse herangezogen werden können.

Modellschätzung. Alle Modelle wurden mithilfe des Programms Mplus 7,4 (Muthén und Muthén 1998–2012) mittels der Marginal Maximum Likelihood Methode (MML) unter Zuhilfenahme der Montecarlo-Integrationsmethode mit 1000 Integrationspunkten geschätzt. In die Analysen zu möglichen Invarianzverstößen, die in Abhängigkeit von Itemmerkmalen auftreten, gingen nur die zum zweiten Messzeitpunkt erfassten Itemantworten der jeweiligen Kompetenzbereiche ein. Die Auswertungen zu den Konsequenzen von Testkontexteffekten basierten auf den zur zweiten Welle erhobenen Itemantworten und Bookletindikatoren sowie den zur ersten Welle vorliegenden PV-Werten. Insgesamt wurden uns vom IQB pro Domäne jeweils 15 PVs zur Verfügung gestellt. Aus diesem Grund wurden die entsprechenden Analysen (Gln. 1 bis 4) für jeden PV wiederholt und die Ergebnisse wurden anschließend gemittelt. Die Standardfehler der Modellparameter in den Gln. 2 und 3 wurden dabei entsprechend der Formel von Rubin (1987) bestimmt. Die entsprechenden Standardfehler wurden ferner dazu genutzt, die Standardfehler der a- und b-Koeffizienten (Gln. 4) und die daraus abgeleiteten adjustierten Kompetenzzuwächse mittels der in Mplus implementierten Delta-Methode zu bestimmen.

Schätzung von Kompetenzniveaus. Die Ergebnisse der Analysen wurden genutzt, um die Metrik der zum zweiten Messzeitpunkt mittels der BISTA-Tests erhobenen Kompetenzen auf die Metrik der vom IQB skalierten PVs (Hecht et al. 2013) zu setzen. Dazu wurden die Kompetenzniveaus zum zweiten Messzeitpunkt auf Grundlage der im ersten Analyseschritt als invariant erachteten Items mittels der PV-Methode geschätzt. Die PV-Schätzung erfolgte je nach Inhaltsdomäne mittels unterschiedlich dimensionierter Messmodelle. Für den Bereich Mathematik wurde ein eindimensionales Rasch-Modell verwendet. In den jeweiligen Naturwissenschaften (Physik, Chemie und Biologie) wurde ein zweidimensionales Modell mit den Dimensionen Fachwissen und Erkenntnisgewinnung spezifiziert.

Die Verwendung der PV-Methode liegt darin begründet, dass die Schülerinnen und Schüler pro Inhaltsdomäne jeweils nur einen kleinen Teil der eingesetzten Testaufgaben bearbeitet haben (vgl. Multi-Matrix-Design in Tab. 1), sodass die Reliabilitäten der zum zweiten Messzeitpunkt generierten Kompetenzwerte im Fall alternativer Schätzverfahren gering ausfallen dürften. Bei der PV-Methode werden neben den Itemantworten weitere Variablen berücksichtigt, die mit den Kompetenzniveaus assoziiert sind (z. B. Kompetenztests in anderen Domänen, familiäre Hintergrundmerkmale usw.). Im vorliegenden Fall wurde bei der Generierung der PVs neben den Itemantworten ein breiter Kranz an Hintergrundvariablen und die Kompetenzausprägungen zum ersten Messzeitpunkt berücksichtigt, wobei wir auf das im Beitrag von

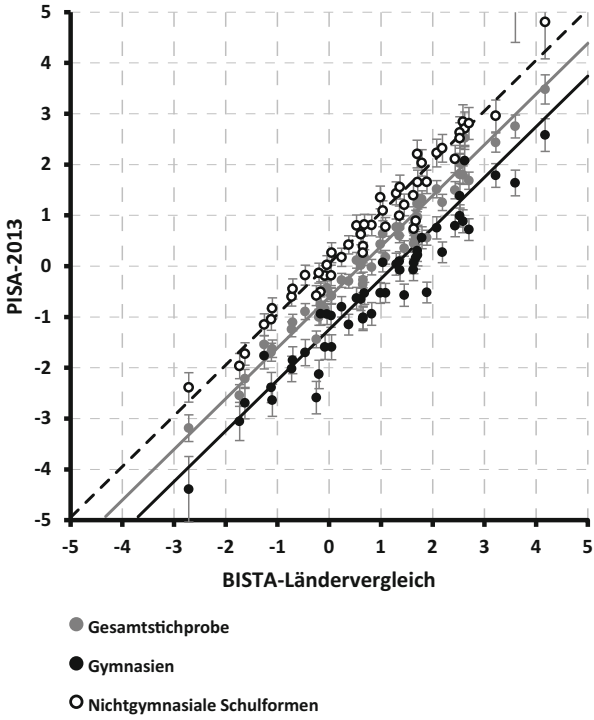

Abb. 1 Übereinstimmung der Itemparameter des Ländervergleichs 2012 (x-Achse) mit den frei geschätzten Itemparametern in der Retesterhebung des PISA-Längsschnitts 2012/2013 (y-Achse) für den Bereich Mathematik. Fehlerbalken indizieren 95 % Konfidenzintervalle

Nagy et al. (2017) verwendete Hintergrundmodell zurückgegriffen haben. Dieses Modell wurde aber im vorliegenden Fall um die zur ersten Welle vorliegenden PV-Werte der BISTA-Kompetenzdimensionen erweitert (s. unten).

Die Generierung der PVs fand mithilfe des Programms ConQuest (Wu et al. 2007) statt. Bei der Interpretation der so generierten Kompetenzschätzungen muss berücksichtigt werden, dass die PVs nicht für den Einfluss möglicher Testkontexteffekte korrigiert wurden. Die Ergebnisse der im zweiten Schritt durchgeführten Analysen erlauben es jedoch, mögliche Verzerrungen abzuschätzen, die auf die Änderung des Testdesigns zurückzuführen sind.

6 Ergebnisse

6.1 Invarianz der Itemparameter über Messzeitpunkte und Schulformen

Im ersten Schritt werden die Ergebnisse zur Übereinstimmung der im Ländervergleich 2012 bestimmten Itemparameter mit den frei geschätzten Parametern des PISA-Längsschnitts 2012/2013 berichtet. Abb. 1 stellt die Zusammenhänge für den Bereich Mathematik dar, wobei die grauen Punkte den Zusammenhang in der Gesamtstichprobe, die schwarzen Punkte den Zusammenhang in der Gruppe der Gymnasiastinnen und Gymnasiasten und die offenen weißen Punkte den Zusammenhang für Schülerinnen und Schüler an nichtgymnasialen Schulformen aufzeigen. Aus der

Abbildung geht hervor, dass alle Zusammenhänge gut mittels einer linearen Funktion mit Einheitssteigung approximiert werden können. Ebenso fielen die Zusammenhänge der Itemparameter in allen Teilstichproben sehr hoch aus (Gesamtstichprobe: r = 0,99, Gymnasien: r = 0,96, nichtgymnasiale Schulformen: r = 0,97). Zudem zeichneten sich keine systematischen Abweichungen von der optimalen linearen Funktion ab. Schließlich belegen die Ergebnisse zu den DIF-Statistiken in Tab. 2, dass die Mathematikitems sowohl in der Gesamtgruppe, als auch in den nach Schulformen aufgeteilten Teilstichproben mittlere DIF-Statistiken im tolerierbaren Bereich lagen und nur ein sehr kleiner Teil der Items starken DIF nach der ETS-Klassifikation aufwiesen. Dieses Ergebnismuster weist somit darauf hin, dass die Itemparameter in allen Teilgruppen als invariant zwischen den Messzeitpunkten angesehen werden können.

