Die Abbildung latenter Merkmale im individuellen Entwicklungsverlauf: Vergleich von IRT-Linkmethoden unter Verwendung Rasch-skalierter Kompetenztestdaten

Inaugural-Dissertation

in der Fakultät Humanwissenschaften

der Otto-Friedrich-Universität Bamberg

vorgelegt von

Luise Fischer

aus Berlin

Falun, den 28.02.2023

Bamberg, 2023

Tag der mündlichen Prüfung: 20.07.2023

Dekan: Universitätsprofessor Dr. Claus-Christian Carbon

Betreuer: Universitätsprofessor Dr. Claus H. Carstensen

Weiterer Gutachter: Dr. Timo Gnambs

URN: urn:nbn:de:bvb:473-irb-910148
DOI: https://doi.org/10.20378/irb-91014

Inhaltsverzeichnis

Einleitende Worte

Als empirische Wissenschaft strebt die Psychologie danach, das Erleben und Verhalten des Menschen in jeder Lebensphase zu beschreiben und zu erklären. Hernach versuchen Psychologinnen und Psychologen Maße zu finden, um die gesamte Lebensspanne in ihrer vollen Entwicklungsbreite abzubilden. Intra- und interindividuelle Entwicklungsverläufe sowie relevante Einflussfaktoren sollen identifiziert und quantifiziert und somit vergleichbar gemacht werden.

Mit diesen ambitionierten Zielen gehen viele und große methodische sowie statistische Herausforderungen einher. Einem kleinen Teil dieser Herausforderungen möchte sich die vorliegende Forschungsarbeit widmen, um die Psychologie in ihrem Bestreben zu unterstützen und den aktuellen Forschungsstand weiter voranzutreiben.

Der Aufbau der vorliegenden Synopse ist wie folgt: Nach einer Begriffsklärung und Einführung in das Thema dieser Qualifikationsarbeit, werden der aktuelle Forschungsstand und die sich daraus ergebenden Fragestellungen dargelegt. Anschließend werden kurz die Inhalte der drei verfassten Beiträge beschrieben und ihr wissenschaftlicher Beitrag diskutiert. Abschließend wird ein Ausblick zu weiterführenden Forschungsfragen gegeben.

Begriffsklärung und Einführung in das Thema

Um das gewählte Forschungsfeld näher zu umreißen, sei der Autorin zunächst ein kleiner Ausflug in die unterschiedlichen psychologischen Skalenniveaus gestattet. Manchen demographischen Merkmalen, welche in der Psychologie häufig als sogenannte Kontextfaktoren erhoben werden, liegt eine Rationalskala zu Grunde. Die Skalen von Merkmalen mit Rationalskalenniveau verfügen über eindeutig definierte Abstände zwischen einzelnen Werten

sowie einen natürlichen Nullpunkt. Als Beispiele hierfür seien Alter, Körpergröße oder Einkommen genannt. Auch abstrakte, also nicht direkt beobachtbare, psychologische Konstrukte (wie z.B. Intelligenz, Kreativität oder Optimismus), sind zumeist metrisch konzeptualisiert und die Psychologie als empirische Wissenschaft strebt danach, die individuelle Ausprägung eines solch abstrakten Konstruktes möglichst genau zu „messen“. Die hierzu üblicherweise eingesetzten Messinstrumente sind Fragebögen und Testverfahren, deren Items als kleinste Bausteine jedoch häufig nur Ordinalskalenniveau haben – und somit weder über klar definierte Abstände zwischen einzelnen Werten, noch über einen klar definierten Nullpunkt verfügen. Sowohl die klassische, als auch die probabilistische Testtheorie initiieren bei ihrer Anwendung ein Intervallskalenniveau (mit klar definierten Abständen zwischen einzelnen Werten). Die Problematik des fehlenden natürlichen Nullpunktes bleibt jedoch bestehen, sodass ein solcher für jede Datenmenge willkürlich festgesetzt wird. Daten, welche von verschiedenen Messzeitpunkten stammen oder von Gruppen, die sich hinsichtlich des interessierenden Merkmals unterscheiden, sind infolgedessen nicht direkt vergleichbar, da die Nullpunkte der ihnen zu Grunde liegenden Skalen verschieden definiert sein können. Als Beispiel diene hier die Zeitverschiebung: Ein Tag auf der Erde hat 24 Stunden, unabhängig davon, ob jemand diesen Tag in Bamberg, Peking oder New York verbringt. Während man jedoch den Arbeitstag um neun Uhr morgens in Bamberg gerade begonnen hat, denkt zeitgleich eine Person in Peking um vier Uhr nachmittags bereits an den Feierabend, während sich eine andere Person in New York um drei Uhr nachts noch in tiefem Schlaf befindet. Um also mehrere Datenmengen mit jeweils willkürlich bestimmten Nullpunkten zu vergleichen, müssen ihre jeweiligen Definitionen des Nullpunktes zuerst übereingestimmt werden. Im Falle unseres Beispiels, in dem Bamberg als Referenzgruppe dienen soll, müsste die Zeit bei Personen in Peking also sieben Stunden

