Robitzsch, Alexander; Lüdtke, Oliver; Köller, Olaf; Kröhne, Ulf; Goldhammer, Frank; Heine, Jörg-Henrik

Herausforderungen bei der Schätzung von Trends in Schulleistungsstudien. Eine Skalierung der deutschen PISA-Daten

formal und inhaltlich überarbeitete Version der Originalveröffentlichung in:
formally and content revised edition of the original source in:
Diagnostica 63 (2017) 2, S. 148-165

Robitzsch, A., Lüdtke, O., Köller, O., Kröhne, U., Goldhammer, F. & Heine, J.-H. (2017). Herausforderungen bei der Schätzung von Trends in Schulleistungsstudien. *Diagnostica, 63*, 148-165.
https://doi.org/10.1026/0012-1924/a000177

Diagnostica

Herausforderungen bei der Schätzung von Trends in Schulleistungsstudien: Eine Skalierung der deutschen PISA-Daten Challenges in estimations of trends in large-scale assessments: A calibration of the German PISA data

--Manuskript-Entwurf--

Manuskriptnummer:	DIA-D-16-00037
Vollständiger Titel:	Herausforderungen bei der Schätzung von Trends in Schulleistungsstudien: Eine Skalierung der deutschen PISA-Daten Challenges in estimations of trends in large-scale assessments: A calibration of the German PISA data
Artikeltyp:	Originalarbeit
Schlüsselwörter:	Schulleistungsmessung; Large-Scale-Assessment; Test-Moduseffekte; Itemkalibrierung Educational measurement; Large scale assessment; Mode effects; Item calibration
Korrespond. Autor:	Alexander Robitzsch IPN Kiel Kiel, GERMANY
Korrespondierender Autor, Zweitinformationen:	
Korrespondierender Autor, Institution:	IPN Kiel
Korrespondierender Autor, zweite Institution:	
Erstautor:	Alexander Robitzsch
Erstautor, Zweitinformationen:	
Reihenfolge der Autoren:	Alexander Robitzsch
	Oliver Lüdtke, Prof. Dr.
	Olaf Köller, Prof. Dr.
	Ulf Kröhne, Dr.
	Frank Goldhammer, Prof. Dr.
	Jörg-Henrik Heine, Dipl.-Psych.
Reihenfolge 'Zweite Informationen' von Autoren:	
Zusammenfassung:	Internationale Schulleistungsstudien wie das Programme for International Student Assessment (PISA) dienen den teilnehmenden Ländern zur Feststellung der Leistungsfähigkeit ihrer Schulsysteme. In PISA wird die Zielpopulation (15jährige Schülerinnen und Schüler) alle drei Jahre getestet. Von besonderer Bedeutung sind dabei die Trendinformationen, die für die Zielpopulation ausweisen, ob sich ihre Leistungen gegenüber denen aus früheren Erhebungen verändert haben. Um solche Trends valide interpretieren zu können, sollten die PISA-Erhebungen unter möglichst vergleichbaren Bedingungen durchgeführt und die verwendeten statistischen Verfahren vergleichbar bleiben. In PISA 2015 wurde erstmalig computerbasiert getestet, zuvor mittels Papier-und-Bleistift-Tests. Es wurde das Skalierungsmodell verändert und in den Naturwissenschaften wurden neue Aufgabenformate eingesetzt. Im vorliegenden Beitrag gehen wir anhand der nationalen PISA-Stichproben von 2000 bis 2015 der Frage nach, inwiefern der Wechsel des Testmodus und der Wechsel des Skalierungsmodells die Interpretation der Trendschätzungen beeinflussen. Die Analysen belegen, dass die Veränderung von Papier-und-Bleistift-Tests auf Computertestung die Trendschätzung für Deutschland verzerrt haben könnte. Keywords: Schulleistungsmessung, Large-Scale-Assessment, Test-Moduseffekte,

	Itemkalibrierung International large-scale assessments, for instance, the Programme for International Student Assessment (PISA) are conducted to provide information on the effectiveness of educational systems. In PISA, the target population of 15-year-old students is assessed every three years. Trends show whether competencies have changed for the target population between PISA cycles. To ensure valid trend information it is necessary to keep the test conditions and statistical methods in all PISA cycles as constant as possible. In PISA 2015, however, several changes were established; the test model changed from paper-pencil to computer tests, scaling methods were changed and new types of tasks were used in science. In this paper, we investigate the effects of these changes on trend estimation in PISA using German data from all PISA cycles (2000 to 2015). Findings suggest that the change from paper-pencil to computer tests could have biased the trend estimation.

Zusammenfassung: Internationale Schulleistungsstudien wie das Programme for International Student Assessment (PISA) dienen den teilnehmenden Ländern zur Feststellung der Leistungsfähigkeit ihrer Schulsysteme. In PISA wird die Zielpopulation (15-jährige Schülerinnen und Schüler) alle 3 Jahre getestet. Von besonderer Bedeutung sind dabei die Trendinformationen, die für die Zielpopulation ausweisen, ob sich ihre Leistungen gegenüber denen aus früheren Erhebungen verändert haben. Um solche Trends valide interpretieren zu können, sollten die PISA-Erhebungen unter möglichst vergleichbaren Bedingungen durchgeführt und die verwendeten statistischen Verfahren vergleichbar bleiben. In PISA 2015 wurde erstmalig computerbasiert getestet; zuvor mittels Papier-und-Bleistift-Tests. Es wurde das Skalierungsmodell verändert und in den Naturwissenschaften wurden neue Aufgabenformate eingesetzt. Im vorliegenden Beitrag gehen wir anhand der nationalen PISA-Stichproben von 2000 bis 2015 der Frage nach, inwiefern der Wechsel des Testmodus und der Wechsel des Skalierungsmodells die Interpretation der Trendschätzungen beeinflussen. Die Analysen belegen, dass die Veränderung von Papier-und-Bleistift-Tests auf Computertestung die Trendschätzung für Deutschland verzerrt haben könnte.

Schlüsselwörter: Schulleistungsmessung; Large-Scale-Assessment; Test-Moduseffekte; Itemkalibrierung

Abstract: International large-scale assessments, for instance, the Programme for International Student Assessment (PISA) are conducted to provide information on the effectiveness of educational systems. In PISA, the target population of 15-year-old students is assessed every 3 years. Trends show whether competencies have changed for the target population between PISA cycles. To ensure valid trend information it is necessary to keep the test conditions and statistical methods in all PISA cycles as constant as possible. In PISA 2015, however, several changes were established; the test model changed from paper-pencil to computer tests, scaling methods were changed and new types of tasks were used in science. In this paper, we

investigate the effects of these changes on trend estimation in PISA using German data from all PISA cycles (2000 to 2015). Findings suggest that the change from paper-pencil to computer tests could have biased the trend estimation.

Keywords: Educational measurement; Large scale assessment; Mode effects; Item calibration

Kurztitel: Trendschätzungen in PISA

Einleitung

Das Programme for International Student Assessment (PISA; Reiss, Sälzer, Schiepe-Tiska & Köller, 2016) erfasst seit 2000 im dreijährigen Zyklus die Kompetenzen von 15-jährigen Schülerinnen und Schülern in den Domänen Mathematik, Leseverstehen und Naturwissenschaften. Basierend auf dem angelsächsischen Grundbildungskonzept (Literacy; OECD, 2016) wird davon ausgegangen, dass hinreichende Kompetenzen in den drei getesteten Bereichen notwendige Voraussetzungen für berufliche und gesellschaftliche Teilhabe sind (vgl. OECD, 2016) und nationale Bildungssysteme Lerngelegenheiten anbieten sollten, in denen Kinder und Jugendliche die entsprechenden Kompetenzen aufbauen können. In diesem Sinne soll PISA auch ein Instrument zur Feststellung der Leistungsfähigkeit von Bildungssystemen sein. Die Leistungsfähigkeit eines Teilnehmerstaats kann durch den sozialen Vergleich, z. B. mit dem Mittelwert aller Organisation for Economic Co-operation and Development (OECD)-Staaten, vorgenommen werden. Durch den PISA-Zyklus mit wiederkehrenden Testungen (alle drei Jahre) ist allerdings der Trend in den Leistungen der 15-jährigen für jeden Teilnehmerstaat informativer. Beispielsweise hatte das enttäuschende Abschneiden deutscher Schülerinnen und Schüler in PISA 2000 (vgl. Baumert et al., 2001) weitreichende Maßnahmen zur Verbesserung der Bildungsqualität zur Folge (Klieme, Jude, Baumert & Prenzel, 2010) und in den nachfolgenden PISA-Erhebungen stiegen die Leistungen deutscher 15-jähriger im Lesen, in der Mathematik und in den Naturwissenschaften bislang kontinuierlich an. Lagen die Leistungen im Jahr 2000 noch in allen drei Domänen signifikant unter dem OECD-Mittelwert, so hatte sich das Bild in 2012 gewandelt und deutsche Schülerinnen und Schüler lagen durchgängig signifikant über dem OECD-Mittelwert (vgl. Prenzel, Sälzer, Klieme & Köller, 2013), was zumindest bildungspolitisch als Ausdruck erfolgreicher Reformen im Bildungssystem interpretiert wurde (vgl. hierzu Ehmke, Klieme & Stanat, 2013).

In PISA 2015 bricht dieser Trend in Mathematik und den Naturwissenschaften ein. In den Naturwissenschaften zeigt sich mit minus 15 Punkten (2012: 524 Punkte; 2015: 509 Punkte) ein dramatischer Abfall der Leistungen innerhalb von drei Jahren, der erklärungsbedürftig ist, folgt man der Argumentation, dass Mittelwertveränderungen auf Länderebene über relativ kurze Zeiträume typischerweise eher klein ausfallen, wenn die Testbedingungen konstant gehalten werden (Beaton, 1988; Mazzeo & von Davier, 2008). Die Deutlichkeit des Rückgangs von 15 Punkten kann auch daran festgemacht werden, dass sich der Leistungszuwachs innerhalb eines Schuljahres typsicherweise in einer Größenordnung von 25 bis 30 Punkten auf der PISA-Skala bewegt (Prenzel et al., 2013). Auch im internationalen Trend (über alle teilnehmenden OECD-Staaten) besteht ein Leistungsabfall von 8 Punkten für die Naturwissenschaften. Es stellt sich die Frage, ob diese relativ hohen Leistungseinbußen Ausdruck eines tatsächlichen Rückgangs der naturwissenschaftlichen Kompetenzen sind oder ob sie nicht (zumindest teilweise) auf die vielen Veränderungen zurückgeführt werden können, die in der PISA 2015 Studie gegenüber den fünf vorherigen Erhebungszyklen (2000, 2003, 2006, 2009, 2012) umgesetzt wurden. In der Auswertung und Durchführung von PISA 2015 wurden vor allem zwei substanzielle Veränderungen vorgenommen. Erstens wurde zur Skalierung der Daten anstatt eines 1PL-Modells (Rasch, 1960), in dem nur Schwierigkeitsparameter für die Items geschätzt werden, ein 2PL-Modell (Birnbaum, 1968) verwendet, das zusätzlich für jedes Item einen Diskriminationsparameter schätzt (Rost, 2004). Zweitens wurde in PISA 2015 der Umstieg von Paper-Based Assessment (PBA) auf Computer-Based Assessment (CBA) vollzogen.

Im vorliegenden Beitrag gehen wir der Frage nach, ob die Mittelwertdifferenzen zwischen PISA 2012 und PISA 2015 womöglich keine Abnahme in der Leistungsfähigkeit des deutschen Schulsystems widerspiegeln, sondern dem Wechsel des Testmodus – von PBA auf CBA – und dem Wechsel des Skalierungsmodells (1PL vs. 2PL) geschuldet sein könnten. Dazu nehmen wir Reanalysen der nationalen PISA-Daten aller sechs Erhebungen vor und

berücksichtigen zur Abschätzung von Effekten des Wechsels im Testmodus auch Daten des 2014 durchgeführten Feldtests für PISA 2015, in dem entsprechende Effekte (PBA vs. CBA) experimentell überprüft wurden.