Abb. 2 fasst analog zur Abb. 1 die Zusammenhänge der Itemparameter der naturwissenschaftlichen Dimensionen zusammen. Aus der Abbildung kann entnommen werden, dass die Zusammenhänge aller Itemparameter durchweg sehr hoch ausfielen und dass alle Zusammenhänge gut mittels einer linearen Funktion mit einer Einheitssteigung beschrieben werden konnten (alle $r > 0{,}94$) und auch in diesen Kompetenzdomänen keine systematisch vom Invarianzkriterium abweichende Itemgruppen identifiziert werden konnten. Wie aus Tab. 2 hervorgeht, befanden sich die mittleren DIF-Statistiken in allen Domänen und Subgruppen im tolerierbaren Bereich. Mit Ausnahme der Kompetenzbereiche „Physik Fachwissen“ und „Chemie Erkenntnisgewinnung“ wurden keine stark von DIF betroffenen Items ermittelt, wobei in diesen Domänen nur 3 von 14 bzw. 2 von 13 Items der DIF-Kategorie C zugeordnet wurden und diese Abweichungen nur in der Gymnasialgruppe auftraten.

Zusammenfassend zeigen die Befunde, dass keine nennenswerten Verstöße gegen das Invarianzkriterium vorlagen. Dieses Befundmuster legt somit nahe, dass die im Rahmen der Ländervergleichsstudie 2012 kalibrierten Itemparameter auch in der Retestwelle des PISA-Längsschnitts verwendet werden können, da sich die relative Ausprägung der Itemschwierigkeiten zueinander zum zweiten Messzeitpunkt der Studie in keiner der betrachteten Subgruppen systematisch von der Abfolge der Itemschwierigkeiten in der Ländervergleichsstudie 2012 unterschied.

7 Auswirkungen nicht-balancierter Testdesigns auf die Zuwachsschätzung

Die in diesem Teilabschnitt vorgestellten Auswertungen beschäftigten sich mit Unterschieden zwischen Zuwachsschätzungen in Abhängigkeit von den zum zweiten Messzeitpunkt bearbeiteten Booklets. Die entsprechenden Auswertungen beruhen auf den zuvor dargestellten Modellen (Gln. 1 bis 4). Aufgrund der eben vorgestellten Ergebnisse wurden in diesen Modellen die Itemparameter der zur zweiten Welle erhobenen BISTA-Items auf die Werte der Ländervergleichsstudie 2012 fixiert.

Die Befunde für die Mathematik-Globalskala sind in Tab. 3 wiedergegeben. Die Kompetenzzuwächse variierten zwischen Booklets (Bookleteffekte in Tab. 3), wobei die Effekte aber eher gering ausfielen. Die relativ schwache Korrelation zwischen Bookleteffekten und der mittleren Position der Itemcluster in den Booklets

Tab. 2 DIF-Statistiken zur Übereinstimmung der Itemparameter der Ländervergleichsstudie 2012 mit den Itemparametern des zweiten Messzeitpunkts des PISA-Längsschnitts 2012/2013 nach Kompetenzdomänen und Schulformen: Korrelation der Itemparameter, mittlerer absoluter DIF (Median in Klammern) und Anzahl Items nach DIF-Kategorie (ETS-Klassifikation)

	Gesamtgruppe		Nichtgymnasiale Schulformen		Gymnasien	
Mathematik						
Korrelation	0,99	–	0,97	–	0,96	–
$M_{\|DIF\|}$ ($Mdn_{\|DIF\|}$)	0,21	(0,18)	0,26	(0,23)	0,30	(0,22)
DIF-Kat. (A;B;C)	(25; 23; 1)		(29; 17; 3)		(17; 17; 5)	
Physik Fachwissen						
Korrelation	0,99	–	0,99	–	0,96	–
$M_{\|DIF\|}$ ($Mdn_{\|DIF\|}$)	0,15	(0,14)	0,16	(0,11)	0,37	(0,31)
DIF-Kat. (A;B;C)	(10; 4; 0)		(11; 3; 0)		(8; 3; 3)	
Physik Erkenntnisgewinnung						
Korrelation	0,99	–	0,98	–	0,97	–
$M_{\|DIF\|}$ ($Mdn_{\|DIF\|}$)	0,15	(0,12)	0,21	(0,20)	0,26	(0,17)
DIF-Kat. (A;B;C)	(9; 3; 0)		(7; 5; 0)		(8; 4; 0)	
Chemie Fachwissen						
Korrelation	0,97	–	0,96	–	0,94	–
$M_{\|DIF\|}$ ($Mdn_{\|DIF\|}$)	0,15	(0,11)	0,20	(0,15)	0,30	(0,27)
DIF-Kat. (A;B;C)	(8; 5; 0)		(7; 6; 0)		(4; 9; 0)	
Chemie Erkenntnisgewinnung						
Korrelation	0,99	–	0,99	–	0,97	–
$M_{\|DIF\|}$ ($Mdn_{\|DIF\|}$)	0,22	(0,21)	0,17	(0,09)	0,30	(0,24)
DIF-Kat. (A;B;C)	(7; 6; 0)		(10; 3; 0)		(8; 3; 2)	
Biologie Fachwissen						
Korrelation	0,99	–	0,98	–	0,96	–
$M_{\|DIF\|}$ ($Mdn_{\|DIF\|}$)	0,15	(0,12)	0,17	(0,13)	0,27	(0,22)
DIF-Kat. (A;B;C)	(8; 2; 0)		(7; 3; 0)		(6; 4; 0)	
Biologie Erkenntnisgewinnung						
Korrelation	0,98	–	0,99	–	0,96	–
$M_{\|DIF\|}$ ($Mdn_{\|DIF\|}$)	0,18	(0,17)	0,11	(0,08)	0,28	(0,26)
DIF-Kat. (A;B;C)	(7; 5; 0)		(11; 1; 0)		(5; 7; 0)	