zurückgestellt werden, während sie bei Personen in New York um sechs Stunden vorgestellt werden müsste. Erst dann kann man einen vergleichbaren Eindruck erhalten, wie Personen ihren Tag um neun Uhr morgens an verschiedenen Orten der Erde verbringen. Der Prozess des Übereinstimmens willkürlich festgesetzter Nullpunkte ist in der Literatur auch als „Verlinken" (engl.: „linking", Kolen & Brennan, 2014) bekannt. Unabhängig vom erfassten psychologischen Merkmal muss gewährleistet sein, dass ebenjenes Merkmal über die Zeit inhaltlich unverändert bleibt, beziehungsweise sich zwischen mehreren Gruppen nicht unterscheidet (siehe auch die Voraussetzungen zum Verlinken weiter unten). Sind diese Voraussetzungen nicht erfüllt, so wären vergleichende Aussagen inhaltlich sinnlos, da die sprichwörtlichen Äpfel und Birnen verglichen würden. Um nochmals auf obiges Beispiel zurückzukommen: Die zuvor angeführten Zeitkorrekturen gelten in der Winterzeit. Wäre gerade Sommerzeit, so müsste die Zeit für Personen in Peking statt sieben nur um sechs Stunden zurückgestellt werden, während die Zeitkorrektur (in Bezug auf die Referenzgruppe Bamberg) für Personen in New York von der Sommer- und Winterzeit unabhängig ist.

Notwendigkeit des Verlinkens

Die Thematik des Verlinkens erhobener Kompetenztestdaten (im Folgenden kurz „Daten" genannt) nimmt einen zentralen Stellenwert in der Bildungsforschung ein: Kohortenvergleiche werden angestellt, um die Auswirkungen von (politischen) Maßnahmen und gesellschaftlichen Veränderungen im Bildungswesen sichtbar zu machen. Die größte internationale Untersuchung zum Bildungsstand 15-jähriger Schülerinnen und Schüler stellt PISA (die Internationale Schulleistungsstudie der OECD) dar. So wurden beispielsweise im PISA-Erhebungszyklus von 2018 600 000 Personen befragt. Alle drei Jahre wird die Befragung

an einer neuen Kohorte wiederholt, um somit Bildungstrends sichtbar zu machen und Vergleiche zwischen Kohorten und Ländern zu ermöglichen. Anders ausgedrückt, wird die Skala der hinzugekommenen Daten an die Skala der bereits erfassten Daten angepasst. Das Verknüpfen von Daten stellt hier also eine Art der Einordnung dar; wobei bereits erfasste Daten den Referenzrahmen darstellen und neue Daten durch den Prozess der Datenverknüpfung relativ zu der bereits vorhandenen Referenzstichprobe eingeordnet werden können.

Die Notwendigkeit, die PISA-Daten unterschiedlicher Erhebungszyklen miteinander zu Verlinken ist klar gegeben: zum einen sind Unterschiede der Verteilungen der zu erfassenden Merkmale zwischen den Kohorten sehr wahrscheinlich; zum anderen aber unterscheiden sich die kohortenspezifischen Testformen der Erhebungszyklen in ihren Aufgaben voneinander. Jedoch lassen sich die im Rahmen der PISA-Studie (sowie vergleichbaren anderen Large Scale Assessments (LSAs), wie z.B., NAEP, TIMMS, PIRLS) gut erprobte Methoden des Datenverlinkens nicht ungeprüft in Kontexte mit anderen Voraussetzungen übertragen – wie zum Beispiel Studien mit Längsschnittdesign, im Rahmen derer individuelle Entwicklungsverläufe untersucht werden sollen. Obwohl der Prozess des Verlinkens statistisch prinzipiell unabhängig ist a) vom Inhalt des erfassten Merkmals, sowie b) vom Stichprobendesign (Längsschnitt oder Querschnitt (z.B. PISA)), liegt der Fokus der vorliegenden Forschungsarbeit dennoch bewusst auf der Erfassung kognitiver Entwicklungsverläufe. Dies ist durch die besonderen Anforderungen an die zugehörigen Erhebungsinstrumente begründet. Während beispielsweise Fragebögen zu Persönlichkeitsmerkmalen häufig in unveränderter Form wiederholt eingesetzt werden (können), so wäre dies bei kognitiven Tests zumeist nicht möglich. Die Gründe hierfür sind mannigfaltig. Sind Veränderungen der kognitiven Leistungsfähigkeit über die Zeit zu