Das Programme for International Student Assessment (PISA)

PISA ist ein Programm der OECD, das den Mitglieds- und Partnerstaaten in regelmäßigen Abständen – alle drei Jahre – indikatorengestützt Informationen über die Leistungsfähigkeit ihrer Bildungssysteme liefern soll. Die Zielpopulation stellen 15-jährige dar. Für diese Altersgruppe ist der Schulbesuch noch obligatorisch, sodass man in den Testungen die Altersgruppe in ihrer gesamten Heterogenität abbilden kann. Die wichtigsten Indikatoren in PISA sind die Leistungen der Schülerinnen und Schüler in den Bereichen Mathematik, Lesen in der Verkehrssprache und Naturwissenschaften. Für alle drei Bereiche gilt, dass sie zu erheblichen Anteilen schulisch vermittelt sind und curriculare Ziele international ähnlich sind. Alle drei Bereiche basieren auf dem angelsächsischen, funktionalen *Literacy*-Konzept, das im Deutschen am besten unter dem Begriff der Grundbildung zu fassen ist. Im Kontext von PISA umfasst funktional im Wesentlichen zwei Aspekte, nämlich die *Anwendbarkeit* für die jetzige und die spätere, nach-schulische Teilhabe an einer Kultur, sowie die *Anschlussfähigkeit* im Sinne kontinuierlichen Weiterlernens über die Lebensspanne.

Die drei untersuchten Domänen wurden von Beginn an unterschiedlich gewichtet, die Hauptdomäne umfasste dabei rund die Hälfte der Aufgaben; die beiden Nebendomänen teilten sich die zweite Hälfte. In den Jahren 2000 und 2009 war Lesen die Hauptdomäne, in den Jahren 2003 und 2012 Mathematik, 2006 die Naturwissenschaften. Beginnend mit PISA 2015 soll die Gewichtung der drei Domänen einander zunehmend angeglichen werden, wobei die Unterteilung in Haupt- und Nebendomäne erhalten bleibt. In PISA 2015 stellen nach PISA 2006 zum zweiten Mal die Naturwissenschaften die Hauptdomäne dar.

Generell verwendet PISA sogenannte Linkitems, die zu mehreren Erhebungen eingesetzt werden. Durch diese gemeinsamen Items zu den unterschiedlichen Zeitpunkten soll

sichergestellt werden, dass eine gemeinsame Metrik in den jeweiligen Domänen über die Zeit etabliert werden kann und Leistungen von 15-jährigen über die verschiedenen PISA-Erhebungen miteinander verglichen werden können (Kolen & Brennan, 2014). So soll überprüft werden, ob sich die Leistungsfähigkeit von Bildungssystemen verbessert oder verschlechtert. Im Folgenden sollen methodischen Herausforderungen diskutiert werden, die mit der Schätzung und Interpretation dieser Trends verbunden sind.

Bestimmung von Trends in PISA

In der Literatur zu Trendanalysen wird der *originale Trend* von einem *marginalen Trend* unterschieden (Gebhardt & Adams, 2007; siehe auch Carstensen, Prenzel & Baumert, 2008; Sachse, Roppelt & Haag, 2016). In der originalen Trendschätzung werden die Änderungen in der mittleren Leistung eines Teilnehmerstaates auf der internationalen Metrik betrachtet. Dazu werden die in jeder PISA-Erhebung auf Basis aller Teilnehmerstaaten gewonnenen internationalen Itemparameter durch ein Linking (Gebhardt & Adams, 2007) oder eine konkurrente Skalierung (ETS, 2015) auf die gemeinsame PISA-Metrik gebracht. Als Referenz für die originale Trendschätzung wird PISA 2000 für Lesen, PISA 2003 für Mathematik oder PISA 2006 für Naturwissenschaften verwendet, in der die Fähigkeitsverteilung aller teilnehmenden Schülerinnen und Schüler jeweils auf einen Mittelwert von 500 und eine Standardabweichung von 100 fixiert wurde.

In der marginalen Trendschätzung wird dagegen die Trendschätzung für einen Teilnehmerstaat unabhängig von den internationalen Itemparametern vorgenommen. Für diese Trendschätzung werden zunächst nationale Itemparameter aus den einzelnen PISA-Erhebungen des Teilnehmerstaates bestimmt, die wieder mithilfe eines Linkings auf eine gemeinsame nationale Metrik gebracht werden, die unabhängig von der internationalen Metrik des originalen Trends ist. Praktisch bedeutet dies, dass der nationale Trend nur auf Basis der über die Erhebungen gemeinsam vorgelegten Linkitems bestimmt wird und (im Gegensatz zur originalen Trendschätzung) nur zu einer Erhebung eingesetzte Items (Nicht-

Linkitems) unberücksichtigt bleiben. Als Referenz für die marginale Trendschätzung in einer Domäne werden üblicherweise der Mittelwert und die Standardabweichung der ersten Erhebung des Teilnehmerstaates verwendet (z. B. für Leseverstehen in Deutschland in PISA 2000: $M = 484$, $SD = 111$).

In den offiziellen Publikationen zur PISA-Studie werden meistens die originalen Trendschätzungen betrachtet, indem einfach die Mittelwerte für einen Teilnehmerstaat aus den querschnittlichen Berichterstattungen subtrahiert werden. Eine Reihe von Studien konnte allerdings zeigen, dass die originalen von den marginalen Trendschätzungen erheblich abweichen können (Carstensen et al., 2008; Gebhardt & Adams, 2007; Robitzsch, 2016, S. 196). Diese Unterschiede zwischen originalem und marginalem Trend lassen sich auf querschnittliches differentielles Itemfunktionieren (Country-DIF) und die Anlage des PISA-Designs zurückführen. Country-DIF führt dazu, dass nationale von den internationalen Itemparametern abweichen und somit der Unterschied zwischen dem Mittelwert eines Teilnahmestaates und dem internationalen Mittelwert davon abhängt, welche Items für den Mittelwertvergleich herangezogen werden. Dies stellt aufgrund des Wechsels von Haupt- und Nebendomänen über die PISA-Erhebungen eine besondere Herausforderung für die Trendschätzung dar. Wenn eine Kompetenzdomäne bei einer Erhebung Hauptdomäne ist, dann wird neben den Linkitems auch eine große Menge von Nicht-Linkitems für diese Domäne administriert. Ist der Country-DIF im Mittel bei den Linkitems bei einem Teilnehmerstaat größer oder kleiner ausgeprägt als bei den Nicht-Linkitems, dann kann sich der Mittelwert dieses Staates auf Basis der Linkitems von dem Mittelwert auf der internationalen Metrik, der auf Basis aller Items bestimmt wird (Linkitems und Nicht-Linkitems), unterscheiden. Dies hat zur Konsequenz, dass die originale von der marginalen Trendschätzung abweicht, da letztere nur die Linkitems berücksichtigt.

Bei der Schätzung von Trends gilt es vor allem zwei Quellen der Unsicherheit zu berücksichtigen. Erstens ist jede Mittelwertschätzung mit einem Schätzfehler im Hinblick auf

eine Population verbunden, da nur eine Stichprobe von Schülerinnen und Schülern eines Teilnehmerstaates zu einem Erhebungszeitpunkt vorliegt. Zweitens kann der Mittelwert eines Teilnehmerstaates gleichermaßen durch Auswahl bestimmter Items zu einem Erhebungszeitpunkt größer oder kleiner ausfallen. Für Trendschätzungen stellt also neben der Auswahl von Schülerinnen und Schülern die Auswahl von Items eine zusätzliche Quelle von Unsicherheit dar, die bei der statistischen Absicherung des Trends miteinfließen sollte (Husek & Sirotnik, 1967; Monseur & Berezner, 2007). PISA weist für den originalen Trend die durch Items bedingte Variabilität in der Trendschätzung einen so genannten Linkfehler aus (OECD, 2009). Dieser Linkfehler beschreibt ausschließlich die Variabilität der internationalen Parameter der Linkitems über mehrere PISA-Erhebungen. Country-DIF geht explizit nicht in die Berechnung dieses Linkfehlers ein. Monseur und Berezner (2007) konnten allerdings mit Reanalysen von PISA-Daten zeigen, dass die durch die Itemauswahl bedingte Unsicherheit in der originalen Trendschätzung für einen Teilnehmerstaat erheblich größer als der international berichtete Linkfehler ausfallen würde, wenn Country-DIF auf den Linkitems in die Berechnung einfließen würde (siehe für ein analytisches Vorgehen Robitzsch, 2016, S. 193ff.). Daraus lässt sich folgern, dass für eine präzise Trendschätzung eine relativ große Menge von Linkitems notwendig ist, um die durch Country-DIF verursachte Variabilitätsquelle zu verringern (Mazzeo & von Davier, 2008). Ist dies wie z. B. für die Domäne Leseverstehen in den PISA-Erhebungen von 2000 bis 2009 nicht der Fall, so können marginale Trendschätzungen zu robusteren Abschätzungen von Kompetenzentwicklungen führen als originale Trendschätzungen (Gebhardt & Adams, 2007; Sachse et al., 2016).

Im internationalen Vorgehen wird nur von partieller Invarianz aller Items über alle Staaten hinweg ausgegangen. Dabei werden für einzelne Staaten die Itemparameter nur dann als invariant über alle Staaten angenommen, wenn der Country-DIF nicht zu groß ausfällt (ETS, 2015; Oliveri & von Davier, 2011), ansonsten werden Itemparameter für einen Staat

frei geschätzt und könnten somit von den internationalen Itemparametern abweichen. Der Vergleich eines Teilnehmerstaates mit dem internationalen Referenzwert beruht dann praktisch auf der Menge der als invariant angenommenen Items, da bei der Betrachtung von mittleren Unterschieden zwischen nationalen und internationalen Itemparametern die Items mit zu großen Abweichungen (Country-DIF) aus der Mittelwertbildung entfernt werden (vergleich für Alternativen Fox & Verhagen, 2010; Muthen & Asparouhov, 2014; Robitzsch, 2016, S. 188ff.). Trotz des Zulassens einzelner nationaler Itemparameter bleibt dennoch die Variabilitätsquelle der Auswahl der Linkitems bestehen und ist eine offene Frage, ob mit dem Vorgehen der Skalierung unter partieller Invarianz im Vergleich zu einem Linking unter vollständiger Nicht-Invarianz (für alle Staaten und Items werden nationale Parameter verwendet) effizientere Mittelwertschätzungen und in dessen Folge effizientere Trendschätzungen erhalten werden.[1]

Zusammenfassend verdeutlichen diese Überlegungen, dass die Schätzung und Interpretation von Trends in PISA selbst dann schon ein komplexes Unterfangen darstellt (Mazzeo & von Davier, 2008), wenn die Testdurchführungsbedingungen über die Erhebungen weitgehend vergleichbar bleiben. Dies lässt vermuten, dass bei größeren Änderungen in den Durchführungsbedingungen von noch instabileren Trendschätzungen auszugehen ist, die eine Interpretation von Veränderungen in den Kompetenzen erschweren. In dem vorliegenden Beitrag sollen die robusteren marginalen Trendschätzungen verwendet werden, um den Einfluss von zwei zentralen Veränderungen, die in PISA 2015 vorgenommenen wurden, auf den Trend für 15-jährige in Deutschland abzuschätzen.

[1] Sachse et al. (2016) zeigten in einer Simulationsstudie, dass bei Existenz von Country-DIF marginale Trendschätzungen effizienter als originale Trendschätzungen sein können. Außerdem erwiesen sich originale Trendschätzungen bei Elimination von Items mit Country-DIF für das Linking (was dem Vorgehen in PISA 2015 entspricht) als weniger effizient im Vergleich zu originalen Trendschätzungen, in denen Items mit Country-DIF nicht eliminiert wurden.

Änderungen in PISA 2015

Im Folgenden sollen zwei zentrale Veränderungen diskutiert werden, die in der PISA 2015 Studie gegenüber den fünf vorherigen Erhebungen (2000, 2003, 2006, 2009, 2012) umgesetzt wurden: die Änderung des Skalierungsmodells und der Umstieg von Paper-Based Assessment (PBA) auf Computer-Based Assessment (CBA).

Änderung des Skalierungsmodells

Large-Scale-Assessment-Studien unterscheiden sich in der Wahl des Skalierungsmodells für Leistungsdaten. So hat die PISA-Studie bis 2012 ein 1PL-Modell zur Skalierung des Kompetenztests verwendet. Auch im deutschen Sprachraum wurde das 1PL-Modell in einer Reihe von Studien eingesetzt, z. B. Deutsch-Englisch-Schülerleistungen-International (DESI; Klieme & Beck, 2007), Überprüfen von Bildungsstandards – Ländervergleich (BISTA; Köller, Knigge & Tesch, 2010) oder National Educational Panel Study (NEPS; Pohl & Carstensen, 2012). Dagegen griffen andere Studien auf das 3PL-Modell zurück, das neben einem Schwierigkeitsparameter noch jeweils einen Diskriminationsparameter und einen Rateparameter für jedes Item vorsieht, z. B. Trends in International Mathematics and Science Study (TIMSS; Wendt, Bos, Selter, Köller, Schwippert & Kasper, 2016), Progress in International Reading Literacy Study (PIRLS; Bos, Hornberg, Arnold et al., 2007) oder National Assessment of Educational Progress (NAEP; NCES, 2013). Ein 2PL-Modell, das für jedes Item einen Schwierigkeits- und einen Diskriminationsparameter postuliert, wurde in PIAAC (Programme for the International Assessment of Adult Competencies; Rammstedt, 2013) eingesetzt und wird auch seit 2015 in PISA für die Skalierung der Leistungstests verwendet (Heine, Mang, Borchert et al., 2016).