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

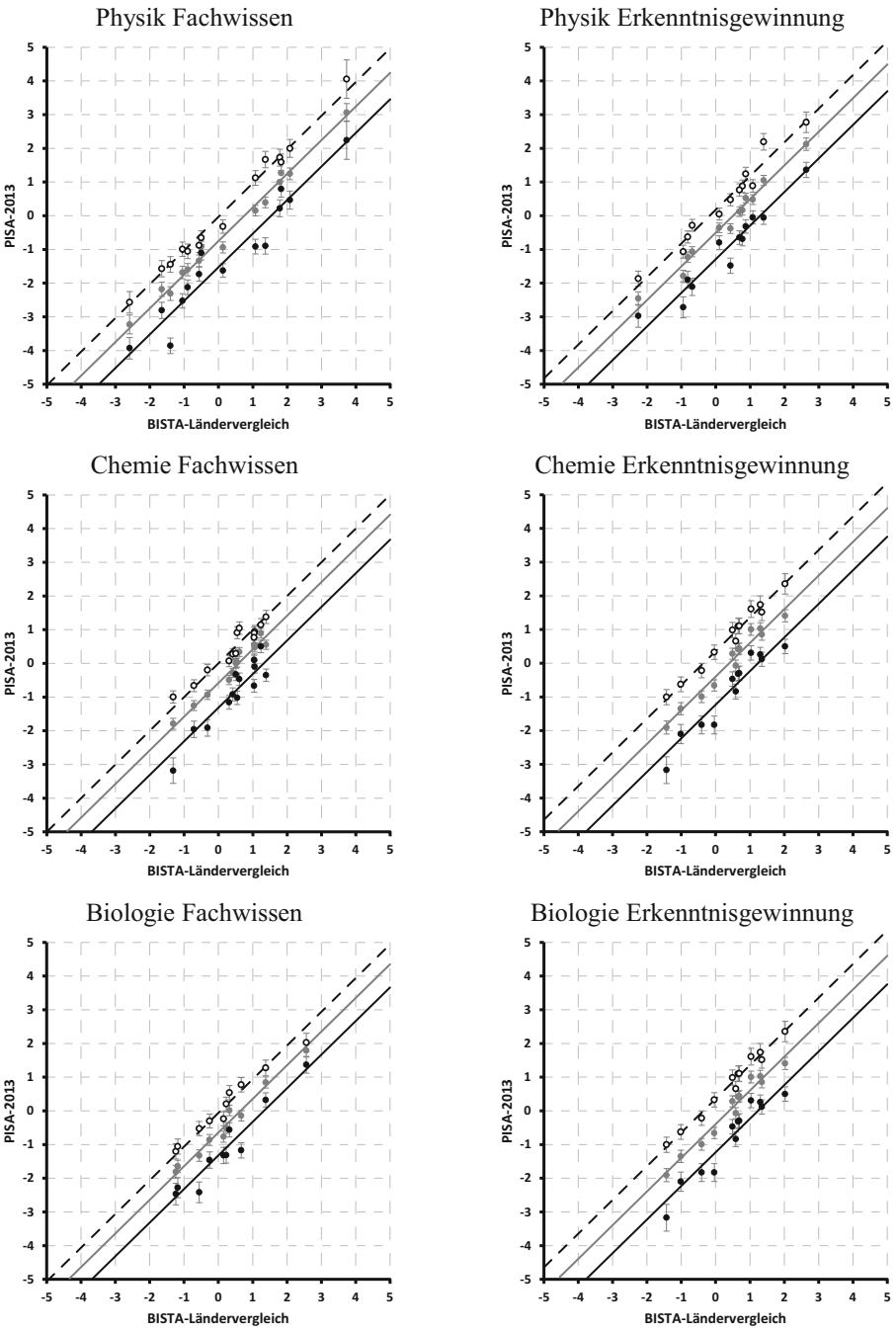

Abb. 2 Übereinstimmung der Itemparameter des Ländervergleichs 2012 (x-Achse) mit den frei geschätzten Itemparametern in der Retesterhebung des PISA-Längsschnitts 2012/2013 (y-Achse) für die naturwissenschaftlichen Bereiche. Fehlerbalken indizieren 95 % Konfidenzintervalle

Tab. 3 Mittlere Leistungszuwächse, um Testkontexteffekte adjustierte Leistungszuwächse (mittels linearer Approximation) sowie Bookleteffekte für den Bereich Mathematik. Angaben für die Gesamtgruppe sowie für die nach Schulformen aufgeteilte Stichprobe

	Gesamtgruppe		Nichtgymnasiale Schulformen		Gymnasien	
	Est	(*SE*)	*Est*	(*SE*)	*Est*	(*SE*)
Mittlerer Zuwachs	0,36	(0,03)***	0,39	(0,03)***	0,32	(0,03)***
Adj. Mittlerer Zuwachs	0,35	(0,03)***	0,37	(0,03)***	0,32	(0,03)***
Bookleteffekte (γ_b)						
Booklet 15 (Pos. = 1,0)	–0,02	(0,07)	0,18	(0,09)**	–0,24	(0,10)**
Booklet 12 (Pos. = 1,0)	–0,01	(0,07)	0,08	(0,09)	–0,06	(0,10)
Booklet 18 (Pos. = 2,0)	0,11	(0,06)*	0,28	(0,08)***	–0,11	(0,08)
Booklet 08 (Pos. = 3,0)	0,14	(0,06)**	0,23	(0,08)***	0,02	(0,08)
Booklet 07 (Pos. = 4,0)	0,24	(0,06)***	0,17	(0,08)**	0,33	(0,08)***
Booklet 13 (Pos. = 4,5)	–0,12	(0,06)**	–0,16	(0,08)**	–0,07	(0,09)
Booklet 09 (Pos. = 5,0)	–0,22	(0,06)***	–0,29	(0,08)***	–0,09	(0,09)
Booklet 16 (Pos. = 5,0)	–0,06	(0,06)	–0,17	(0,09)*	0,05	(0,08)
Booklet 14 (Pos. = 5,0)	–0,04	(0,11)	–0,29	(0,16)*	0,19	(0,14)
Booklet 06 (Pos. = 5,0)	–0,02	(0,06)	–0,03	(0,08)	–0,01	(0,09)
Corr(γ, $Pos.$)	–0,62	–	–0,86	–	0,50	–

Pos. mittlere Position der Itemcluster in Booklets
* $p \leq 0{,}10$, ** $p \leq 0{,}05$, *** $p \leq 0{,}01$

($r = -0{,}62$), lässt darauf schließen, dass die Bookleteffekte neben Positionseffekten auch andere Einflüsse des Testkontexts wiederspiegelten (z. B. Effekt der Domänenabfolge). Die relativ schwache Abhängigkeit der Bookleteffekte von den Itemclusterpositionen, verbunden mit der Tatsache, dass die über Booklets gemittelten Clusterpositionen nahe an der mittleren Position lagen, führte dazu, dass der mittels der linearen Funktion adjustierte Kompetenzzuwachs (0,35, $p < 0{,}01$) nahezu deckungsgleich zum über Booklets gemittelten Zuwachs in der Gesamtstichprobe ausfiel (0,36, $p < 0{,}01$). Wie aus Tab. 3 entnommen werden kann, unterschied sich die Stärke der Bookleteffekte zwischen Schulformen. Diese Effekte waren in nichtgymnasialen Schulformen stärker ausgeprägt und spiegelten offenbar auch Positionseffekte wieder (Korrelation zwischen Bookleteffekten und Clusterpositionen von $r = -0{,}86$ in Tab. 3). Trotz der indizierten Positionseffekte fielen der unadjustierte und der linear adjustierte mittlere Kompetenzanstieg – aufgrund der im Mittel

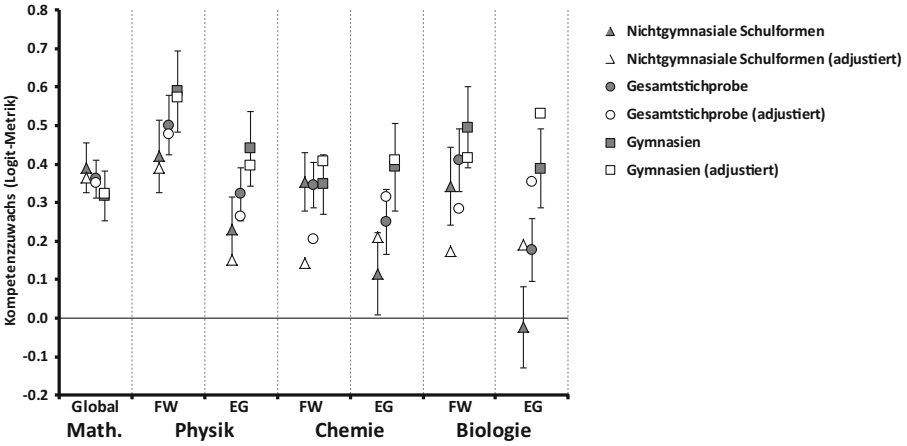

Abb. 3 Mittlere Kompetenzzuwächse nach Kompetenzdomänen in der Gesamtstichprobe, in nichtgymnasialen Schulformen und in Gymnasien. Ergebnisse für nicht-adjustierte Zuwächse (inklusive 95 % Konfidenzintervalle) und linear adjustierte Zuwächse. *FW* Fachwissen, *EG* Erkenntnisgewinnung

balancierten Itemclusterpositonen in dieser Gruppe – nahezu identisch aus. Eine graphische Zusammenfassung dieser Befunde findet sich in Abb. 3.