erwarten, so sollte die Testschwierigkeit stets dem durchschnittlichen aktuellen Fähigkeitsniveau angepasst sein (im Weiteren wird hierfür der englische Begriff „test targeting“ verwendet), um eine hohe Messgenauigkeit zu garantieren und die Motivation der Probanden aufrecht zu erhalten. Weitere Aspekte betreffen Erinnerungseffekte und eine eventuell stark veränderliche Testausdauer der Probanden. Insofern kommen bei der Erfassung kognitiver Entwicklungsverläufe verschiedene Testformen zum Einsatz, die sich sowohl in ihrer Aufgabenzusammensetzung als auch ihrer Testlänge unterscheiden können. Ein Vergleich der erreichten Summenwerte, welche von verschiedenen Messzeitpunkten stammen, wäre deshalb nicht direkt als Maß der Veränderungsmessung des erhobenen Merkmals interpretierbar.

Im Gegensatz zu Kohortenvergleichen in LSAs gibt es keine belastbaren Ergebnisse zum Verlinken angrenzender Messzeitpunkte bei Studien mit begrenzter Itemzahl und potentiell großen Fähigkeitsveränderungen. Infolgedessen fehlen Leitlinien und praktisch orientierte Anleitungen.

Item Response Theory

Beide in der Psychologie üblichen Testtheorien (klassische [Lienert, 1998] sowie probabilistische Testtheorie [de Ayala, 2022]) bieten Möglichkeiten, um Datenmengen, welche sich hinsichtlich der Verteilung des interessierenden Merkmals unterscheiden, miteinander zu verknüpfen. Die vorliegende Forschungsarbeit beschäftigt sich mit dem Verlinken im Kontext der probabilistischen Testtheorie, welche im englischen Sprachraum als item response theory (IRT) bezeichnet wird. Mittels eines IRT-Modells werden Antworten einer Person auf Items mit dem zu Grunde liegenden latenten Konstrukt in Beziehung gesetzt, indem Personenfähigkeit und Itemschwierigkeit auf derselben Skala (logit Skala) verortet werden. Dieser auch als Skalieren

bekannte Schätzprozess (häufig mittels Maximum Likelihood Methode) mündet in Personen- und Itemparameter. Je nach inhaltlicher Konzeption eines latenten Merkmals sind die am häufigsten verwendeten Messmodelle (Kim et al., 2020) das ein-, zwei- oder drei-parameter logistische Modell (1PLM [Rasch, 1980], 2PLM und 3PLM [Birnbaum, 1968]). Dabei resultieren zumeist ein personenspezifischer Parameter und jeweils ein bis drei itemspezifische Parameter. Von den drei genannten Modellen, ist das 3PLM das globalste, in welchem die Wahrscheinlichkeit p_{ij} von Person i ein Item j zu lösen definiert ist als

$$p_{ij}\left(X = 1 \middle| \theta_i, a_j, b_j, c_j\right) = c_j + \left(1 - c_j\right)\frac{\exp\left[a_j\left(\theta_i - b_j\right)\right]}{1 + \exp\left[a_j\left(\theta_i - b_j\right)\right]}, \quad (1)$$

wobei „exp“ die natürliche Exponentialfunktion darstellt. In Gleichung (1) ist Θ als Personenfähigkeit definiert, während a, b und c Itemcharakteristika beschreiben. Der Nenner standardisiert Gleichung (1), sodass eine Lösungswahrscheinlichkeit p_{ij} zwischen 0 und 1 resultiert. Die Funktionsgraphen der Items (sog. „item characteristic curves“ (ICCs), s. Abb. 1) veranschaulichen die Lösungswahrscheinlichkeit in Abhängigkeit der Personenfähigkeit bei vorliegenden Itemparametern a, b und c.