Für die Wahl eines 2PL- oder 3PL-Modells werden in der psychometrischen Literatur vor allem der häufig bessere Modellfit (d. h. ein besserer Fit der Item-Response Funktion) angeführt (Oliveri & von Davier, 2011, 2014). Insbesondere wenn, wie in Large-Scale-Assessments häufig der Fall, Items mit unterschiedlichen Antwortformaten (z. B. multiple-

choice und offene Antwortformate) vorliegen, weisen Items eine unterschiedliche Reliabilität auf und führen somit zu unterschiedlichen Diskriminationsparametern (Mazzeo & von Davier, 2008). Das 3PL-Modell besitzt den zusätzlichen Vorteil, dass Rateverhalten bei Multiple-Choice-Items modelliert werden kann, was in Large-Scale-Assessments häufig gegenüber dem 2PL-Modell zu einem besseren Modellfit führt (Aitkin & Aitkin, 2011)

Für die Wahl eines 1PL-Modells spricht allerdings, dass jedes Item mit dem gleichen Gewicht in die Bestimmung des Fähigkeitswerts einfließt. Dies steht im Gegensatz zu einer eher datengetriebenen Gewichtung der Items beim 2PL- oder 3PL-Modell, in denen reliablere Items mit einem stärkeren Gewicht berücksichtigt werden. Es kann argumentiert werden, dass die Gleichgewichtung von Items zu einer besseren Abbildung des Testframeworks führen kann als eine Gewichtung von Items auf Basis der Passung des Modells (Brennan, 2001). Folgt man dieser Perspektive, dann ist die Frage nach dem Skalierungsmodell nicht einfach über einen Modellvergleich zu lösen, also keine rein empirische Frage, sondern es ist theoretisch zu klären, wie die einzelnen Items gewichtet werden sollen. Empirisch stellt sich allerdings die Frage, wie stark Befunde sich unterscheiden, wenn ein 1PL- oder 2PL-Modell für die Skalierung von Leistungsdaten verwendet wird. Macaskill (2008) nutzte PISA-Daten von 31 Teilnehmerstaaten aus PISA 2003 und PISA 2006 und verglich Mittelwerte und Trendschätzungen der Staaten sowohl unter dem 1PL- als auch dem 2PL-Modell. Dabei zeigten sich für PISA 2006 für die absolute Differenz der Mittelwerte aus dem 1PL- und dem 2PL-Modell im Mittel kleinere, aber für einige Staaten (insbesondere in Lesen) größere Abweichungen, obwohl die Korrelationen der Ländermittelwerte aus dem 1PL- und dem 2PL-Modell hoch ausfielen (Naturwissenschaften: $M = 2.2$, $Max = 4.7$, $r > .999$; Mathematik: $M = 1.0$, $Max = 3.0$, $r = .999$; Lesen: $M = 2.7$, $Max = 12.9$, $r = .992$). Anhand der Daten aus Macaskill (2008) lässt sich außerdem zeigen dass mit einem Wechsel des Skalierungsmodells von 1PL (PISA 2003) zu 2PL (PISA 2006) keine deutlich anderen Trendschätzungen resultieren würden, als wenn die Trends auf dem 1PL beruhen würden.

Moduswechsel

Gründe für einen Wechsel des Modus von Papier (PBA) auf Computer (CBA) können aus diagnostischer Sicht darin bestehen, innovative Aufgabenformate zu realisieren (Parshall, Harmes, Davey & Pashley, 2010), die Messeffizienz zu erhöhen (van der Linden, 2005) oder zusätzlich zu Ergebnisdaten auch Prozessdaten zu sammeln (Goldhammer, Naumann, Rölke, Stelter & Tóth, im Druck). Bei einem Moduswechsel in Studien mit Längsschnitt auf Populationsebene (z. B. PISA) oder Individualebene (z. B. National Educational Panel Study, NEPS) stellt sich aus psychometrischer Sicht die Herausforderung, die Vergleichbarkeit der Messungen zwischen den Modi sicherzustellen, da ansonsten keine stabile Trendfortsetzung möglich ist (vgl. Mazzeo & von Davier, 2008).

Die zentrale Frage, ob der Wechsel des Modus die psychometrischen Eigenschaften der Messung beeinflusst (Moduseffekt, s. Kröhne & Martens, 2011), wurde im Kontext internationaler Large-Scale-Assessments bereits in PIAAC 2012 untersucht (OECD, 2013a; Yamamoto, 2012). Dies war erforderlich, da Teilnehmerinnen und Teilnehmer mit mangelnder Computervertrautheit die Aufgaben auf Papier lösten und außerdem ein Vergleich mit den Ergebnissen der früheren PBA-Erwachsenenstudien ALL (Adult Literacy and Lifeskills Survey) und IALS (International Adult Literacy Survey, OECD) möglich sein sollte (OECD, 2013b). Auf nationaler Ebene erfolgt im NEPS (Artelt, Weinert & Carstensen, 2013) von Moduseffekt-Link-Studien begleitet der Umstieg von PBA auf CBA seit 2012. Auch in den zentralen Bildungsvergleichsstudien der International Association for the Evaluation of Educational Achievement (IEA) werden gerade CBA-Komponenten ergänzt (ePIRLS, 2016) bzw. wird von PBA auf CBA umgestellt (eTIMSS, 2019).

Aus Metanalysen ist ersichtlich, dass Richtung und Stärke des Moduseffekts von unterschiedlichen Faktoren, beispielsweise dem Gegenstandsbereich oder der Art der Testzusammenstellung, abhängen können (siehe Kingston, 2009; Wang, Jiao, Young, Brooks & Olson, 2008); ein weiteres Beispiel ist die Abhängigkeit vom Antwortformat

(Bennett, Braswell, Oranje, Sandene, Kaplan & Yan, 2008). Daraus folgt, dass in jeder Studie eine eigene empirische Überprüfung von Moduseffekten erforderlich ist, insofern diese als Folge einer unbekannten Mischung von Einzeleffekten geänderter Messeigenschaften angenommen werden (vgl. Kröhne & Martens, 2011).

Im 2014 durchgeführten Feldtest zu PISA 2015 wurden Moduseffekte überprüft, indem Schülerinnen und Schüler einer Schule zufällig einer Bedingung zugewiesen wurden, in der sie Aufgaben auf Papier vorgelegt bekamen oder einer Bedingung, in der sie die computerisierten PBA-Aufgaben bearbeiteten. Unter der Annahme zufallsäquivalenter Gruppen können auftretende Unterschiede in den Testleistungen auf Unterschiede in den Modi (CBA vs. PBA) zurückgeführt werden. Die Analyse der Feldtestdaten aller Teilnehmerstaaten ergab, dass in einem 2PL-Modell die Itemdiskriminationen zwischen den Modi nur geringfügig variierten, es aber bezüglich. der Itemschwierigkeiten Moduseffekte gab (ETS, 2015). Insgesamt erwiesen sich die CBA-Items als schwieriger. Unter der Annahme, dass zu diesem durchschnittlichen Moduseffekt auf Testebene nur eine Teilmenge von Items beitragen, wurde mithilfe von CBA-Items ohne Schwierigkeitsänderung gegenüber PBA (invariante Items) eine gemeinsame Skala gebildet. CBA-Items mit Moduseffekt auf die Schwierigkeit (nicht invariante Items) dürfen sich dagegen in der Skalierung hinsichtlich ihrer Schwierigkeit von PBA-Items unterscheiden. Für die Trendbestimmung ist zentral, dass der Moduseffekt mit den als invariant angenommenen Items in der internationalen Auswertung des Feldtests tatsächlich vollständig eliminiert werden konnte. Hinzu kommt die in den internationalen Analysen des Feldtests nicht weiter geprüfte Annahme, dass keine Interaktion des Modus mit Teilnehmerstaat vorliegt (ETS, 2015).

Fragestellungen

Der vorliegende Beitrag untersucht anhand der deutschen Stichproben der seit 2000 durchgeführten PISA-Erhebungen, inwiefern sich die international berichteten originalen Trendschätzungen für Deutschland in den drei Kompetenzdomänen (Naturwissenschaften,

Mathematik und Lesen) mit den marginalen Trendschätzungen reproduzieren lassen. Im Mittelpunkt stehen die folgenden Forschungsfragen.

Moduseffekte

Im ersten Schritt soll der Frage nachgegangen werden, wie sich die Trendschätzungen dadurch verändern können, dass der Testadministrationsmodus im Jahre 2015 umgestellt wurde, nämlich weg von Papier-und-Bleistift-Test hin zu computerbasierten Tests. Aus der Forschungsliteratur (z. B. Kröhne & Martens, 2011) ist bekannt, dass sich die Richtung von Moduseffekten, d. h. ob ein Testmodus Aufgaben erleichtert oder erschwert, nicht eindeutig ist. In diesem Sinne sollte letztendlich jede Umstellung von papierbasierter auf computerbasierte Testung empirische Studien zur Folge habe, die eine Abschätzung der Höhe des Moduseffekts erlauben. Im Rahmen der PISA 2015 Feldtestung ist genau solch eine Studie in den Teilnehmerstaaten durchgeführt worden. Die Auswertung der Daten über alle Staaten hinweg führte in allen drei Domänen (Lesen, Mathematik und Naturwissenschaften) zu einer Teilmenge invarianter Linkitems, auf deren Basis der Moduseffekt in 2015 kontrolliert werden sollte. Dieses Vorgehen ignoriert Interaktionseffekte Country × Modus, d. h. macht die starke (ungeprüfte) Annahme, dass die Moduseffekte für alle Teilnehmerstaaten identisch sind. Wir nehmen diese Problematik auf und fragen, wie durch das von der OECD gewählte Vorgehen zur Behandlung möglicher Moduseffekte, die Trendschätzungen für Deutschland beeinflusst wurden.

Effekte des Skalierungsmodells

Zweitens soll analysiert werden, inwieweit Trendschätzungen in PISA durch das Skalierungsmodell (1PL vs. 2PL) moderiert werden. Während in den PISA-Erhebungen 2000 bis 2012 die Itemkalibrierung auf der Basis des 1PL geschah, kam in PISA 2015 erstmalig das 2PL zum Einsatz. Letzteres weist üblicherweise einen besseren Modellfit auf, nimmt auf der anderen Seite aber auch eine Gewichtung der Items bei der Fähigkeitsschätzung vor, wohingegen die Items im 1PL mit identischer Gewichtung in die Fähigkeitsschätzung

eingehen. Anhand einer Serie von Skalierungsläufen gehen wir der Frage nach, ob sich diese unterschiedlichen Ansätze in variierenden Trendschätzungen widerspiegeln.

Unterschiede zwischen originalen und marginalen Trendschätzungen

Schließlich soll der Frage nachgegangen werden, wie groß die Unterschiede in den Trendschätzungen sind, wenn sich die Analysen nicht auf die internationalen Datensätze beziehen, sondern eine Beschränkung auf die deutschen Datensätze stattfindet. Die Forschungsliteratur (z. B. Monseur & Berezner, 2007) zeigt hier deutlich, dass erhebliche Abweichungen entstehen können, wenn die für die Trendschätzung verwendeten Items differenzielle Itemfunktionen in einzelnen Staaten haben. In PISA wird üblicherweise versucht, solche Effekte zu vermeiden, doch gelingt dieses nicht perfekt (vgl. auch Artelt & Baumert, 2004).

Studie 1: Untersuchung von Moduseffekten anhand der deutschen Feldtest-Studie für PISA 2015

Im PISA 2015 Feldtest wurden Schülerinnen und Schüler einer Schule zufällig einer Bedingung zugewiesen, in der sie Aufgaben auf Papier (PBA) oder dieselben Aufgaben auf dem Computer (CBA) vorgelegt bekamen. Im Folgenden soll anhand der deutschen Stichprobe des Feldtests überprüft werden, ob sich in Deutschland Moduseffekte für die Domänen Naturwissenschaften, Mathematik und Leseverstehen zeigten.