Die Ergebnisse für die Kompetenzdomänen Physik-Fachwissen und Physik-Erkenntnisgewinnung sind in Tab. 4 wiedergegeben. Für beide Teilbereiche fanden sich in der Gesamtstichprobe statistisch signifikante Bookleteffekte, die sich mit Positionseffekten kompatibel erwiesen (Korrelationen zwischen Bookleteffekten und Clusterpositionen von $r \leq -0{,}90$ in der Gesamtstichprobe in Tab. 4). Die Bookleteffekte waren in beiden Teildomänen erneut in der Gruppe der nichtgymnasialen Schulformen stärker ausgeprägt. Die durch die Bookleteffekte indizierten Positionseffekte wirkten sich jedoch nicht nennenswert auf die Zuwachsschätzungen für den Bereich Physik-Fachwissen aus, da die entsprechenden Itemcluster über alle Booklets hinweg nahe an der mittleren Itemclusterposition vorgelegt wurden. Wie aus Abb. 3 entnommen werden kann, lagen die entsprechenden linear adjustierten Kompetenzzuwächse nahe an den mittleren Kompetenzänderungen und wurden von den entsprechenden 95 % Konfidenzintervallen eingeschlossen.

Für den Kompetenzbereich Physik-Erkenntnisgewinnung zeigte sich, dass sich die lineare Adjustierung der mittleren Kompetenzzuwächse, je nach betrachteter Stichprobe, unterschiedlich stark auswirkte. Da die entsprechenden Items dieser Kompetenzdomäne im Mittel leicht in die vordere Testhälfte verschoben waren, zeichnete sich eine Überschätzung des mittleren Kompetenzanstiegs in der Gesamtgruppe ab (adjustierter mittlerer Zuwachs ist geringer als mittlerer Zuwachs in Tab. 4), jedoch wurde die adjustierte Zuwachsschätzung vom 95 % Konfidenzintervall des mittleren Zuwachses eingeschlossen (Abb. 3). Die Unterschiede zwischen nicht-adjustierten und adjustierten Zuwächsen fielen in der Gruppe der nichtgymnasialen Schulen stärker aus, befanden sich aber immer noch im Konfidenzintervall des nicht-adjustierten Schätzers.

Tab. 4 Mittlere Leistungszuwächse, um Testkontexteffekte adjustierte Leistungszuwächse (mittels linearer Approximation) sowie Bookleteffekte für die Bereiche Physik-Fachwissen und Physik-Erkenntnisgewinnung. Angaben für die Gesamtgruppe, sowie für die nach Schulformen aufgeteilte Stichprobe

	Gesamtgruppe		Nichtgymnasiale Schulformen		Gymnasien	
	Est	(*SE*)	*Est*	(*SE*)	*Est*	(*SE*)
Physik-Fachwissen						
Mittlerer Zuwachs	0,50	(0,04)***	0,42	(0,05)***	0,59	(0,05)***
Adj. Mittlerer Zuwachs	0,48	(0,04)***	0,39	(0,05)***	0,57	(0,06)***
Bookleteffekte (γ_b)						
Booklet 11 (Pos. = 1,0)	0,19	(0,07)***	0,28	(0,08)***	0,14	(0,09)
Booklet 05 (Pos. = 3,0)	0,02	(0,06)	0,06	(0,08)	–0,10	(0,09)
Booklet 12 (Pos. = 3,0)	0,06	(0,06)	0,00	(0,08)	0,16	(0,08)*
Booklet 04 (Pos. = 6,0)	–0,26	(0,07)***	–0,35	(0,09)***	–0,20	(0,10)**
Corr(γ, $Pos.$)	–0,99	–	–0,99	–	–0,80	–
Physik-Erkenntnisgewinnung						
Mittlerer Zuwachs	0,32	(0,04)***	0,23	(0,04)***	0,44	(0,05)***
Adj. Mittlerer Zuwachs	0,26	(0,04)***	0,15	(0,05)***	0,40	(0,05)***
Bookleteffekte (γ_b)						
Booklet 04 (Pos. = 1,0)	0,25	(0,07)***	0,27	(0,08)***	0,19	(0,09)**
Booklet 13 (Pos. = 2,0)	0,12	(0,07)*	0,20	(0,08)***	0,09	(0,10)
Booklet 15 (Pos. = 3,0)	0,05	(0,07)	0,07	(0,08)	0,06	(0,11)
Booklet 10 (Pos. = 4,0)	–0,26	(0,07)***	–0,23	(0,08)***	–0,28	(0,09)***
Booklet 05 (Pos. = 5,0)	–0,15	(0,07)**	–0,30	(0,09)***	–0,06	(0,09)
Corr(γ, $Pos.$)	–0,90	–	–0,97	–	–0,76	–

Pos. mittlere Position der Itemcluster in Booklets
* $p \leq 0{,}10$, ** $p \leq 0{,}05$, *** $p \leq 0{,}01$

Die Bereiche Chemie-Fachwissen und Chemie-Erkenntnisgewinnung waren ebenfalls von Bookleteffekten betroffen, die mit Positionseffekten kompatibel waren (Tab. 5). In beiden Fällen waren die entsprechenden Effekte auf die Zuwachsschätzung in der Gruppe der nichtgymnasialen Schulformen stärker ausgeprägt. Für den Bereich Chemie-Fachwissen zeigte sich, dass die ermittelten Kompetenzzuwächse aufgrund des zum zweiten Messzeitpunkt implementierten Testdesigns womöglich überschätzt werden. Diese Überschätzung ist auf die relativ frühe Positionierung der entsprechenden Items im Test und die deutlich ausgeprägten Positionseffekte zurückzuführen. Wie aus Abb. 3 entnommen werden kann, sind die Kompetenzzuwächse in der Gesamtgruppe und der Gruppe der nichtgymnasialen Schulformen

Tab. 5 Mittlere Leistungszuwächse, um Testkontexteffekte adjustierte Leistungszuwächse (mittels linearer Approximation) sowie Bookleteffekte für die Bereiche Chemie-Fachwissen und Chemie-Erkenntnisgewinnung. Angaben für die Gesamtgruppe sowie für die nach Schulformen aufgeteilte Stichprobe

	Gesamtgruppe		Nichtgymnasiale Schulformen		Gymnasien	
	Est	(*SE*)	*Est*	(*SE*)	*Est*	(*SE*)
Chemie-Fachwissen						
Mittlerer Zuwachs	0,35	(0,03)***	0,35	(0,04)***	0,35	(0,04)***
Adj. Mittlerer Zuwachs	0,20	(0,05)***	0,14	(0,07)**	0,41	(0,05)***
Bookleteffekte (γ_b)						
Booklet 13 (Pos. = 1,0)	0,09	(0,06)	0,08	(0,07)	0,12	(0,08)
Booklet 04 (Pos. = 2,0)	0,08	(0,06)	0,04	(0,07)	0,08	(0,07)
Booklet 11 (Pos. = 2,0)	0,09	(0,06)	0,17	(0,07)***	0,03	(0,07)
Booklet 15 (Pos. = 2,0)	–0,04	(0,06)	0,05	(0,07)	–0,10	(0,08)
Booklet 05 (Pos. = 4,0)	–0,22	(0,06)***	–0,35	(0,07)***	–0,13	(0,07)*
Corr(γ, $Pos.$)	–0,90	–	–0,89	–	–0,79	–
Chemie-Erkenntnisgewinnung						
Mittlerer Zuwachs	0,25	(0,04)***	0,12	(0,05)**	0,39	(0,06)***
Adj. Mittlerer Zuwachs	0,31	(0,05)***	0,21	(0,06)***	0,41	(0,06)***
Bookleteffekte (γ_b)						
Booklet 05 (Pos. = 2,0)	0,26	(0,07)***	0,35	(0,09)***	0,09	(0,10)
Booklet 04 (Pos. = 4,0)	–0,08	(0,07)	–0,08	(0,09)	–0,05	(0,10)
Booklet 10 (Pos. = 5,0)	–0,13	(0,07)*	–0,15	(0,09)	–0,07	(0,10)
Booklet 14 (Pos. = 6,0)	–0,06	(0,08)	–0,12	(0,09)	0,03	(0,10)
Corr(γ, $Pos.$)	–0,85	–	–0,91	–	–0,50	–