Der Fähigkeitsparameter Θ_i für Person i liegt zwischen $-\infty < \Theta < \infty$ und wird oftmals einer Normalverteilung folgend skaliert, mit einem Mittelwert von 0 und einer Stadardabweichung von 1 (Kolen & Brennan, 2014). In diesem Fall liegen die Werte fast aller Personen zwischen -3 und 3 logits. Die Parameter a_j, b_j und c_j kennzeichnen item j. Der Itemschwierigkeitsparameter b_j verortet ein Item auf der logit Skala und liegt häufig ebenfalls im Bereich zwischen -3 und 3. Er markiert den Wendepunkt der ICC auf der x-Achse (s. Abb. 1), und markiert somit die Stelle der größten Steigung, welche sich bei $b_j = \Theta$ findet. Genau diese

Steigung beschreibt der Itemdiskriminationsparameter a_j. Im Falle von $a_j = 0$ wäre die ICC eine horizontale Linie, weshalb alle Personen – unabhängig von ihrer Fähigkeit – dieselbe Lösungswahrscheinlichkeit hätten. Die Diskrimination liegt häufig im Bereich 0 bis 2.5. Der Parameter c_j liegt zwischen 0 und 1 und repräsentiert die Ratewahrscheinlichkeit, wie sie durch multiple choice Aufgaben gegeben ist. Wird der Rateparameter c_j auf 0 fixiert, so vereinfacht sich das 3PLM zum 2PLM:

$$p_{ij}(X = 1|\theta_i, a_j, b_j) = \frac{\exp[a_j(\theta_i - b_j)]}{1 + \exp[a_j(\theta_i - b_j)]}. \quad (2)$$

Weiteres fixieren des Diskriminationsparameters a_j (üblicherweise auf den Wert 1), mündet in das 1PLM, auch bekannt als Rasch Modell:

$$p_{ij}(X = 1|\theta_i, b_j) = \frac{\exp(\theta_i - b_j)}{1 + \exp(\theta_i - b_j)}. \quad (3)$$

Im Rasch Modell erfolgt – im Gegensatz zu 2PLM und 3PLM – die Schwierigkeitsschätzung bei Geltung des Modells unabhängig von der zu Grunde liegenden Stichprobe (Fischer & Molenaar, 2012). Der Vergleich zweier Itemschwierigkeiten würde also immer zu demselben Ergebnis führen, unabhängig davon, welche Personen es bearbeitet haben.

Je komplexer das Modell (d. h. je größer die Anzahl der zu schätzenden Itemparameter), umso größer sollte die verwendete Stichprobe sein, um stabile (d. h. möglichst fehlerfrei geschätzte) Itemparameterwerte zu erhalten. Die Genauigkeit der Item- und Personenparameterschätzung wird neben Stichprobengröße und Itemzahl vom test targeting, also der Überlappung der Verteilungen von Stichprobenfähigkeit und Itemschwierigkeit, beeinflusst. Dabei wird die Schätzung der Schwierigkeitsparameter durch das test targeting mehr beeinflusst