Methode

Die Auswertungen beruhen auf einer Teilstichprobe des im Frühjahr 2014 in Deutschland durchgeführten Feldtests zur PISA 2015 Studie mit $N = 517$ Schülerinnen und Schülern im PBA-Modus und $N = 506$ Schülerinnen und Schülern im CBA-Modus. Die Schülerinnen und Schüler innerhalb der 39 Schulen wurden zufällig den Bedingungen PBA bzw. CBA zugewiesen, wobei jede Teilnehmerin/ jeder Teilnehmer Items in zwei der drei Domänen (z. B. Naturwissenschaft und Lesen) bearbeitete. Aufgrund der zufälligen

Zuordnung zu den Bedingungen können auftretende Unterschiede in den Testleistungen kausal auf Unterschiede in den Modi (CBA vs. PBA) zurückgeführt werden.

Die zufällige Zuordnung des Administrationsmodus zu Schülerinnen und Schülern innerhalb von Schulen konnte anhand der Variablen Geschlecht (d = .05, SE = .03), Alter (d = .04, SE = .06) und Klassenstufe (d = .00, SE = .02) überprüft werden.

Die Skalierung der Leistungsdaten der Feldteststudie erfolgte mit einem 1PL-Modell für dichotome und polytome Items (Partial-Credit-Modell; Rost, 2004)[2]. Die Stichproben der beiden Administrationsmodi (CBA und PBA) wurden zunächst separat skaliert. Der mittlere Moduseffekt wurde durch ein anschließendes *Mean-Mean-Linking* der Itemschwierigkeiten bestimmt (Kolen & Brennan, 2014). Die Effektgröße des Moduseffekts d für eine Kompetenzdomäne wird durch die Relativierung der mittleren Differenz an der Standardabweichung der entsprechenden Fähigkeit im PBA-Modus bestimmt. Zusätzlich wurde die Standardabweichung für die Differenz der Itemschwierigkeiten zwischen den beiden Testmodi ermittelt (DIF-Standardabweichung; Camilli & Penfield, 1997). Die Standardfehler wurden mithilfe einer Double-Jackknife-Methode (Xu & von Davier, 2010) berechnet, in der sowohl die mit der Auswahl von Personen als auch der von Items verbundene Unsicherheit berücksichtigt wird. Dazu wurden die 39 Schulen als Jackknife-Zonen für die Auswahl der Personen verwendet. Für die Auswahl der Items wurden Testlets als Jackknife-Zonen verwendet, da einzelne Items häufig mit einem gemeinsamen Stimulus präsentiert wurden (Testlets; Monseur & Berezner, 2007). Für Naturwissenschaften wurden 28 Testlets, für Mathematik 38 Testlets und für Leseverstehen 24 Testlets als Jackknife-Zonen verwendet. Im Rahmen der Jackknife-Methode wurde zusätzlich eine Bias-Korrektur vorgenommen (Cameron & Trivedi, 2005).

[2] Für die Anpassung eines 2PL-Modells erwies sich die Stichprobengröße als zu klein. Die Stichprobengrößen pro Item lagen im PBA-Modus zwischen N = 108 und N = 125 (mittleres N = 116.9) sowie im CBA-Modus zwischen N = 96 und N = 115 (mittleres N = 108.4).

Die OECD hat auf eine länderspezifische Auswertung der Feldtestdaten verzichtet und nur Analysen durchgeführt, in denen die Daten aller Staaten zusammengefasst wurden (ETS, 2015; Heine et al., 2016). In diesen Analysen wurden Items identifiziert, die unter den beiden Bedingungen (CBA und PBA) dieselben statistischen Eigenschaften aufwiesen (sogenannte invariante Items). Diese invarianten Items sollten – zumindest auf internationaler Ebene – nicht von einem Moduseffekt betroffen sein. Motiviert durch dieses Vorgehen haben wir das Mean-Mean-Linking der Itemschwierigkeiten für die deutschen Feldtestdaten unter zwei Bedingungen durchgeführt. In einer ersten Analyse wurden jeweils alle Items einer Domäne für das Mean-Mean-Linking berücksichtigt, also auch die Items, die in der Auswertung der internationalen Stichprobe der Feldtestdaten als nicht invariant identifiziert wurden (alle Items). In einer zweiten Analyse wurde das Linking für jede Domäne jeweils nur auf Basis der laut OECD invarianten Items durchgeführt (invariante Items). Mithilfe dieser zweiten Analyse konnte überprüft werden, inwiefern die in der internationalen Auswertung invarianten Items in der deutschen Feldteststichprobe von einem mittleren Moduseffekt betroffen sind.

Für alle folgenden statistischen Auswertungen wurde die Software R (R Core Team, 2016) sowie die R-Pakete TAM (Kiefer, Robitzsch & Wu, 2016) und sirt (Robitzsch, 2016) verwendet.

Ergebnisse

Tabelle 1 zeigt die Ergebnisse für die Prüfung eines Moduseffekts in den drei Domänen Naturwissenschaften, Mathematik und Lesen für den deutschen Feldtest. Zunächst werden die Ergebnisse auf Basis von allen Items dargestellt (alle Items), die im Feldtest administriert wurden. Für alle drei Kompetenzdomänen zeigte sich ein negativer Effekt des CBA-Modus im Vergleich zum PBA-Modus, d. h. Aufgaben am Computer fielen schwerer aus als auf dem Papier. Dabei war der Moduseffekt in Naturwissenschaften und Mathematik statistisch signifikant von Null verschieden. Insgesamt waren die Effekte von $d = -.23$

(Naturwissenschaften), $d = -.14$ (Mathematik) bzw. $d = -.13$ (Lesen) substanziell. Differenzen in den Moduseffekten zwischen den Kompetenzdomänen fielen nicht statistisch signifikant aus (Wald-Test: $\chi^2 = 1.39$, $df = 2$, $p = .50$), d. h. der Moduseffekt bestand unabhängig von der untersuchten Kompetenzdomäne. Des Weiteren zeigte die Standardabweichung der itemspezifischen Moduseffekte (SD_{Modus}), dass die Differenz der Itemschwierigkeiten zwischen den beiden Testmodi erheblich über die Items variierte und somit durch den Wechsel des Testmodus nicht lediglich eine konstante Verschiebung in den Itemschwierigkeiten induziert wurde. Es wird deutlich, dass die auf differentielle Moduseffekte zurückzuführende Variabilität der Itemschwierigkeiten besonders stark für die Domäne Leseverstehen und nur relativ schwach für die Naturwissenschaften ausgeprägt war (siehe für eine Klassifikation von Effektstärken Camilli & Penfield, 1997).

--- hier ewa Tabelle 1 einfügen ---

In einer zweiten Analyse wurden im Linking nur die laut OECD invarianten Items berücksichtigt. Auch bei Einschränkung auf diese Itemmenge zeigten sich Effekte des Testmodus, wobei sie insgesamt etwas schwächer ausfielen und nur noch für die Domäne Naturwissenschaften signifikant von Null verschieden waren. Dies legt nahe, dass zumindest für Deutschland durch die von der OECD ausgewählten invarianten Items Moduseffekte nur unvollständig adjustiert werden. Die Standardabweichung der itemspezifischen Moduseffekte fiel etwas geringer aus, blieb aber in den Domänen Mathematik und Leseverstehen weiterhin bedeutsam.

Studie 2: Trendschätzungen in PISA für Deutschland mit nationalen Skalierungen

Im Folgenden soll für die deutsche PISA Stichprobe untersucht werden, wie sensitiv die Trendschätzungen gegenüber der in PISA 2015 vorgenommenen Veränderung des Skalierungsmodells und des Administrationsmodus sind. Dazu wurden für alle drei Kompetenzdomänen die deutschen Stichproben der PISA-Erhebungen mithilfe verschiedener

Ansätze sowohl mit als auch ohne Berücksichtigung des in der Feldstudie identifizierten Moduseffekts skaliert.

Methode

Tabelle 2 gibt einen Überblick über die Stichproben der verwendeten PISA-Studien. Aufgeführt sind jeweils die Zahlen der Schülerinnen und Schüler, denen Items einer Domäne in einer Erhebung administriert wurden. Dabei gilt es zu beachten, dass nur dann allen Schülerinnen und Schülern die Items einer Domäne vorgelegt wurden, wenn sie Hauptdomäne in der entsprechenden Erhebung war. Zusätzlich wird die Anzahl der insgesamt in einer Domäne zu einer Erhebung vorgelegten Items aufgeführt. Eine Teilmenge dieser Items wurde in unseren Analysen als Linkitems eingesetzt. Beispielsweise wurden in Naturwissenschaften 2006 insgesamt 103 Items vorgelegt, wovon 77 Items als Linkitems verwendet wurden. Ein Linkitem muss in mindestens zwei PISA-Erhebungen administriert worden sein (siehe Appendix A im elektronischen Supplement für eine vollständige Dokumentation aller verwendeten Linkitems).

--- hier etwa Tabelle 2 einfügen ---

Für die Trendschätzungen wurde in Analogie zum internationalen Vorgehen jeweils die Erhebung als Ausgangspunkt gewählt, bei der eine Leistungsdomäne zum ersten Mal eine Hauptdomäne war (Naturwissenschaften: PISA 2006, Mathematik: PISA 2003, Lesen: PISA 2000). Für die Behandlung eines möglichen Moduseffekts können bei der Skalierung zwei Strategien unterschieden werden. In der ersten Vorgehensweise wurde der Feldtest nicht in der Skalierung berücksichtigt. Diese Strategie ist für die Domäne Naturwissenschaft in der linken Grafik in Abbildung 1 dargestellt. Es wird für die marginale Trendschätzung angenommen, dass für die Items, die im CBA-Modus (PISA 2015) und in mindestens einer der früheren Erhebungen vorgelegt wurden, kein mittlerer Moduseffekt besteht. Diese Strategie wurde sowohl für alle Items als auch die in der internationalen Auswertung des Feldtests als invariant identifizierten Items durchgeführt.

--- hier etwa Abbildung 1 einfügen---

In der zweiten Strategie wurde die Vergleichbarkeit der CBA- und PBA-Items mithilfe der Feldteststudie hergestellt (siehe rechte Grafik Abbildung 1). Unter der Annahme zufallsäquivalenter Gruppen im Feldtest können Unterschiede in den Testleistungen zwischen den beiden Gruppen ausschließlich auf Unterschiede zwischen Items im CBA- und PBA-Modus zurückgeführt werden. Die Items der früheren PISA-Erhebungen (bis einschließlich 2012) in einer Domäne wurden dann mit den Items des Feldtests im PBA-Modus verlinkt. Analog wurde ein Linking der Items aus PISA 2015 mit den Items des Feldtests im CBA-Modus vorgenommen. Durch dieses Vorgehen wird eine gemeinsame Metrik für alle PISA-Erhebungen in einer Domäne etabliert, die mögliche Moduseffekte auf allen Items mit Rückgriff auf den deutschen Feldtest als Brückenstudie berücksichtigt (*bridge study*; Mazzeo & von Davier, 2008).

Für die technische Durchführung der Skalierung wurde zwischen konkurrenten Skalierungen und separaten Skalierungen mit anschließendem Linking unterschieden (Kolen & Brennan, 2014). Bei der konkurrenten Skalierung wurden unter der Annahme invarianter Itemparameter die einzelnen Erhebungen als Gruppen in einem Mehrgruppen-IRT-Modell behandelt (Bock & Moustaki, 2007). In der separaten Skalierung wurden die einzelnen Erhebungen zunächst getrennt skaliert und anschließend die Itemparameter entweder in einem simultanen Linking nach dem von Haberman (2009) vorgeschlagenen Regressionsansatz oder in einem schrittweisen Linking aufeinanderfolgender Erhebungen verlinkt (*chain linking*; Kolen & Brennan, 2004). Zur Bestimmung der Fähigkeitsverteilungen wurden für alle Skalierungsmodelle jeweils 20 Plausible Values ohne weitere Kovariaten im Hintergrundmodell gezogen (von Davier & Sinharay, 2014). Alle Skalierungen und die Analysen mit den Plausible Values wurden unter Berücksichtigung der Stichprobengewichte

durchgeführt[3]. Insgesamt ergaben sich somit unter zusätzlicher Berücksichtigung der Wahl eines 1Pl- oder 2PL-Modells als Skalierungsmodell die folgenden 12 Methoden:

In der Methode C1 wurde eine konkurrente Skalierung (concurrent calibration) nach dem 1PL-Modell ohne Berücksichtigung der deutschen Feldteststudie auf Basis aller Items vorgenommen. Die Methode C2 beruht auf einer konkurrenten Skalierung nach dem 2PL-Modell (Generalized-Partial-Credit-Modell; Rost, 2004) ohne Berücksichtigung des Feldtests. In der Methode H1 wurde eine separate Skalierung nach dem 1PL-Modell vorgenommen. Anschließend wurden die aus den Skalierungen erhaltenen Itemschwierigkeiten mit dem Regressionsansatz von Haberman (2009) verlinkt. Dabei wurden die Linkitems aller PISA-Studien – ohne Berücksichtigung der deutschen Feldteststudie – verwendet. Die Methode H2 ist analog zur Methode H1, nur dass die separate Skalierung mithilfe des 2PL-Modells vorgenommen wurde und dann im nächsten Schritt die Itemschwierigkeiten und Diskriminationsparameter mithilfe der Methode von Haberman (2009) verlinkt wurden. In der Methode S1 wurde wieder zunächst eine separate Skalierung nach dem 1PL-Modell durchgeführt. Das Linking der verschiedenen Studien wurde dann allerdings schrittweise durchgeführt. Dabei wird jeweils eine nachfolgende Studie (z. B. 2009) auf eine davorliegende Studie (z. B. 2006) mithilfe eines Mean-Mean-Linkings verlinkt.