Pos. mittlere Position der Itemcluster in Booklets
*$p \leq 0{,}10$, **$p \leq 0{,}05$, *** $p \leq 0{,}01$

besonders stark von Verzerrungen betroffen, da die linear adjustierten Kompetenzanstiege deutlich außerhalb der 95 % Konfidenzintervalle der unadjustierten mittleren Zuwächse liegen. Die Verzerrung scheint in der Gymnasialgruppe aufgrund der schwächer ausgeprägten Positionseffekte (Tab. 5) geringer auszufallen. Tatsächlich befindet sich der linear adjustierte mittlere Zuwachs deutlich innerhalb des 95 % Konfidenzintervalls des nicht-adjustierten Schätzers (Abb. 3).

Die Zuwachsschätzungen für Biologie hingen besonders deutlich vom Testdesign ab. Wie aus Tab. 6 hervorgeht, waren die Bereiche Fachwissen und Erkenntnisgewinnung in allen betrachteten Schülergruppen von Bookleteffekten betroffen, die mit Positionseffekten kompatibel ausfielen. Erneut deutete sich an, dass die entspre-

Tab. 6 Mittlere Leistungszuwächse, um Testkontexteffekte adjustierte Leistungszuwächse (mittels linearer Approximation) sowie Bookleteffekte für die Bereiche Biologie-Fachwissen und Biologie-Erkenntnisgewinnung. Angaben für die Gesamtgruppe sowie für die nach Schulformen aufgeteilte Stichprobe

	Gesamtgruppe		Nichtgymnasiale Schulformen		Gymnasien	
	Est	(*SE*)	*Est*	(*SE*)	*Est*	(*SE*)
Biologie-Fachwissen						
Mittlerer Zuwachs	0,41	(0,04)***	0,34	(0,05)***	0,50	(0,05)***
Adj. Mittlerer Zuwachs	0,28	(0,05)***	0,18	(0,06)***	0,42	(0,06)***
Bookleteffekte (γ_b)						
Booklet 05 (Pos. = 1,0)	0,26	(0,07)***	0,33	(0,08)***	0,12	(0,09)
Booklet 12 (Pos. = 2,0)	0,14	(0,06)**	0,18	(0,09)**	0,13	(0,08)*
Booklet 11 (Pos. = 3,0)	0,00	(0,07)	0,05	(0,08)	0,01	(0,09)
Booklet 04 (Pos. = 5,0)	–0,41	(0,07)***	–0,56	(0,10)***	–0,27	(0,09)***
Corr(γ, $Pos.$)	–0,99	–	–0,98	–	–0,96	–
Biologie-Erkenntnisgewinnung						
Mittlerer Zuwachs	0,18	(0,04)***	–0,02	(0,05)	0,39	(0,05)***
Adj. Mittlerer Zuwachs	0,36	(0,05)***	0,19	(0,07)***	0,53	(0,07)***
Bookleteffekte (γ_b)						
Booklet 04 (Pos. = 3,0)	0,27	(0,07)***	0,28	(0,08)***	0,25	(0,08)***
Booklet 14 (Pos. = 4,0)	0,07	(0,07)	0,16	(0,09)*	0,00	(0,08)
Booklet 05 (Pos. = 6,0)	–0,15	(0,07)**	–0,28	(0,09)***	–0,07	(0,09)
Booklet 10 (Pos. = 6,0)	–0,19	(0,07)***	–0,16	(0,09)*	–0,19	(0,09)**
Corr(γ, $Pos.$)	–0,99	–	–0,98	–	–0,91	–

Pos. mittlere Position der Itemcluster in Booklets
*$p \leq 0{,}10$, **$p \leq 0{,}05$, ***$p \leq 0{,}01$

chenden Effekte in der Gruppe der nichtgymnasialen Schulformen am stärksten ausgeprägt waren. Da die Items zum Biologie-Fachwissen am zweiten Messzeitpunkt zum Anfang des Tests hin verschoben waren, deutete sich an, dass die mittleren Kompetenzzuwächse eine Überschätzung der tatsächlichen Zuwächse darstellen. In allen betrachteten Teilgruppen lagen die linear adjustierten Kompetenzzuwächse unterhalb der nicht-adjustierten Mittelwerte (Tab. 6). Der adjustierte Wert wurde nur in der Gymnasialgruppe vom 95 % Konfidenzintervall des nicht-adjustierten Schätzers umschlossen. Da die Inhalte des Tests zur Biologie-Erkenntnisgewinnung zum zweiten Messzeitpunkt stärker zum Ende des Tests hin verschoben waren, deutete sich an, dass die mittleren Zuwächse die tatsächlichen Kompetenzgewinne unterschätzten. Der Vergleich der nicht-adjustierten mit den adjustierten Zuwächsen indizierte,

dass die Verzerrung in allen betrachteten Gruppen essenziell war, da die adjustierten Zuwächse nicht von den 95 % Konfidenzintervallen der nicht-adjustierten Schätzer eingeschlossen wurden (Abb. 3).

Die hier vorgestellten Ergebnisse legen in der Gesamtschau nahe, dass die mittels der BISTA-Tests erfassten Kompetenzgewinne je nach Kompetenzbereich und betrachteter Schulform unterschiedlich stark vom zum zweiten Messzeitpunkt implementierten Testdesign beeinflusst wurden. Die Kompetenzzuwächse in den Domänen Mathematik (Globalskala) und Physik-Fachwissen erweisen sich als robust, da die mittleren Zuwächse in allen Gruppen nahezu identisch zu den linear adjustierten Zuwächsen waren (Abb. 3). In der Gruppe der Gymnasiastinnen und Gymnasiasten erwiesen sich mit Ausnahme des Bereichs Biologie-Erkenntnisgewinnung alle weiteren Zuwachsschätzungen als vergleichsweise robust, da die linear adjustierten Werte in die Konfidenzbereiche der nicht-adjustierten Werte fielen. Demgegenüber sind die Kompetenzzuwächse in den Domänen Chemie-Fachwissen, Biologie-Fachwissen und Biologie-Erkenntnisgewinnung in der Gesamtgruppe sowie in der Gruppe der nichtgymnasialen Schulformen mit großer Vorsicht zu interpretieren, da die adjustierten Zuwächse deutlich von den nicht-adjustierten Werten abwichen (Abb. 3).

8 Schätzung von Plausible Values

Die Schätzung der Kompetenzzuwächse geschah mittels der PV-Methode, wobei die IRT-Skalierung analog zur Skalierung der Kompetenztests in der Ländervergleichsstudie 2012 getrennt für jede Kompetenzdomäne durchgeführt wurde (Hecht et al. 2013). Aufbauend auf die im ersten Teilabschnitt berichteten Ergebnisse, wurden die Itemparameter zum zweiten Messzeitpunkt des PISA-Längsschnitts 2012/2013 auf die Werte der Ländervergleichsstudie fixiert. Wie bereits erwähnt, wurde von einer Adjustierung für mögliche Testkontexteffekte zum zweiten Messzeitpunkt abgesehen.