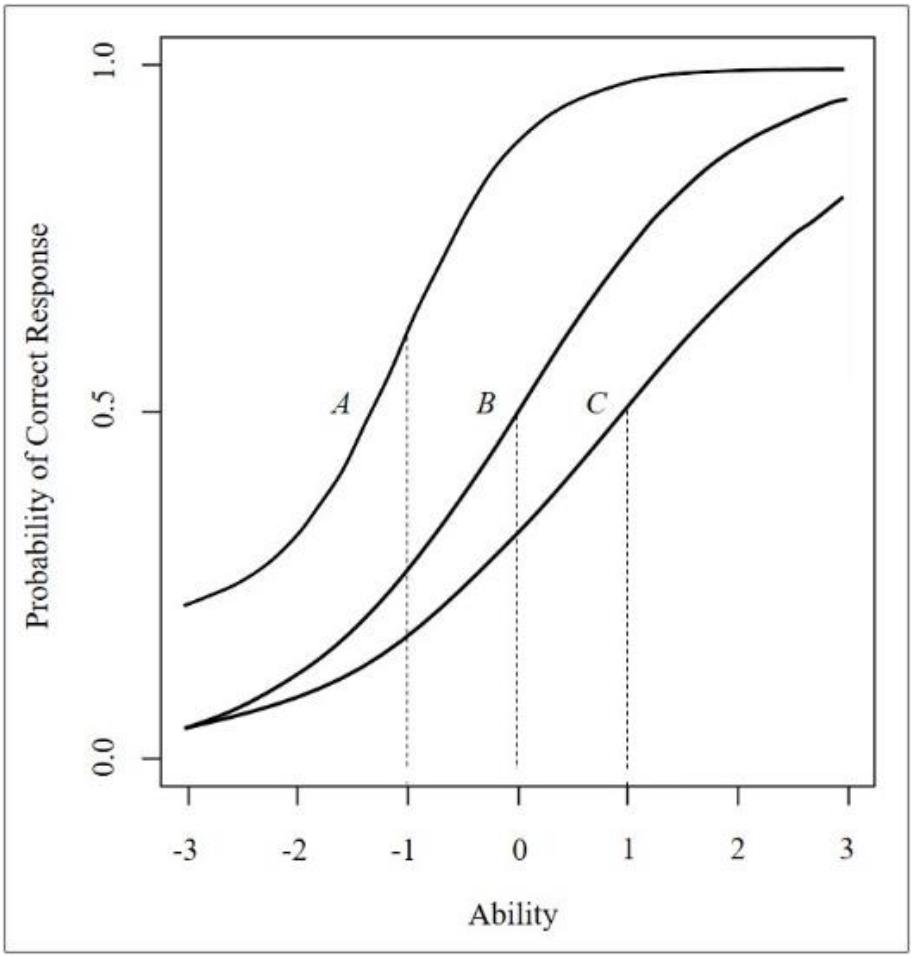
als die der Personenparameter (Svetina et al., 2013). Je genauer die Itemparameterschätzung ist, desto stärker stimmen die aus dem Linkprozess resultierende empirische Veränderung und die wahre Veränderung des latenten Merkmals überein. Jedoch: Nicht allein bloße Stichprobengröße definiert die Auswahl des passenden Messmodells – auch inhaltliche Aspekte spielen bei der Modellwahl eine Rolle. Fest steht: Je weniger Itemparameter geschätzt werden, umso weniger Stellschrauben stehen zur Verfügung, um die zu Grunde liegenden Daten im Rahmen eines Modells passgenau zu beschreiben.

Im Kontext der IRT ist die Modellpassung oft eine graduelle Entscheidung (Meijer & Tendeiro, 2015): Ein Modell passt mehr oder weniger; kaum aber kann ein Modell die Daten perfekt beschreiben. Das Dogma der Modellpassung scheint in der Praxis so stark, dass umgangssprachlich sogar die Richtung der Passung umgekehrt wird: Die Daten würden (nicht) zum Modell passen. Eine ungünstige Modellpassung zeigt sich unter anderem in Abweichungen der geschätzten Itemparameter von ihren wahren Werten. Trügerischerweise können diese Abweichungen bei empirischen Daten jedoch kaum identifiziert werden. So würde die Anwendung des 1PLM auf nicht eindimensionale Daten dazu führen, dass das Ignorieren der sich unterscheidenden Diskriminationsparameter sich in einer Verschätzung der Itemschwierigkeitsparameter niederschlägt (Humphry, 2018). In der Folge würde die Verwendung fehlerbehafteter Itemparameter im Linkprozess zu einer Abweichung zwischen geschätzter und wahrer Veränderung des latenten Merkmals führen. Häufiger betroffen von dem beschriebenen Dilemma zwischen Stichprobengröße und Modellwahl sind Studien mit begrenzten Ressourcen: Aus methodischer Sicht legt eine geringere Stichprobengröße die Wahl

eines weniger komplexen Modells nahe. Diese Wahl beinhaltet jedoch das Risiko einer schlechteren Passung zwischen den vorliegenden Daten und dem gewählten Modell.

Abbildung 1

Darstellung unterschiedlicher Itemfunktionsgraphen (ICCs), modelliert mit 1PLM, 2PLM und 3PLM



Anmerkung. Die Lösungswahrscheinlichkeiten der Items A, B, und C in Abhängigkeit von der Personenfähigkeit wird modelliert mittels 1PLM ($b_B = 0$), 2PLM ($a_C = 0.75$, $b_C = 1$) und 3PLM ($a_A = 1.75$, $b_A = -1$, $c_A = 0.2$). Die vertikale gestrichelte Linie markiert die Itemschwierigkeit auf der x-Achse (siehe Text für mehr Information).