In der Methode C1I wurden im Gegensatz zur Methode C1 bei der konkurrenten Skalierung nur die Items über die Studien gleichgesetzt, die von der OECD als invariant identifiziert wurden. Die nicht invarianten Items erhielten 2015 (unter CBA) andere Itemparameter als in den Wellen zuvor (bis 2012 unter PBA). Die Methode C2I ist analog zur Methode C1I, mit dem Unterschied, dass dass ein 2PL-Modell bei der konkurrenten Skalierung verwendet wurde. Es entspricht somit weitgehend der Auswertungsstrategie der PISA-Erhebung 2015 (siehe Heine et al., 2016). In der Methode H1I wurde wie in Methode H1 ein Linking nach der Methode von Haberman (2009) durchgeführt, wobei lediglich die

[3] Für die Feldtestdaten lagen keine Stichprobengewichte vor. Diese Schülerinnen und Schüler wurden deshalb gleichgewichtet.

von OECD als invariant identifizierten Items über die Studien eingesetzt wurden. Dieses Vorgehen entspricht weitgehend der Auswertungsstrategie in den PISA-Erhebungen 2000 bis 2012. Die Methode H2I geht analog zur Methode H1I vor, mit dem Unterschied, dass dass ein 2PL-Modell für die separate Skalierung verwendet wurde.

In der Methode C1F wurde eine konkurrente Skalierung nach dem 1PL-Modell unter Berücksichtigung der deutschen Feldteststudie vorgenommen. Für alle im PBA-Modus eingesetzten Items (Erhebungen 2000 bis 2012 sowie im deutschen Feldtest) wurde in der Skalierung von invarianten Itemschwierigkeiten ausgegangen. Ebenso wurde Invarianz für die im CBA-Modus eingesetzten Items angenommen. Eine gemeinsame Metrik aller Erhebungen wurde durch die Spezifikation gleicher Fähigkeitsverteilungen der PBA- und CBA-Stichproben im Feldtest etabliert. Aufgrund der kleinen Stichprobengrößen pro Item im Feldtest wurde auf die Anwendung des 2PL-Modells verzichtet. In der Methode H1F wurde eine separate Skalierung mithilfe des 1PL-Modells durchgeführt und anschließend die Itemschwierigkeiten nach dem Regressionsansatz von Haberman (2009) unter Berücksichtigung der deutschen Feldteststudie verlinkt. Dazu wurde der Regressionsansatz sowohl für die Items im PBA-Modus als auch die Items im CBA-Modus angewendet. Die gemeinsame Metrik wurde dann wieder über die Annahme gleicher Fähigkeitsverteilungen der PBA- und CBA-Stichproben im Feldtest erhalten (siehe rechte Grafik Abbildung 1). Die Methode S1F führte zunächst eine separate Skalierung mithilfe des 1PL-Modells durch. Im Anschluss wurde ein schrittweises Linking sowohl für die Items im PBA-Modus als auch CBA-Modus durchgeführt. Analog zur Methode H1F wurde eine gemeinsame Metrik durch die Annahme gleicher Fähigkeitsverteilung der PBA- und CBA-Stichproben im Feldtest etabliert.

Die Bestimmung der Standardfehler für die Trendschätzungen folgte dem internationalen Vorgehen. Die auf die Auswahl der Personen zurückzuführende Unsicherheit wurde durch die Balanced-Repeated-Replication (BRR)-Methode auf Basis der im

Originaldatensatz vorhandenen 80 Replikationszonen durchgeführt (Heine et al., 2016). Linkfehler zur Erfassung der mit der Auswahl von Items verbundenen Unsicherheit wurden durch ein Jackknife von Items bestimmt, wobei analog zum Feldtest Testlets als Jackknife-Zonen verwendet wurden (Naturwissenschaften 28 Testlets, Mathematik 38 Testlets und Leseverstehen 24 Testlets). Der Standardfehler (SE_{tot}) ergibt sich durch Addition der quadrierten Standardfehler auf der Personen- und Itemseite (Heine et al., 2016).

Ergebnisse

Für das bessere Verständnis der später berichteten marginalen Trendschätzungen ist es instruktiv, zunächst die aus separaten Skalierungen der deutschen Stichproben mithilfe eines 1PL-Modells gewonnenen Itemschwierigkeiten deskriptiv zu betrachten (siehe Tabelle 3). Dazu wurden die Fähigkeiten zu jeder Erhebung zentriert (Mittelwerte der Verteilung gleich 0), sodass Veränderungen in den mittleren Schwierigkeiten mit einer Änderung der mittleren Leistung verbunden sind. Items, die in denselben PISA-Erhebungen vorgelegt wurden, wurden zu Itemgruppen zusammengefasst. In Mathematik können z. B. zwei Itemgruppen unterschieden werden (bezeichnet mit M1A und M1B). Die 31 Items der Gruppe M1A kamen in allen Erhebungen zwischen 2003 und 2015 zum Einsatz. Es wird ersichtlich, dass sich die mittlere Itemschwierigkeit dieser Gruppe von 2003 nach 2012 deutlich verringerte (–.18 – .01 = –.19 logits) und demzufolge ein positiver Trend in der mathematischen Kompetenz über die 9 Jahre zu verzeichnen ist. Dagegen stieg die mittlere Itemschwierigkeit von 2012 nach 2015 sowohl in Itemgruppe M1A als auch der Gruppe M1B (nur 2012 und 2015 verwendete Items) wieder an, was mit einem Abfall der mathematischen Leistung über diesen Zeitraum verbunden ist. Dieser Abfall von 2012 nach 2015 sollte aber an der Differenz der mittleren Itemschwierigkeiten zwischen dem PBA-Modus und dem CBA-Modus (Feldtest 2014) relativiert werden. Es fällt auf, dass nach einer Adjustierung der Unterschiede in Schwierigkeiten zwischen den Modi (z. B. für PISA 2015 Itemgruppe M1A: –0.07 + (–

0.10) = –.17), nahezu keine Differenzen mehr in den mittleren Itemschwierigkeiten zwischen 2012 und 2015 bestehen.

--- hier etwa Tabelle 3 einfügen---

Des Weiteren zeigte sich, dass die Trendschätzungen teilweise stark davon abhängen, auf welchen Itemgruppen die Veränderungen in den Kompetenzen betrachtet werden. Dies wird besonders in der Domäne Leseverstehen für die Trendschätzung von 2000 nach 2009 deutlich. So betrug für die Itemgruppe R1A zwischen 2000 und 2009 der Unterschied in den Itemschwierigkeiten –0.28, wohingegen er für die Itemgruppe R1B mit –0.12 erheblich geringer ausfiel. Der Zuwachs im Leseverstehen wäre somit für die Itemgruppe R1A deutlich stärker ausgeprägt als für die Itemgruppe R1B. Diese Unterschiede haben direkte Konsequenzen für die Ergebnisse der eingesetzten Linking-Methoden. Da bei einem schrittweisen Linking (Methode S1) von 2000 nach 2009 über die Studien 2003 und 2006 immer nur Items herangezogen werden, die in nachfolgenden Studien auftreten, wird im schrittweisen Ansatz die Trendschätzung nur auf Basis der Itemgruppe R1A ermittelt. Bei einem gemeinsamen Linking (Methode H1) gehen dagegen beide Itemgruppen (R1A und R1B) in die Trendschätzung ein, sodass mit der Methode H1 ein weniger stark ausgeprägter Trend für Leseverstehen resultieren würde als mit der Methode S1.

Tabelle 4 zeigt die Ergebnisse der Trendschätzungen für die Naturwissenschaften (Mittelwerte und Trend für 2012 nach 2015). In der Zeile „Original“ sind die in den internationalen Berichten der jeweiligen PISA-Erhebungen aufgeführten Leistungswerte berichtet. Es wird deutlich, dass die Ergebnisse für die konkurrenten Skalierungen mit den Methoden C1 und C2 auf Basis aller Items weitgehend mit den internationalen Befunden übereinstimmen (z. B. für PISA 2015 wurden 509 Punkte für Deutschland berichtet und die Analysen mit den deutschen Stichproben ergeben jeweils 508 Punkte). Auch die separaten Skalierungen mit anschließendem Linking führten sowohl im schrittweisen Ansatz (Methode S1) als auch nach dem Regressionsansatz von Haberman (Methoden H1 und H2) mit jeweils

511, 506 und 501 Punkten in PISA 2015 zu einer ähnlichen Trendschätzung. Des Weiteren zeigt sich, dass die Trendschätzung relativ stabil gegenüber der Wahl eines 1PL-Modells oder 2PL-Modells ist. Die Methoden, die nur auf den laut OECD invarianten Items beruhen (Methoden C1I, C2I, H1I und H2I) führen mit 506 bis 513 Punkten zu leicht höher ausgeprägten Trendschätzungen. Der Trend ist allerdings immer noch negativ.

---hier etwa Tabelle 4 einfügen---

Alle neun Methoden, in denen ein möglicher Moduseffekt gar nicht oder nur auf einer Teilmenge von nicht-invarianten Items berücksichtigt wurde, führten somit zur Schätzung eines statistisch signifikanten negativen Leistungstrends in den Naturwissenschaften. Wenn in den Trendschätzungen dagegen die deutsche Feldteststudie bzw. der für diese Stichprobe identifizierte Moduseffekt statistisch berücksichtigt wurde, zeigt sich ein leicht positiver, aber nicht signifikanter Trend (PISA 2015: 528, 528 und 531 Punkte). Dies gilt relativ unabhängig davon, ob eine konkurrente (C1F) oder eine separate Skalierung (H1F und S1F) vorgenommen wurde.

Für die Domäne Mathematik (siehe Tabelle 5) zeigten sich im Gegensatz zum international berichteten Trend bei Berücksichtigung des deutschen Feldtests (und somit möglicher Moduseffekte für Deutschland) nahezu keine Veränderungen der Leistung in PISA 2015 gegenüber dem Abschneiden in PISA 2012. Auch die Wahl des Skalierungsmodells (1PL vs. 2PL) besitzt für die Mathematikleistung nahezu keinen Einfluss. Es fällt allerdings auf, dass eine Beschränkung der Trendschätzung auf die laut OECD invarianten Items zu einer stärkeren Korrektur führte, die mit ähnlichen Schlussfolgerungen wie die Analyse unter Hinzunahme des deutschen Feldtests verbunden ist. Des Weiteren wird deutlich, dass die nationalen Trends teilweise stark von den international berichteten Mittelwerten abweichen (z. B. in PISA 2006 werden original 504 Punkte berichtet und in der marginalen Trendschätzung ohne Feldtest liegt er zwischen 512 und 517 Punkten).

--- hier etwa Tabelle 5 einfügen---

Die Befunde für das Leseverstehen (siehe Tabelle 6) erweisen sich im Vergleich zu den beiden anderen Domänen als etwas instabiler (d. h. stärker sensitiv gegenüber der Wahl der Auswertungsmethode). Während sich im international berichteten Trend von 2012 nach 2015 nahezu keine Veränderung zeigt, ist für die marginalen Trendschätzungen ohne Berücksichtigung des Feldtests auf Basis aller Items ein deutlicher Abfall zu verzeichnen (Methoden C1, C2, H1, H2 und S1). Wurden dagegen nur die invarianten Items gewählt oder auf den Feldtest zurückgegriffen, dann zeigte sich eine positive Trendschätzung für Lesen von 2012 nach 2015. Des Weiteren fällt auf, dass in Lesen die Analysen mit den deutschen PISA-Stichproben auch in den früheren PISA-Erhebungen zu teilweise erheblichen Abweichungen vom internationalen Trend führen. Diese Abweichungen könnten darauf zurückgeführt werden, dass sich der mittlere Country-DIF für Deutschland auf den Linkitems und den Nicht-Linkitems unterschied. Insgesamt stehen aber auch für das Leseverstehen die Analysen im Einklang mit der Annahme, dass die Wahl des Skalierungsmodells (1PL- vs. 2PL-Modell) nur einen relativ geringen Einfluss besitzt und die Behandlung möglicher Moduseffekte (sowohl durch die Beschränkung auf invariante Items als auch die Adjustierung um den im Feldtest ermittelten Moduseffekt) zu einem eher positiv ausgeprägten Leistungstrend führt.