Die Skalierung der Kompetenztests geschah unter Rückgriff auf die im Rahmen des Ländervergleichs 2012 erzeugten PVs, die die anfänglichen Kompetenzniveaus der Schülerinnen und Schüler ausweisen. Insgesamt lagen pro Dimension jeweils 15 PVs vor, die bei der Abschätzung der Kompetenzniveaus auf Grundlage der BISTA-Tests im Jahr 2013 berücksichtigt wurden. In die dimensionsweise Skalierung gingen jeweils diejenigen Schülerinnen und Schüler ein, für die PVs auf der entsprechenden Domäne zum ersten Messzeitpunkt vorlagen und/oder die zum zweiten Messzeitpunkt ein Testbooklet bearbeitet hatten, das Items zur entsprechenden Kompetenzdimension beinhaltete. Neben den PVs der BISTA-Dimensionen, die zur ersten Welle erhoben wurden, entsprachen die Hintergrundmodelle dem Hintergrundmodell der PISA-Tests, das auch im Beitrag von Nagy et al. (2017) verwendet wurde.

Bei der Erstellung der Hintergrundmodelle musste dem Umstand Rechnung getragen werden, dass die Ausgangsleistungen in Form von 15 PVs vorlagen, die zum Teil fehlende Werte beinhalteten. Die fehlenden Werte auf den PVs lagen im Erhebungsdesign der Ländervergleichsstudie begründet (vgl. Hecht et al. 2013). Wir

haben das Problem gelöst, indem wir die fehlenden Werte auf den PVs unter Berücksichtigung der Variablen des PISA-Hintergrundmodells mittels einer einfachen Imputation unter Verwendung des Expectation-Maximization-Algorithmus ersetzt haben. Die Imputationen erfolgten getrennt für jeden PV, sodass für jede zu skalierende Kompetenzdimension 15 Hintergrundmodelle vorlagen. Die dimensionsspezifischen Hintergrundmodelle wurden im zweiten Schritt verwendet, um PVs auf den BISTA-Dimensionen zu erzeugen, wobei für jedes der 15 Hintergrundmodelle jeweils 5 PVs erzeugt wurden (d. h. $15 \times 5 = 75$ PVs). Die resultierenden PVs können im Sinne von genesteten multiplen Imputationen verstanden werden, deren Kombination spezifische Routinen voraussetzt (Weirich et al. 2014).

Die auf Grundlage der PVs ermittelten deskriptiven Befunde sind in Tab. 7 dargestellt. Die EAP/PV-Reliabilitäten waren insgesamt befriedigend (gemittelt über 15 Skalierungsläufe), fielen aber in den Gymnasien durchweg höher aus. Auffällig ist weiterhin, dass die Reliabilität der Mathematik Globalskala am geringsten ausfiel, obwohl diese mit der größten Itemzahl erhoben wurde. Dies könnte darauf zurückzuführen sein, dass dieser Kompetenzbereich mittels vergleichsweise heterogener Testinhalte erhoben wurde (d. h. Items zu unterschiedlichen Leitideen; Blum et al. 2010). Erwartungsgemäß lagen zu beiden Wellen große Schulformunterschiede zugunsten der Gymnasiastinnen und Gymnasiasten vor. Mit Ausnahme der Kompetenzdimensionen Chemie Erkenntnisgewinnung und Biologie Fachwissen, für die in der Gruppe der nichtgymnasialen Schulformen keine Leistungsgewinne festgestellt wurden, zeigten sich für alle Kompetenzdimensionen in beiden Teilpopulationen Zuwächse. Die Fachleistungen wiesen in der Gesamtgruppe mittlere Stabilitäten auf, die in den Teilgruppen geringer ausfielen.

Bei der Interpretation der mittels der PVs ermittelten Kompetenzzuwächse muss berücksichtigt werden, dass diese von den Effekten des zum zweiten Messzeitpunkt nicht-balancierten Testdesigns (vgl. Tab. 1) betroffen sind. Dies zeigt sich zum Beispiel darin, dass die in Tab. 7 dargestellten mittleren Zuwächse sehr gut mit den in den Tab. 3 bis 6 berichteten nicht-adjustierten Kompetenzveränderungen übereinstimmen.

9 Zusammenfassung und Diskussion

Gegenstand des vorliegenden Beitrags war die IRT-Skalierung und die Untersuchung der Messeigenschaften der im PISA-Längsschnitt 2012/2013 eingesetzten BISTA-Tests. Die Ergebnisse liefern in der Gesamtschau robuste Hinweise dafür, dass die im Rahmen der Retesterhebung des PISA-Längsschnitts 2012/2013 kalibrierten Itemparameter keine systematischen Verstöße gegen die Messinvarianzannahme aufweisen, die auf den Inhalt der eingesetzten Items zurückzuführen sind. Diese Schlussfolgerung ergab sich aus den Befunden, die eine hohe Übereinstimmung der Itemparameter der Ländervergleichsstudie 2012 mit den frei geschätzten Itemparametern der zweiten Welle des PISA-Längsschnitts erbrachten. Besonders hervorzuheben gilt, dass die Übereinstimmung in allen betrachteten Schulformen und für alle Kompetenzdimensionen durchweg hoch ausfiel. Diese Ergebnisse unterstützen somit die Sichtweise, dass die BISTA-Tests in den betrachteten Schul-

Tab. 7 Deskriptive Statistiken der PVs der Kompetenzdimensionen nach Schulform und in der Gesamtgruppe

	Nichtgymnasiale Schulformen							Gymnasien							Gesamtgruppe					
	M_{T1}	(SD_{T1})	M_{T2}	(SD_{T2})	$r_{T1,T2}$	*MD*	*Rel*	M_{T1}	(SD_{T1})	M_{T2}	(SD_{T2})	$r_{T1,T2}$	*MD*	*Rel*	M_{T1}	(SD_{T1})	M_{T2}	(SD_{T2})	$r_{T1,T2}$	*MD*
MATH	–0,42	(0,91)	–0,03	(1,03)	0,67	0,39	0,70	1,09	(0,97)	1,39	(1,13)	0,71	0,29	0,77	0,27	(1,20)	0,62	(1,28)	0,79	0,35
PH-FW	–0,36	(0,89)	0,05	(1,04)	0,44	0,41	0,84	0,91	(0,90)	1,48	(1,26)	0,51	0,57	0,96	0,21	(1,09)	0,70	(1,35)	0,63	0,48
PH-EG	–0,39	(0,87)	–0,19	(1,03)	0,58	0,20	0,81	0,83	(0,90)	1,33	(1,19)	0,48	0,50	0,88	0,16	(1,07)	0,50	(1,34)	0,68	0,33
CH-FW	–0,35	(0,89)	–0,05	(0,91)	0,55	0,31	0,87	0,90	(0,90)	1,22	(0,87)	0,58	0,32	0,93	0,22	(1,08)	0,53	(1,09)	0,71	0,31
CH-EG	–0,40	(0,91)	–0,35	(1,23)	0,52	0,04	0,86	0,93	(0,93)	1,33	(1,31)	0,41	0,41	0,91	0,20	(1,13)	0,41	(1,52)	0,64	0,21
BI-FW	–0,43	(0,97)	0,00	(1,26)	0,44	0,43	0,92	0,93	(0,95)	1,31	(1,05)	0,30	0,38	0,99	0,20	(1,17)	0,60	(1,34)	0,55	0,41
BI-FW	–0,35	(0,90)	–0,41	(1,26)	0,49	–0,06	0,90	0,81	(0,84)	1,19	(1,16)	0,40	0,38	0,98	0,18	(1,05)	0,32	(1,45)	0,62	0,14

MATH Mathematik Globalskala, *PH* Physik, *CH* Chemie, *BI* Biologie, *FW* Fachwissen, *EG* Erkenntnisgewinnung, *MD* Mittlere Leistungszuwächse, *Rel.* EAP/PV-Reliabilität.

formen und Jahrgangsstufen individuelle Unterschiede in denselben Kompetenzen erfassen. Damit erfüllen die Tests eine wichtige Voraussetzung für die Interpretation von quer- und längsschnittlichen Zusammenhängen der Kompetenzwerte mit leistungsrelevanten Schülermerkmalen (Meredith 1993).