In einer idealen Welt wäre die Auswahl eines Modells alleinig begründet durch die zu Grunde liegende Fragestellung, denn der Anspruch der Eindimensionalität seitens des Rasch Modells ist eine vornehmlich inhaltliche Forderung, welcher sich aus dem theoretischen Rahmen eines latenten Merkmals begründet. So muss die inhaltliche Bedeutung eines latenten Merkmals über mehrere Messzeitpunkte hinweg unverändert bleiben, um belastbare Aussagen über die Entwicklung ebendiesen latenten Merkmals treffen zu können. An dieser Schnittstelle methodisch und inhaltlich orientierter Disziplinen wird deutlich, dass die zuarbeitende Methodik wichtige Werkzeuge für die Beantwortung inhaltlicher Fragestellungen bereitstellt – methodische Werkzeuge, deren Entwicklung erst durch die Nachfrage von inhaltlich motivierten Forschungsfragen angestoßen worden ist. Als exemplarisches Beispiel sei hier die Fragestellung des Entwicklungsverlaufs der mathematischen Kompetenz vom Eintritt in die Grundschule bis zum Erwerb der mittleren Reife angeführt, wie sie im Rahmen des Nationalen Bildungspanels (NEPS; Blossfeld, 2011, NEPS-Netzwerk, 2022) untersucht wird. Das Verlinken der Messzeitpunkte ist hierbei die Anwendung einer Methode, deren inhaltliche Sinnhaftigkeit jedoch in der Mathematikdidaktik und Entwicklungspsychologie begründet liegt.

Linkdesign

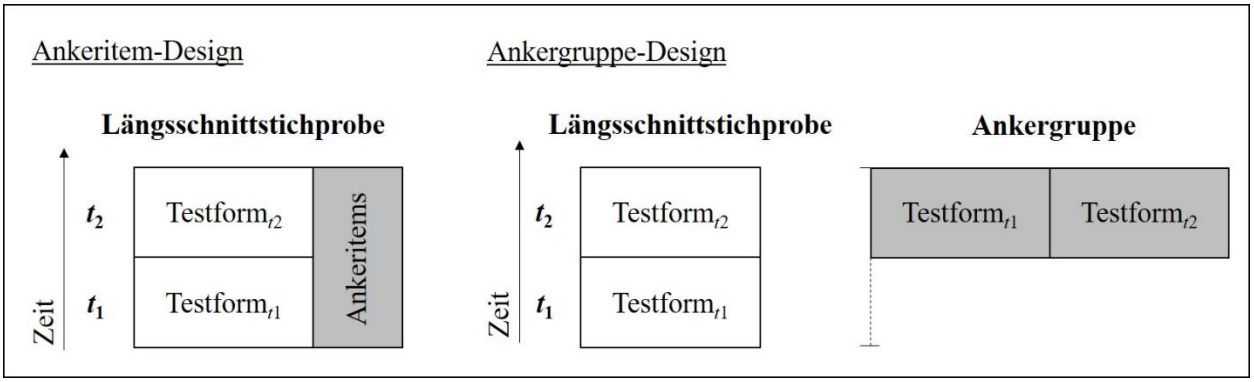
Daten einer mehrfach getesteten Stichprobe, die sich aufgrund von Kompetenzentwicklung in ihrer Fähigkeitsverteilung des interessierenden Merkmals unterschieden, gelten statistisch gesehen als „nonequivalent groups“ (von Davier, Carstensen, & von Davier, 2006) und somit nicht mehr als dieselbe Stichprobe. Um dennoch die Daten mehrerer Messzeitpunkte miteinander zu verknüpfen, müssen im Studiendesign Ankerpunkte gesetzt werden, die ein späteres Übereinstimmen der willkürlich definierten Nullpunkte der

messzeitpunkt-spezifischen Skalen ermöglichen (Pohl, Haberkorn, & Carstensen, 2015). Die häufigste Form von Ankerpunkten stellen Items dar, die in allen Testformen der zu verlinkenden Messzeitpunkte unverändert eingesetzt werden (Kolen & Brennan, 2014; siehe Abbildung 2). Diese sogenannten Ankeritems dienen als Grundlage, anhand derer die Linkinformation zum Verknüpfen von Messzeitpunkten gewonnen wird.