--- hier etwa Tabelle 6 einfügen---

Abbildung 2 fasst die Befunde der verschiedenen Methoden zur marginalen Trendschätzung in Abhängigkeit der Berücksichtigung möglicher Moduseffekte zusammen. Dargestellt sind die Mittelwerte sowie Minimum und Maximum (vertikale graue Balken) über drei Methodengruppen: Methoden, die den Feldtest nicht berücksichtigen und alle Items für das Linking verwenden, Methoden, die den Feldtest nicht berücksichtigen und nur die invarianten Items verwenden, sowie Methoden, die den Feldtest berücksichtigen. Für die Naturwissenschaften zeigt sich deutlich, dass nur die Methoden, die die Feldtestdaten berücksichtigen, zu einem positiven Trend über alle vier Erhebungen führen, und ansonsten vor 2015 nahezu keine Unterschiede zwischen originalem und marginalem Trend bestehen.

Auch in Mathematik und Lesen besteht ein positiver Trend über die Erhebungen, wenn der Feldtest in der marginalen Trendschätzung herangezogen wird. Es bestehen allerdings stärkere Abweichungen zwischen originalen und marginalen Trends.

---hier etwa Abbildung 2 einfügen---

Diskussion

Große internationale Schulleistungsstudien verfolgen das Ziel, die Leistungsfähigkeit von Bildungssystemen im internationalen Vergleich zu untersuchen. Eine besondere Bedeutung kommt dabei den Trendschätzungen zu, in denen sich abbilden soll, ob und in welchem Ausmaß sich die Leistungsfähigkeit eines Bildungssystems verändert hat. Auf bildungspolitischer Ebene werden solche Trends (im positiven Fall) genutzt, um post hoc eigenes Handeln im bildungspolitischen Bereich zu legitimieren oder (im negativen Fall) neue Reformmaßnahmen zu initiieren. Die Glaubwürdigkeit und Belastbarkeit solchen Handelns wie auch der Ergebnisse hängt entscheidend davon ab, wie valide bzw. robust die vorgenommenen Trendschätzungen sind. Der vorliegende Beitrag untersuchte daher anhand einer Analyse der deutschen PISA-Stichproben, wie sich Trendschätzungen verändern können, wenn der Testadministrationsmodus (Papier vs. Computer) oder das Analysemodell (1PL- vs. 2PL-Modell) über den Erhebungszeitraum verändert werden. Während die Wahl des Analysemodells nur einen geringen Einfluss auf die Trendschätzungen besaß, legen die Analysen unter Verwendung der Daten deutschen Feldteststudie nahe, dass der für Deutschland berichtete Abfall in den Leistungsdomänen von 2012 nach 2015 (zumindest zum Teil) durch einen Wechsel des Testadministrationsmodus verursacht sein könnte. In der öffentlichen und bildungspolitischen Diskussion der PISA-2015-Ergebisse sollte daher berücksichtigt werden, dass der Leistungsabfall möglicherweise überhaupt nicht aufgetreten wäre, wenn die Testung erneut mit Papier-und-Bleistift-Tests stattgefunden hätte. Wir wollen im Folgenden die potenziellen Limitationen unserer Befunde diskutieren und auch andere Faktoren reflektieren, welche den Trend moderiert haben könnten.

Belastbarkeit der Daten der deutschen Feldteststudie

In der internationalen Auswertung der PISA 2015 Studie wurde von der OECD auf eine länderspezifische Auswertung der Feldtestdaten verzichtet und die einzelnen Länderstichproben für die weiteren Analysen zu einer Gesamtstichprobe zusammengefasst. Auf Basis dieser Gesamtstichprobe wurden dann Itemanalysen und Moduseffekte untersucht (Heine et al., 2016). Ohne Frage lässt sich dieses Vorgehen legitimieren, da die länderspezifischen Stichproben relativ klein für belastbare Schlussfolgerungen über die Ausprägung von Moduseffekten waren. Zweifelsohne wäre eine noch größere Stichprobe in der Feldteststudie erstrebenswert gewesen, da sie zu insgesamt präziseren Schätzungen von Effektgrößen geführt hätte. In kleineren Stichproben sind die Schätzungen der Effekte natürlich variabler, was sich auch in den größeren Standardfehlern für die Analysen mit der deutschen Stichprobe der Feldtestdaten widerspiegelt. Die Moduseffekte ließen sich allerdings trotz einer relativ kleinen Stichprobe für zwei der drei Domänen (Naturwissenschaften, Mathematik) in den deutschen Feldtestdaten inferenzstatistisch absichern. Des Weiteren wurde ein 1PL-Modell, das geringere Anforderungen an die Stichprobengröße stellt als das 2PL, zur Abschätzung der Moduseffekte verwendet.

Zweitens kann man gegen eine länderspezifische Auswertung der Feldtestdaten anführen, dass es sich nicht um eine repräsentative Stichprobe für das jeweilige Land handele. So gilt für die deutsche Feldtest-Stichprobe, dass sie keine für Deutschland repräsentative Zufallsstichprobe ist, da nicht aus allen Bundesländern Schulen ausgewählt wurden. Obwohl dies durchaus als eine Einschränkung angesehen werden kann, erscheint nicht ganz nachvollziehbar, warum dies zu substanziellen systematischen Verzerrungen in der Schätzung von Moduseffekten führen sollte. Aufgrund des experimentellen Designs werden systematische, auf Schulen zurückzuführende Effekte (die nicht mit dem Moduseffekt korrelieren) ausgeschlossen, da die zufällige Zuweisung des Testmodus zu den Schülerinnen und Schülern innerhalb von Schulen vorgenommen wurde.

Drittens wird die Belastbarkeit der Feldteststudie möglicherweise auch dadurch eingeschränkt, dass sie von vielen Staaten zur Prüfung von Testabläufen (z. B. Abschätzung der Testzeit, Umsetzbarkeit von computerbasierter Erhebung etc.) verwendet wurde und deshalb hinsichtlich der Durchführungsbedingungen nicht direkt vergleichbar mit dem PISA-Haupttest war. Dagegen ließe sich einwenden, dass dies auch für die internationale Auswertung der Feldtestdaten zutreffen würde und es wenig plausibel erscheint, dass sich systematische Verzerrungen auf Einzelstaatenebene (wenn sie denn vorliegen sollten) über die Staaten ausmitteln sollten.

Änderung des Science Frameworks in PISA 2015

Eine alternative Erklärung des vergleichsweise starken Abfalls der Leistung in Naturwissenschaft von 2012 nach 2015 könnte in der Änderung des Science Framework liegen und der damit verbundenen Administration von neuen Aufgabentypen in Naturwissenschaften in PISA 2015. Diese Aufgaben sind interaktiver gestaltet und zielen stärker auf Aspekte des naturwissenschaftlichen Experimentierens ab als die bisherigen Aufgabentypen. In weiteren Analysen sind wir der Frage nachgegangen, inwiefern die neuen Aufgabentypen dazu geführt haben könnten, dass sich die Schwierigkeiten der alten Aufgaben (bis 2012) im PISA-Test 2015 verändert haben. Dazu wurde der deutsche Trend in den Naturwissenschaften nur auf Basis der alten Items untersucht und nur Testhefte verwendet, in denen die alten Items vor den neuen Items administriert wurden. Insgesamt zeigten sich keine größeren Abweichungen zu den in Tabelle 4 berichteten Trends (siehe zusätzliche Analysen im Appendix B des elektronischen Supplements). Dies spricht gegen die Annahme, dass der abfallende Trend in den Naturwissenschaften primär durch den Wechsel des Aufgabenmaterials verursacht sei. Auf nationaler Ebene erwiesen sich zudem die durch Linkitems (alte Aufgaben) und neue Items gebildeten Dimensionen als nahezu perfekt korreliert.

Unterschiede von originalem und marginalem Trend für Deutschland

In den Reanalysen der deutschen PISA-Erhebungen zeigten sich, insbesondere in den Domänen Mathematik und Lesen, teilweise deutliche Unterschiede zwischen dem originalen und marginalen Trend. Bei der Interpretation dieser Unterschiede gilt es zu berücksichtigen, dass wir in den Reanalysen mit der deutschen Stichprobe in einigen technischen Details von dem internationalen Vorgehen abgewichen sind. Im Gegensatz zu den internationalen Analysen haben wir bei der Ziehung der Plausible Values keine weiteren Kovariaten im Hintergrundmodell berücksichtigt. Außerdem haben wir uns in unseren Analysen jeweils auf die Teilstichproben von Schülerinnen und Schüler zu einer Erhebung beschränkt, die Items einer Domäne vorgelegt bekamen. In der internationalen Auswertung von PISA werden auch Plausible Values für Schülerinnen und Schüler in einer Domäne generiert, wenn sie aufgrund des komplexen Testdesigns gar keine Items in dieser Domäne bearbeitet haben.

Des Weiteren muss betont werden, dass die marginalen Trendschätzungen, obwohl sie robuster sind, die originalen Trendschätzungen nicht ersetzen können (Gebhardt & Adams, 2007). Nur die originale Trendschätzung erlaubt es, dass die Trends verschiedener Teilnehmerstaaten direkt miteinander verglichen werden. Mit den marginalen Trendanalysen des vorliegenden Beitrags kann also nicht das Ziel verfolgt werden, die international berichteten Trendanalysen zu ersetzen. Allerdings demonstrieren unsere Analysen, wie sensitiv Trendschätzungen gegenüber Veränderungen des Testmodus sein können, und dass mögliche Moduseffekte auch im originalen Trend zu substantiellen Einbußen in der gemessenen Kompetenz sowohl im internationalen Trend als auch in anderen Teilnehmerstaaten in PISA 2015 geführt haben könnten. Die hier berichteten Befunde illustrieren weiteren Forschungsbedarf zur methodischen Berücksichtigung und zur nationalen sowie internationalen inhaltlichen Interpretation von Moduseffekten (vgl. Heine et al., 2016).

Fazit

Mit PISA 2015 wurde der notwendige Schritt im Testmodus weg vom Papier zum Computer vollzogen. Neben dieser Umstellung wurde auch das Skalierungsmodell geändert. Da bisher

keine Analysen vorliegen, ob beide Veränderungen Effekte auf die deutschen Trendschätzungen hatten, sind wir anhand der deutschen Daten ab PISA 2000 beiden Fragen detaillierter nachgegangen und kommen zu bemerkenswerten Ergebnissen. Die Befunde machen für Deutschland das Folgende deutlich:

- Der Wechsel des Skalierungsmodells steht nicht in Zusammenhang mit dem Abfall der Leistungen in Mathematik und den Naturwissenschaften.
- Die Feldtestdaten aus dem Jahr 2014 machen deutlich, dass PISA-Aufgaben für deutsche 15-jährige im Mittel schwerer werden, wenn sie mittels Computer administriert werden (sogenannter Modus-Effekt).
- Der negative Effekt der Computeradministration auf die Leistungen der 15-jährigen zeigt sich in allen drei Domänen.
- Unter der Annahme, dass die Moduseffekte in PISA 2015 bei den deutschen Schülerinnen und Schülern genauso stark ausgeprägt waren wie in 2014, würde sich für die Veränderungen zwischen 2012 und 2015 ergeben, dass die Leistungen in Mathematik und den Naturwissenschaften unverändert geblieben wären und sich im Lesen leicht verbessert hätten.

Die internationalen Berichte zu Leistungsveränderungen zwischen 2012 und 2015 sollten daher nur vorsichtig interpretiert werden. Es lässt sich nicht ausschließen, dass beispielsweise die eher geringere Vertrautheit deutscher 15-jähriger mit Computern im schulischen Kontext ein Absinken der mathematischen und naturwissenschaftlichen Kompetenzen verursacht haben könnte.

Literatur

Aitkin, M. & Aitkin, I. (2011). *Statistical modeling of the National Assessment of Educational Progress*. New York: Springer.

Artelt, C. & Baumert, J. (2004). Zur Vergleichbarkeit von Schülerleistungen bei Leseaufgaben unterschiedlichen sprachlichen Ursprungs. *Zeitschrift für Pädagogische Psychologie*, *18*, 171–185.

Artelt, C., Weinert, S. & Carstensen, C. H. (2013). Assessing competencies across the lifespan within the German National Educational Panel Study (NEPS) – Editorial. *Journal for Educational Research Online, 5* (2), 5–14.