Die Analysen zu der Belastbarkeit der mittels der BISTA-Tests ermittelten Kompetenzzuwächse lieferten jedoch Hinweise dafür, dass aufgrund des zum zweiten Messzeitpunkt verwendeten Testdesigns mit Verzerrungen zu rechnen ist. Hier zeigte sich, dass die verschiedenen Kompetenzbereiche und Schulformen unterschiedlich stark von Verzerrungen betroffen waren. Die Ergebnisse indizierten, dass der Mathematiktest und der Test zur Domäne Physik-Fachwissen belastbare Abschätzungen von absoluten Kompetenzänderungen erlauben. Beiden Tests ist gemeinsam, dass deren Inhalte zu beiden Messzeitpunkten im Mittel nahe an der mittleren Position vorgelegt wurden. Mit mäßigen Verzerrungen bei der Abschätzung absoluter Kompetenzzuwächse ist im Fall der Tests zu den Bereichen Physik-Erkenntnisgewinnung (tendenzielle Überschätzung) und Chemie-Erkenntnisgewinnung (tendenzielle Unterschätzung) zu rechnen, wobei die Gruppe der Schülerinnen und Schüler an nichtgymnasialen Schulformen stärker von den Verzerrungen betroffen waren. Von deutlichen Verzerrungen der mittleren Trendschätzungen kann in den Domänen Chemie-Fachwissen und Biologie (Fachwissen und Erkenntnisgewinnung) ausgegangen werden.

Die im Rahmen der Ländervergleichsstudie kalibrierten Itemparameter wurden zur Erzeugung dimensionsspezifischer PVs verwendet. Die entsprechenden PVs wiesen in den betrachteten Schulformen befriedigende bis gute Reliabilitäten auf, die jedoch für die Mathematik-Globalskala etwas geringer ausfielen. Dieser Befund ist insgesamt erwartungskonform, da dieser Bereich mittels vergleichsweise heterogener Testhinhalte erhoben wurde, die für jeweils unterschiedliche Leitideen des Mathematikunterrichts stehen (Blum et al. 2010). Die im Vergleich zum Ländervergleich 2012 geringeren Reliabilitäten der globalen Mathematikkompetenz sind dabei auch auf die getrennt nach Schulform durchgeführte Skalierung des Tests zurückzuführen. Durch die Gruppenaufteilung fällt die Variabilität der mathematischen Kompetenz geringer aus als in der Gesamtstichprobe, sodass die Reliabilitäten innerhalb der Gruppen zwangsläufig geringer als die Reliabilität in der Gesamtstichprobe sind.

Aus einer psychometrischen Perspektive haben die hier vorgestellten Befunde wichtige Implikationen für die Nutzung der skalierten PVs. Wie zuvor beschrieben, lieferten unsere Analysen Hinweise dafür, dass sich die Bedeutung der mittels der domänenspezifischen BISTA-Tests erfassten individuellen Unterschiede in den Kompetenzbereichen nicht zwischen Schulformen und Messzeitpunkten unterscheidet. Die vorliegenden Daten ermöglichen es somit, korrelative Zusammenhänge (z. B. Kovarianzen und Regressionskoeffizienten) für die Kompetenzbereiche zu untersuchen und deren Höhe zwischen Gruppen und Messzeitpunkten zu vergleichen. Des Weiteren konnten mit Hilfe von Analysen zu den Effekten des zum zweiten Messzeitpunkt eingesetzten Testdesigns diejenigen Kompetenzbereiche identifiziert werden, für die robuste Rückschlüsse über absolute Veränderungen möglich sind (Mathematik und Physik-Fachwissen). Im Fall der anderen Kompetenzbereiche

sollten absolute Veränderungen (d. h. Differenzwerte über die Zeit) aufgrund der ermittelten Verzerrungen mit großer Vorsicht interpretiert werden.

10 Schulformunterschiede in Kompetenzzuwächsen

Obwohl der Fokus des vorliegenden Betrags in erster Linie psychometrisch motiviert war, erbrachten die Analysen auch inhaltlich relevante Befunde zu den Zuwächsen in den mittels der BISTA-Tests erfassten Kompetenzbereichen. Bei aller Unsicherheit der Zuwachsschätzungen, die sich aus dem nicht-balancierten Design ergaben, zeigte sich insgesamt ein Befundmuster, wonach sowohl im Fach Mathematik als auch in den getesteten naturwissenschaftlichen Kompetenzbereichen substanzielle Leistungszuwächse in der 10. Jahrgangsstufe erreicht wurden, und zwar sowohl in gymnasialen als auch in nichtgymnasialen Bildungsgängen. Dieses qualitative Befundmuster erwies sich auch auf Ebene der nicht-adjustierten Kompetenzzuwächse in den nichtgymnasialen Schulformen als relativ robust, da lediglich für den Bereich Biologie-Erkenntnisgewinnung kein statistisch signifikanter Kompetenzanstieg festgestellt wurde. Nach der linearen Adjustierung der Kompetenzzuwächse wurden in der Gesamtstichprobe für alle Bereiche positive und statistisch signifikante Kompetenzgewinne festgestellt (vgl. Tab. 3 bis 6).

Die auf Grundlage der nicht-adjustierten Kompetenzzuwächse ausgewiesenen Schulformunterschiede indizierten, dass die Lerngewinne im Bereich Mathematik vergleichbar ausfielen. In den meisten naturwissenschaftlichen Kompetenzbereichen lagen höhere Lerngewinne an Gymnasien vor, wobei die Höhe der Unterschiede vergleichsweise stark zwischen den Kompetenzbereichen streute. Nach der gruppenweisen Adjustierung der Kompetenzzuwächse reduzierte sich die Variabilität der Schulformunterschiede in den Zuwächsen in den naturwissenschaftlichen Bereichen (Abb. 3). Die entsprechenden Ergebnisse liefern somit Hinweise dafür, dass die Kompetenzzuwächse in den naturwissenschaftlichen Domänen vergleichsweise homogen zugunsten der Gymnasien ausfielen, wobei auch für die nichtgymnasialen Schulformen in allen erfassten Kompetenzbereichen durchgehend Kompetenzzuwächse ermittelt wurden.