Im Kontext der IRT wird die „Schwierigkeit“ eines Items als weitestgehend unveränderliche Eigenschaft desselben angenommen. So wird gewährleistet, dass eine intraindividuelle Entwicklung klar als solche erkannt wird, da bei Fixieren der Itemschwierigkeit eine auftretende Veränderung auf die Personenfähigkeit zurückgeführt werden kann. Nicht immer können Items wiederholt eingesetzt werden. Kurze Zeitabstände zwischen Messzeitpunkten könnten Erinnerungseffekte begünstigen und die Probanden potentiell irritieren (Pohl & Carstensen, 2013). Sehr einprägsame Items (z.B. bei Lesetests) scheinen ebenso wenig geeignet für den wiederholten Einsatz. Deshalb kann alternativ eine Testvorgabe beider Testformen (von Messzeitpunkten t_1 und t_2) zu einem einzelnen Messzeitpunkt an einer zusätzlichen Stichprobe erfolgen (Vale, 1986). Diese Stichprobe, oder auch Ankergruppe, wird eigens zum Zweck der Datenverknüpfung gezogen (siehe Abbildung 2) und sollte aus derselben Population wie die Längsschnittstichprobe stammen. Die Ankergruppe sollte also in ihren demographischen Eigenschaften wie auch in der Verteilung des interessierenden latenten Merkmals mit der Längsschnittstichprobe vergleichbar sein (Pohl & Carstensen, 2012).

Abbildung 2

Veranschaulichung zweier Linkdesigns als Grundlage des Verlinkens zweier Messzeitpunkte t_1 und t_2



Anmerkung. Die Linkinformation (dargestellt als graue Rechtecke) kann mittels verschiedener Linkdesigns gewonnen werden. Im Ankeritem-Design wird eine Teilmenge von Items (sogenannte Ankeritems) zu beiden Messzeitpunkten, t_1 und t_2, eingesetzt. Im Ankergruppe-Design werden die Testformen beider Messzeitpunkte, t_1 und t_2, zu einem einzelnen Messzeitpunkt einer zusätzlich gezogenen Stichprobe (Ankergruppe) vorgegeben. Dieser Messzeitpunkt kann entweder t_1 oder t_2 entsprechen oder zeitlich dazwischenliegen.

IRT-Linkmethoden

Basierend auf dem Linkdesign, wird die gesammelte Linkinformation zum Verknüpfen der Messzeitpunkte herangezogen. Die eingesetzte IRT-Linkmethode (im Weiteren Linkmethode genannt) bestimmt hierbei, wie die Linkinformation verwendet wird. Trotz identischer Linkinformation kann demnach das Ergebnis des Linkprozesses – also die resultierende mittlere Veränderung des latenten Merkmals zwischen zwei verknüpften Messzeitpunkten – unterschiedlich ausfallen. Unabhängig von der Anzahl der zu verknüpfenden Messzeitpunkte lassen sich einstufige und zweistufige Linkmethoden unterscheiden.

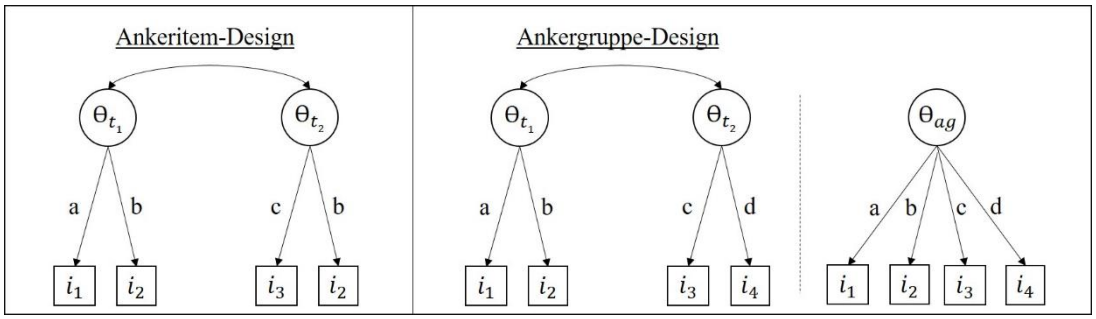
Einstufige Linkmethoden. Diese sind dadurch charakterisiert, dass die Parameter des Ankeritems *i,* welches wiederholt eingesetzt wird (z. B. i_{t1}, i_{t2}, i_{t3}) nicht unabhängig voneinander

geschätzt werden. Zu den einstufigen Linkmethoden zählen beispielsweise die Concurrent Calibration (CC) und die Fixed Parameter Calibration (FPC).