Baumert, J., Klieme, E., Neubrand, M., Prenzel, M., Schiefele, U., Schneider, W. et al. (Hrsg.). (2001). *PISA 2000: Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich*. Opladen: Leske + Budrich.

Beaton, A. E. (1988). *Disentangling the NAEP 1985-86 reading anomaly*. *Technical report, ETS*. Princeton, New Jersey, USA.

Bennett, R. E., Braswell, J., Oranje, A., Sandene, B., Kaplan, B. & Yan, F. (2008). Does it matter if I take my mathematics test on computer? A second empirical study of mode effects in NAEP. *The Journal of Technology, Learning, and Assessment, 6* (9). Verfügbar unter http://www.jtla.org

Birnbaum, A. (1968). Some latent trait models. In F. M. Lord & M. R. Novick, (Eds.), *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Bock, R. D. & Moustaki, I. (2007). Item response theory in a general framework. In C. R. Rao & S. Sinharay (Eds.), *Handbook of Statistics*, *Volume 26: Psychometrics* (pp. 469–513). North Holland: Elsevier.

Bos, W., Hornberg, S., Arnold, K.-H., Faust, G., Fried, L., Lankes, E.-M. et al. (Hrsg.). (2007). *IGLU 2006. Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich*. Münster: Waxmann

Brennan, R. L. (2001). Some problems, pitfalls, and paradoxes in educational measurement. *Educational Measurement: Issues and Practice, 20* (4), 6–18.

Cameron, A. C. & Trivedi, P. K. (2005). *Microeconometrics*. New York: Cambridge University Press.

Camilli, G. & Penfield, D. A. (1997). Variance estimation for differential test functioning based on Mantel-Haenszel statistics. *Journal of Educational Measurement*, *34*, 123–129.

Carstensen, C. H., Prenzel, M. & Baumert, J. (2008). Trendanalysen in PISA: Wie haben sich die Kompetenzen in Deutschland zwischen PISA 2000 und PISA 2006 entwickelt? *Zeitschrift für Erziehungswissenschaften, 10*, 11–34.

von Davier, M. & Sinharay, S. (2014). Analytics in international large-scale assessments: Item response theory and population models. In L. Rutkowski, M. von Davier & D. Rutkowski (Eds.), *Handbook of international large-scale assessment* (pp. 155–174). Boca Raton: CRC Press.

Ehmke, T., Klieme, E. & Stanat, P. (2013). Veränderungen der Lesekompetenz von PISA 2000 nach PISA 2009. Die Rolle von Unterschieden in den Bildungswegen und in der Zusammensetzung der Schülerschaft. *Zeitschrift für Pädagogik, 59* [Beiheft], 132–150.

Fox, J.-P. & Verhagen, A. J. (2010). Random item effects modeling for cross-national survey data. In E. Davidov, P. & Schmidt & J. Billiet (Eds.), *Cross-cultural analysis: Methods and applications* (pp. 461–482). London: Routeledge Academic.

Gebhardt, E. & Adams, R. J. (2007). The influence of equating methodology on reported trends in PISA. *Journal of Applied Measurement*, *8*, 305–322.

Goldhammer, F., Naumann, J., Rölke, H., Stelter, A. & Tóth, K. (in press). Relating product data to process data from computer-based competence assessment. In D. Leutner, J.

Fleischer, J. Grünkorn & E. Klieme (Eds.), *Competence assessment in education: Research, models and instruments.* Heidelberg: Springer.

Haberman, S. J. (2009). *Linking parameter estimates derived from an item response model through separate calibrations*. ETS Research Report ETS RR-09-40. Princeton, ETS.

Heine, J.-H., Mang, J., Borchert, L., Gomolka, J., Kröhne, U., Goldhammer, F. & Sälzer, C. (2016). Kompetenzmessung in PISA 2015. In K. Reiss, C. Sälzer, A. Schiepe-Tiska & O. Köller (Hrsg.). *PISA 2015: Eine Studie in Kontinuität und Wandel* (S. xxx–xxx). Münster: Waxmann.

Husek, T. R. & Sirotnik, K. (1967). *Item sampling in educational research* (CSEIP Occasional Report No. 2). Los Angeles: University of California.

Educational Testing Service (ETS) (2015). *PISA 2015 Field trial analysis report: Outcomes of the cognitive assessment* [interner Bericht]. Princeton, NJ.

Kiefer, T., Robitzsch, A. & Wu, M. (2016). *TAM: Test analysis modules*. R package version 1.995-0. Verfügbar unter http://CRAN.R-project.org/package=TAM

Kingston, N. M. (2009). Comparability of computer- and paper-administered multiple-choice tests for K–12 populations: A synthesis. *Applied Measurement in Education, 22*, 22–37.

Köller, O., Knigge, M. & Tesch, B. (Hrsg.) (2010). *Sprachliche Kompetenzen im Ländervergleich*. Münster: Waxmann.

Klieme, E. & Beck, B. (2007). *Sprachliche Kompetenzen. Konzepte und Messung. DESI-Studie (Deutsch Englisch Schülerleistungen International).* Weinheim: Beltz

Klieme, E., Jude, N., Baumert, J. & Prenzel, M., (2010). PISA 2000–2009: Bilanz der Veränderungen im Schulsystem. E. In Klieme, C. Artelt, J. Hartig, N. Jude, O. Köller, M. Prenzel, W. Schneider & P. Stanat (Hrsg.), *PISA 2009. Bilanz nach einem Jahrzehnt* (S. 277–300). Münster: Waxmann.

Kolen, M. J. & Brennan, R. L. (2014). *Test equating, scaling, and linking*. New York: Springer.

Kröhne, U. & Martens, T. (2011). Computer-based competence tests in the national educational panel study: The challenge of mode effects. *Zeitschrift für Erziehungswissenschaft, 14* (2), 169–186.

Macaskill, G. (2008, September). Alternative scaling models and dependencies. TAG(0809)6a Report. Paper presented at the 2008 TAG Meeting in Sydney, Australia. Retrieved October 28, 2016, from https://www.acer.edu.au/files/macaskill_alternativescalingmodelsdependenciespisa(1).pdf

Mazzeo, J. & von Davier, M. (2008). Review of the Programme for International Student Assessment (PISA) test design: Recommendations for fostering stability in assessment results. *Education Working Papers EDU/PISA/GB (2008)*, *28*, 23–24.

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex surveys. *Psychometrika*, *56*, 177–196.

Monseur, C. & Berezner, A. (2007). The computation of equating errors in international surveys in education. *Journal of Applied Measurement, 8*, 323–335.

Muthén, B. & Asparouhov, T. (2014). IRT studies of many groups: The alignment method. *Frontiers in Psychology | Quantitative Psychology and Measurement, 5*, 978.

National Center for Education Statistics (NCES) (2013). *The nation's report card: NAEP 2012. Trends in academic progress*. Washington, D.C: Institute of Education Science, U.S. Department of Education.

Organisation for Economic Co-operation and Development (OECD) (2009). *PISA 2006 technical report*. Paris: OECD Publishing.

Organisation for Economic Co-operation and Development (OECD) (2013a). *Technical report of the survey of adult skills (PIAAC)*. Paris: OECD Publishing.

Organisation for Economic Co-operation and Development (OECD) (2013b). *The survey of adult skills: Reader's companion*. Paris: OECD Publishing.

Oliveri, M. E. & von Davier, M. (2011). Investigation of model fit and score scale comparability in international assessments. *Psychological Test and Assessment Modeling*, *53*, 315–333.

Oliveri, M. E. & von Davier, M. (2014). Toward increasing fairness in score scale calibrations employed in international large-scale assessments. *International Journal of Testing*, *14*, 1–21.

Parshall, C. G., Harmes, J. C., Davey T., & Pashley, P. J. (2010). Innovative item types for computerized testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 215–230). New York, NY: Springer.

Pohl, S. & Carstensen, C. H. (2012): *NEPS technical report – Scaling the data of the competence tests* (NEPS Working Paper No. 14). Bamberg: Otto-Friedrich-Universität, Nationales Bildungspanel.

Prenzel, M., Sälzer, C., Klieme, E. & Köller, O. (Hrsg.). (2013). *PISA 2012: Fortschritte und Herausforderungen in Deutschland*. Münster: Waxmann.

R Core Team (2016). *R: A language and environment for statistical computing*. Vienna, Austria. Verfügbar unter https://www.R-project.org/

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

Rammstedt, B. (Hrsg.). (2013). *Grundlegende Kompetenzen Erwachsener im internationalen Vergleich*. *Ergebnisse von PIAAC 2012*. Münster: Waxmann.

Reiss, K., Sälzer, C., Schiepe-Tiska, A. & Köller, O. (Hrsg.). (2016). *PISA 2015: Eine Studie in Kontinuität und Wandel*. Münster: Waxmann.

Robitzsch, A. (2016). *Essays zu methodischen Herausforderungen im Large-Scale Assessment.* Dissertation, Humboldt-Universität zu Berlin. Zugriff am 28.10.2016 unter edoc.hu-berlin.de/dissertationen/robitzsch-alexander-2015-10-27/PDF/robitzsch.pdf

Robitzsch, A. (2016). *sirt: Supplementary item response theory models*. R package version 1.12-2. Verfügbar unter http://CRAN.R-project.org/package=sirt

Rost, J. (2004). *Lehrbuch Testtheorie – Testkonstruktion*. Bern: Huber.

Sachse, K. A., Roppelt, A. & Haag, N. (2016). A comparison of linking methods for estimating national trends in international comparative large-scale assessments in the presence of cross-national DIF. *Journal of Educational Measurement, 53* (2)*,* 152–171.

van der Linden, W. J. (2005). *Linear models for optimal test design*. New York: Springer.

Wang, S., Jiao, H., Young, M. J., Brooks, T., & Olson, J. (2008). Comparability of computer-based and paper-and-pencil testing in K–12 reading assessments: A meta-analysis of testing mode effects. *Educational and Psychological Measurement, 68*, 5–24.

Wendt, H., Bos, W., Selter, C., Köller, O., Schwippert, K. & Kasper, D. (Hrsg.). (2016). *TIMSS 2015: Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich*. Münster: Waxmann.

Wu, M. (2010). Measurement, sampling, and equating errors in large-scale assessments. *Educational Measurement: Issues and Practice*, *29* (4), 15–27.

Xu, X. & von Davier, M. (2010). *Linking errors in trend estimation in large-scale surveys: A case study* (ETS Research Report RR10-10). Princeton: ETS.

Yamamoto, K. (2012). *Outgrowing the mode effect study of paper and computer based testing.* Zugriff am 21. Oktober 2016. Verfügbar unter http://www.umdcipe.org/conferences/EducationEvaluationItaly/COMPLETE_PAPERS/Yamamoto/YAMAMOTO.pdf

Tabelle 1. Ergebnisse der Feldtest-Studie 2014 zu Effekten des Testmodus aus dem 1PL-Modell

Bereich	*N*		Alle Items					Invariante Items (lt. OECD)				
				D		SD_{Modus}			*d*		SD_{Modus}	
	PBA	CBA	*I*	*Est*	*SE*	*Est*	*SE*	*I*	*Est*	*SE*	*Est*	*SE*
Naturwissenschaften	340	338	77	**−.23**	.08	.17	.05	56	**−.17**	.08	.13	.06
Mathematik	345	340	66	**−.14**	.07	.31	.05	36	−.09	.07	.27	.08
Lesen	349	334	82	−.13	.10	.43	.05	47	−.06	.09	.25	.07

Anmerkungen. OECD = Organisation for Economic Co-operation and Development, *N* = Anzahl der Schülerinnen und Schüler, *I* = Anzahl der Items für den Vergleich, PBA = papierbasierte Administration, CBA = computerbasierte Administration, Modus = Administrationsmodus, *d* = Effektgröße für Moduseffekt CBA vs. PBA (negativer Effekt indiziert Nachteil von CBA), *Est* = Schätzung, *SE* = Standardfehler, SD_{Modus} = Standardabweichung der Differenz der Itemschwierigkeiten. Fett dargestellte *d*-Werte sind statistisch signifikant ($p < .05$).