Bei der Interpretation der Befunde gilt es jedoch zu berücksichtigen, dass die Ergebnisse auf einfachen linearen Approximationen beruhen und somit nur tentative Richtwerte liefern. Ebenso gilt, dass die hier vorgestellten Abschätzungen der Kompetenzzuwächse lediglich die in der aktuellen Studie vorliegenden Abweichungen von einem hinsichtlich Positionen balancierten Testdesign berücksichtigen. Nicht berücksichtigt wurden längsschnittliche Unterschiede in der Höhe von Positionseffekten und andere Formen von Testkontexteffekten (z. B. Effekte der Domänenabfolge; Harris 1991). Insofern die für die PISA-Tests ermittelten Befunde (Heine et al. 2017) auch für die BISTA-Tests gelten, muss davon ausgegangen werden, dass die Stärke der Positionseffekte in den BISTA-Tests in der Retesterhebung ansteigt und dass dieser Anstieg in leistungsschwächeren Gruppen besonders stark akzentuiert ist. Vor dem Hintergrund dieser Befundlage liegt es nahe davon auszugehen, dass die hier vorgenommene Adjustierung der Kompetenzzuwächse konservativ ausfällt.

Bei aller Unsicherheit der Zuwachsschätzungen ergaben sich über alle getesteten Kompetenzbereiche hinweg Hinweise für nennenswerte Kompetenzgewinne von der Klassenstufe 9 zu 10. Diese Ergebnisse weichen deutlich von denen ab, die mit Hilfe von entsprechenden Instrumenten aus PISA 2012 gefunden wurden (vgl. Nagy et al. 2017). Wir haben im Einleitungsteil argumentiert, dass sich die Tests aus den IQB-Ländervergleichen an den Bildungsstandards orientieren und daher in den Ländern einen vergleichsweise engen Bezug zu den Lehrplänen und vermutlich auch zu den realisierten Lerngelegenheiten im Unterricht haben. Die berichteten Zuwächse stützen diese Argumentation und machen deutlich, dass Tests, die explizit auf nationalen Standards basieren, sensibler für entsprechende Lerngelegenheiten sind und bei Fragen der Effektivität deutscher Schulen zum Einsatz kommen sollten.

11 Implikationen für die Gestaltung von Schulleistungsstudien im Längsschnitt

Die im vorliegenden Beitrag ermittelten Ergebnisse haben wichtige Implikationen für die Umsetzung von Testdesigns in längsschnittlich angelegten Schulleistungsstudien. Insbesondere belegen die Auswertungen die Bedeutung gut konstruierter Testdesigns für die Bestimmung robuster Trendschätzungen. Im optimalen Fall sollten die Testdesigns derart konfektioniert sein, dass sie die Adjustierung von Testkontexteffekten und gegebenenfalls deren Identifikation zu jedem Messzeitpunkt ermöglichen. Optimale Designs sollten zudem zu jedem Messzeitpunkt mehrere Sätze von Ankeritems umfassen, die kaum von Störeffekten beeinflusst sind und somit die längsschnittliche Verlinkung der Kompetenzskalen gewährleisten. Eine Möglichkeit, diese Vorgabe in rotierten Matrix-Designs (Frey et al. 2009) umzusetzen, besteht darin, die längsschnittliche Verlinkung der Kompetenzdimensionen über die an den ersten Itemclusterpositionen vorgelegten Items vorzunehmen (vgl. Nagy et al. 2017).

Von diesem Design sind alternative Testdesigns zu unterscheiden, die auf die Konstanz der Testkonfiguration setzen und dabei nur eine einzige Testform verwenden. Ein solches Design schließt eine Zu- oder Abnahme von Testkontexteffekten (z. B. Positionseffekten), die sich aufgrund der geänderten Testkonfiguration ergeben, aus. Im Vergleich zum zuvor skizzierten Design ermöglicht es jedoch nicht die Adjustierung von Kompetenzzuwächsen um Veränderungen der Reaktionen der getesteten Personen auf die vorgelegte Testkonfiguration (vgl. Nagy et al. 2017). Im schwächsten Fall umfasst das Testdesign lediglich jeweils eine messzeitpunktspezifische Testform, die mittels Ankeritems über die Messzeitpunkte verlinkt werden. Dieses Design ist typisch für den Großteil der längsschnittlichen Schulleistungsstudien und ermöglicht weder die direkte Kontrolle von Testkontexteffekten, die sich aus der Änderung der Itempositionierung ergeben, noch die Kontrolle von Veränderungen von individuellen Reaktionen auf den Testkontext.

Das im PISA-Längsschnitt 2012/2013 verwendete Testdesign für die BISTA-Tests kann zwischen den beiden letztgenannten Designoptionen lokalisiert werden, da es aufgrund der implementierten Testheftrotation die Abschätzung von Effekten der Veränderungen der Itempositionen ermöglicht und somit eine näherungswei-

se Adjustierung von Kompetenzzuwächsen erlaubt, die in den für die empirische Bildungsforschung typischen Testdesigns nicht möglich ist. Insgesamt legen die hier vorgestellten Befunde nahe (vgl. auch Nagy et al. 2017), dass die Erfassung von Kompetenzzuwächsen mit großen Herausforderungen bezüglich des Testdesigns verbunden ist, die es in zukünftigen Längsschnittstudien noch stärker zu berücksichtigen gilt.

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.

Blum, W., Drücke-Noe, C., Hartung, R., & Köller, O. (2010). *Bildungsstandards Mathematik: konkret* (4. Aufl.). Berlin: Cornelsen.

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

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

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.

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

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

Hecht, M., Roppelt, A., & Siegle, T. (2013). Testdesign und Auswertung des Ländervergleichs. In H. A. Pant, P. Stanat, U. Schroeders, A. Roppelt, T. Siegle & C. Pöhlmann (Hrsg.), *IQB-Ländervergleich 2012. Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I* (S. 391–402). Münster: Waxmann.

Heine, J.-H. et al. (2017). Empirische Grundlage und Stichprobenausfall im PISA-Längsschnitt-2012–2013. *Zeitschrift für Erziehungswissenschaft*. doi:10.1007/s11618-017-0756-0.

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Hillsdale NJ: Erlbaum.

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.

Klieme, E., Avenarius, H., Blum, W., Döbrich, P., Gruber, H., Prenzel, M., Reiss, K., Riquarts, K., Rost, J., Tenorth, H.-E., & Vollmer, H. (2003). *Zur Entwicklung nationaler Bildungsstandards – Eine Expertise*. Frankfrut a. M.: Deutsches Institut für Internationale Pädagogische Forschung.

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.

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.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*, 525–543.

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.

Muthén, L. K., & Muthén, B. O. (2012). *Mplus user's guide* (7. Aufl.). Los Angeles CA: 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.

Nagy, G., Heine, J. H., & Köller, O. (2017). IRT-Skalierung der PISA-Tests im PISA-Längsschnitt 2012/2013: Auswirkungen von Testkontexteffekten auf die Zuwachsschätzung. *Zeitschrift für Erziehungswissenschaft*. doi:10.1007/s11618-017-0749-z.

Organization for Economic Cooperation and Development (OECD) (2014). *PISA 2012 technical report*. Paris: OECD Publishing.

Pant, H. A., Stanat, P., Schroeders, U., Roppelt, A., Siegle, T., & Pöhlmann, C. (2013). *IQB-Ländervergleich 2012. Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I*. Münster: Waxmann.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Kopenhagen: Paedagogiske Institut.

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.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.

Wagner, H., Schöps, K., Hahn, I., Pietsch, M., & Köller, O. (2014). Konzeptionelle Äquivalenz von Kompetenzmessungen in den Naturwissenschaften zwischen NEPS, IQB-Ländervergleich und PISA. *Unterrichtswissenschaft*, *42*, 301–320.

Weirich, S., Haag, N., Hecht, M., Böhme, K., Siegle, T., & Lüdtke, O. (2014). Nested multiple imputation in large-scale assessments. *Large-scale Assessments in Education*, *2*, 1–18.

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation*, *31*, 114–128.

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). *ACERConQuest, Version 2.0: generalized item response modelling software*. Camberwell VIC: Australian council for Educational Research.