Tabelle 2. Stichproben der verwendeten PISA-Studien für die Linking-Studie

Bereich	Studie	Modus	*N*	# Items	# Linkitems	
					alle	invariant lt. OECD
Naturwissen-schaften	2006	PBA	4881	103	77	56
	2009	PBA	3477	53	52	40
	2012	PBA	3505	52	52	40
	2014	PBA	340	91	77	56
	2014	CBA	338	91	77	56
	2015	CBA	6501	181	77	56
Mathematik	2003	PBA	4656	84	31	16
	2006	PBA	3795	48	31	16
	2009	PBA	3503	35	31	16
	2012	PBA	4971	84	66	36
	2014	PBA	345	70	66	36
	2014	CBA	340	68	66	36
	2015	CBA	2739	69	66	36
Lesen	2000	PBA	5060	128	35	19
	2003	PBA	2555	27	24	13
	2006	PBA	2701	28	24	13
	2009	PBA	4975	100	82	47
	2012	PBA	3470	43	42	23
	2014	PBA	349	85	82	47
	2014	CBA	334	85	82	47
	2015	CBA	2746	87	82	47

Anmerkungen. PISA = Programme for International Student Assessment, OECD = Organisation for Economic Co-operation and Development, Studie = verwendete PISA-Studie. „2014“ bezeichnet den PISA-Feldtest für PISA 2015. Modus = Administrationsmodus, PBA = papierbasiert, CBA = computerbasiert, *N* = Stichprobenumfang, # Items = Anzahl der verwendeten Items in Skalierung, # Link-Items = Anzahl der verwendeten Items in gemeinsamem Linking. Dabei kennzeichnet die Spalte „invariant“ die Anzahl der laut OECD ausgezeichneten invarianten Items.

Tabelle 3. Übersicht über die verwendeten Linkitems mit mittleren Itemschwierigkeiten aus dem 1PL-Modell in den verschiedenen PISA-Studien

Bereich	Item-gruppe	# Items	2000 PBA	2003 PBA	2006 PBA	2009 PBA	2012 PBA	2015 CBA	2014 PBA vs. CBA
Naturwissen-schaften	S2A	52	—	—	−0.40	−0.41	−0.49	−0.34	−0.24
	S2B	25	—	—	−0.29	—	—	−0.13	−0.23
Mathematik	M1A	31	—	0.01	−0.11	−0.14	−0.18	−0.07	−0.10
	M1B	35	—	—	—	—	−0.08	0.17	−0.25
Lesen	R1A	24	−0.34	−0.32	−0.41	−0.62	—	−0.53	−0.34
	R1B	11	−0.97	—	—	−1.09	—	−1.10	−0.37
	R2A	39	—	—	—	−0.57	−0.66	−0.56	−0.07
	R2B	8	—	—	—	0.00	—	−0.18	0.12

Anmerkungen. PISA = Programme for International Student Assessment, Itemgruppe = Bezeichnung der Gruppe der Linkitems, die in verschiedenen PISA-Studien auftritt. # Items = Anzahl der Items in Itemgruppe, PBA = papierbasierte Administration, CBA = computerbasierte Administration, 2014 PBA vs. CBA = Differenz in mittleren Itemschwierigkeiten zwischen den Administrationsmodi PBA und CBA im PISA Feldtest 2014. Negative Werte kennzeichnen, dass Items einer Itemgruppe im PBA-Modus leichter als im CBA-Modus sind.

Tabelle 4. Trendschätzung in Naturwissenschaften (Mittelwerte für Deutschland sowie Trendschätzung von 2012 nach 2015)

	Methode	2006	2009	2012	2015	Trend 2012 → 2015			
						Est	SE_{tot}	SE_p	SE_i
Original		516	520	524	509	−15	5.6	4.0	3.9
ohne Feldtest (alle Items)	C1	516	519	523	508	−15	6.5	3.7	5.3
	C2	516	518	524	508	−16	6.1	3.6	4.9
	H1	516	515	522	506	−16	6.3	3.7	5.2
	H2	516	516	522	501	−21	6.6	3.6	5.5
	S1	516	517	524	511	−13	6.4	3.7	5.2
ohne Feldtest (invariante Items)	C1I	516	519	523	513	−10	6.8	3.8	5.6
	C2I	516	519	524	513	−11	6.7	3.6	5.6
	H1I	516	517	523	513	−10	6.8	3.7	5.7
	H2I	516	518	524	506	−18	6.4	3.6	5.3
mit Feldtest	C1F	516	520	524	528	4	8.0	3.7	7.0
	H1F	516	516	522	528	6	8.1	3.8	7.2
	S1F	516	517	524	531	7	8.3	3.8	7.4

Anmerkungen. *Est* = Schätzung, SE_{tot} = Standardfehler aufgrund Variabilität von Personen und Items, SE_p = Standardfehler aufgrund Variabilität von Personen, SE_i = Standardfehler aufgrund Variabilität von Items (Linkfehler).

Tabelle 5. Trendschätzung in Mathematik (Mittelwerte für Deutschland sowie Trendschätzung von 2012 nach 2015)

	Methode	2003	2006	2009	2012	2015	Trend 2012 → 2015			
							Est	SE_{tot}	SE_p	SE_i
Original		503	504	513	514	506	−8	5.4	4.1	3.5
ohne Feldtest (alle Items)	C1	503	513	515	522	510	−12	5.7	3.7	4.4
	C2	503	515	517	524	511	−13	4.8	3.5	3.3
	H1	503	512	514	521	505	−16	5.5	3.6	4.2
	H2	503	517	521	528	515	−13	5.6	3.5	4.5
	S1	503	512	514	518	503	−15	5.7	3.6	4.4
ohne Feldtest (invariante Items)	C1I	503	513	515	522	521	−1	8.5	3.7	7.7
	C2I	503	515	517	524	522	−2	7.2	3.5	6.3
	H1I	503	512	514	521	512	−9	6.0	3.6	4.8
	H2I	503	516	521	528	524	−4	6.6	3.5	5.6
mit Feldtest	C1F	503	512	515	516	518	2	6.0	3.6	4.8
	H1F	503	512	514	514	515	1	7.2	3.6	6.2
	S1F	503	512	514	518	517	−1	7.1	3.6	6.1

Anmerkungen. *Est* = Schätzung, SE_{tot} = Standardfehler aufgrund Variabilität von Personen und Items, SE_p = Standardfehler aufgrund Variabilität von Personen, SE_i = Standardfehler aufgrund Variabilität von Items (Linkfehler).

Tabelle 6. Trendschätzung in Lesen (Mittelwerte für Deutschland sowie Trendschätzung von 2012 nach 2015)

	Methode	2000	2003	2006	2009	2012	2015	Trend 2012 → 2015			
								Est	SE_{tot}	SE_p	SE_i
Original		484	491	495	497	508	509	1	6.7	4.1	5.3
ohne Feldtest (alle Items)	C1	484	479	488	504	510	504	−6	7.5	4.6	6.0
	C2	484	484	492	502	506	501	−5	8.4	4.9	6.8
	H1	484	478	487	501	509	502	−7	6.9	4.7	5.0
	H2	484	473	478	491	504	495	−9	10.3	4.9	9.0
	S1	484	482	490	510	516	505	−11	8.2	4.7	6.8
ohne Feldtest (invariante Items)	C1I	484	478	488	507	513	523	10	9.5	4.7	8.3
	C2I	484	483	491	506	511	521	10	10.7	4.9	9.4
	H1I	484	477	486	504	512	517	5	7.6	4.7	5.9
	H2I	484	473	477	496	508	510	2	11.4	5.0	10.2
mit Feldtest	C1F	484	480	489	499	501	512	11	9.5	4.7	8.3
	H1F	484	479	488	499	505	516	11	9.6	5.0	8.2
	S1F	484	482	490	510	516	528	12	9.2	5.0	7.7

Anmerkungen. *Est* = Schätzung, SE_{tot} = Standardfehler aufgrund Variabilität von Personen und Items, SE_p = Standardfehler aufgrund Variabilität von Personen, SE_i = Standardfehler aufgrund Variabilität von Items (Linkfehler).

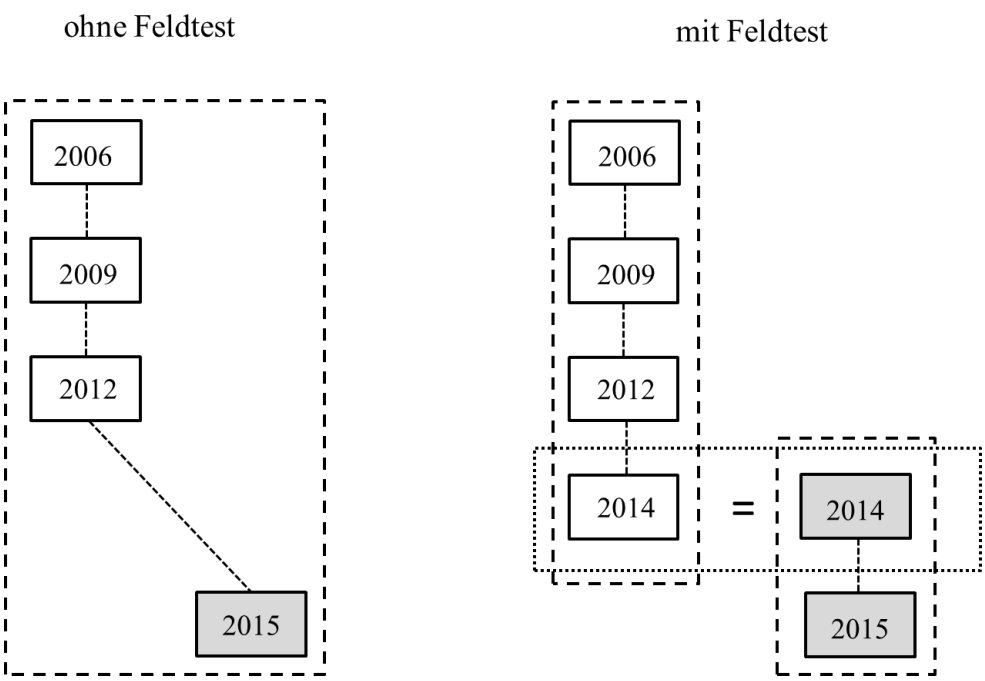

Abbildung 1. Schematische Darstellung der marginalen Trendschätzung für Naturwissenschaften ohne Berücksichtigung (linke Grafik) und mit Berücksichtigung (rechte Grafik) der Daten der deutschen Feldteststudie von 2014

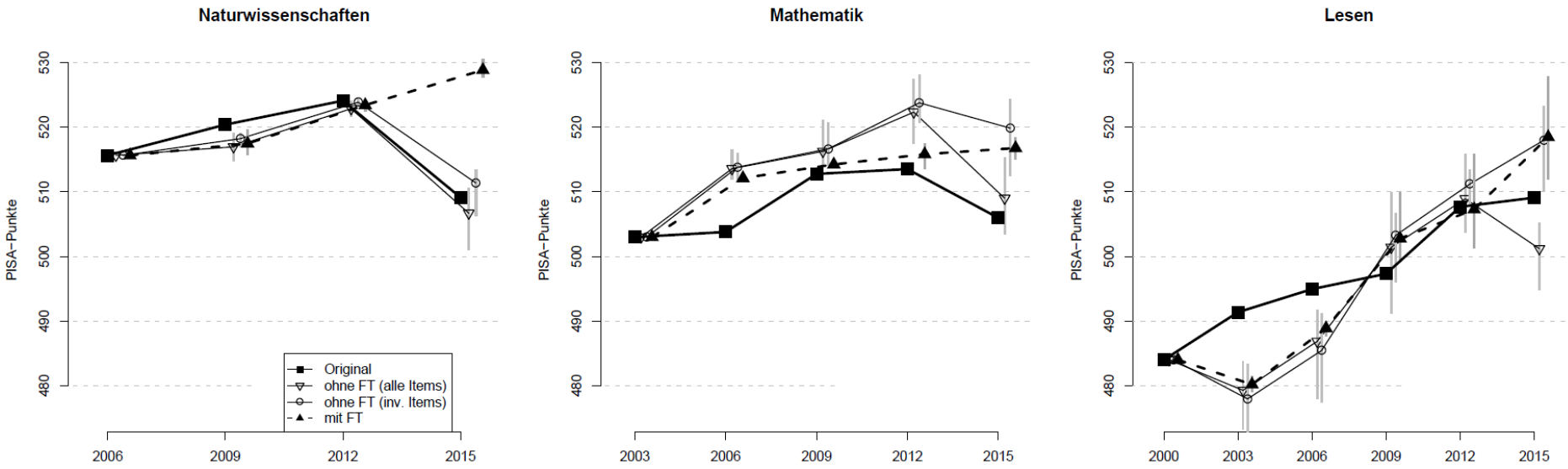

Abbildung 2. Originale und verschiedene Strategien marginaler Trendschätzungen in den Naturwissenschaften, Mathematik und Lesen. Für die marginalen Strategien wird der Mittelwert aller Methoden dargestellt (z. B. bei mit FT: Mittelwert über C1F, H1F und S1F). Das Minimum und Maximum wird durch die vertikalen grauen Balken angegeben. FT = Feldtest, inv. = invariant