Concurrent Calibration. Die Daten beider Testformen werden zu einem Datensatz zusammengefügt und in einem mehrdimensionalen Modell skaliert (siehe Abbildung 3). Hierbei dürfen die Messzeitpunkte korrelieren. Im Falle eines Ankeritem-Designs werden die Ankeritemparameter verschiedener Messzeitpunkte (d. h. Dimensionen) gleichgesetzt. Das IRT-Modell wird identifiziert durch die Fixierung des Stichprobenmittelwertes des ersten Messzeitpunktes auf null. Dadurch reflektieren darauffolgende Messungen eine Kompetenzveränderung gemessen an dem ersten Messzeitpunkt. Im Falle eines Ankergruppen-Designs werden zusätzlich die Daten der Ankergruppe zu dem Datensatz hinzugefügt (s. Abb. 3). Das verwendete Modell ist also ein mehrdimensionales IRT-Mehrgruppenmodell. Während jeder Messzeitpunkt der Längsschnittstichprobe auf einer einzelnen Dimension lädt (und dadurch unterschiedliche Mittelwerte der latenten Variable zulässt), laden alle Testformen der Ankergruppe auf einer einzigen Dimension. Dies ist bedingt dadurch, dass keine Fähigkeitsunterschiede in der Ankergruppe zwischen den Testformen auftreten können. Die Itemparameter zwischen den Gruppen werden gleichgesetzt. Auch hier wird der Mittelwert der latenten Variable des ersten Messzeitpunktes der Längsschnittstichprobe auf null fixiert um das Modell zu identifizieren.

Abbildung 3

Verlinken zweier Messzeitpunkte mittels Concurrent Calibration bei verschiedenen Linkdesigns



Anmerkung. $\theta_{t_1}, \theta_{t_2}$ und θ_{ag} repräsentieren die latente Fähigkeit zum jeweiligen Messzeitpunkt. Durch fixieren von θ_{t_1} auf null wird das Modell identifiziert. Die Buchstaben a – d repräsentieren die gleichgesetzten Itemparameter ik der verschiedenen Testformen (siehe Text für eine detailliertere Beschreibung).
ag = Ankergruppe.

Fixed parameter calibration. Die Parameter der Ankeritems, welche von dem ersten Messzeitpunkt stammen, werden in allen nachfolgenden Messzeitpunkten fixiert (also nicht geschätzt; Kim, 2006). Infolgedessen werden nur testformspezifische Items geschätzt, während die Ankeritems vom ersten Messzeitpunkt die Referenzskala definieren. Im Falle eines Ankergruppe-Designs werden die Itemparameter, welche aus der Skalierung der Ankergruppe stammen, als fixe Itemparameter in die Längsschnittstichprobe übernommen. Das Fixieren von Itemparametern impliziert automatisch eine identische Skala bei allen nachfolgenden Messzeitpunkten. Diese Linkmethode ist sehr strikt und lässt keinen Spielraum für messzeitpunktspezifische Unterschiede in den Itemparametern zwischen den Messzeitpunkten zu.

Zweistufige Linkmethoden. Bei den zweistufigen Linkmethoden werden in einem ersten Schritt die verschiedenen Messzeitpunkte unabhängig voneinander skaliert, um danach – in einem zweiten Schritt – die Übereinstimmung der willkürlich definierten Nullpunkte der zu

verknüpfenden messzeitpunkt-spezifischen Skalen vorzunehmen. Sollen beispielsweise Messzeitpunkte der Klassen 1 bis 3 miteinander verlinkt werden, so wird erst der unabhängig skalierte Messzeitpunkt von Klasse 2 mit Klasse 1 verlinkt. Anschließend wird der unabhängig skalierte Messzeitpunkt von Klasse 3 mit Klasse 2 verlinkt. Dabei machen sich die Linkmethoden die Invarianz von IRT Skalen gegenüber linearen Transformationen (Kolen & Brennan, 2014), wie z. B.

$$\theta^*_{i,t_2} = \theta_{i,t_2} + \mathrm{B}, \tag{4}$$

zunutze. Gleichung (4) gilt, wenn dieselbe lineare Transformation auch auf den Itemparameter b_j angewandt wird. Merke, dass Gleichung (4) für das 1PLM gilt und für komplexere Messmodelle angepasst werden muss (das würde an dieser Stelle jedoch zu weit führen; interessierte Lesende finden eine detailliertere Darstellung hierzu in Anhang A). Die additive Komponente B wird aus den geschätzten Parametern (welche auf der Längsschnittstichprobe beruhen) der zur Linkinformation beitragenden Items berechnet und anschließend zu jedem Parameter der zu verlinkenden Skala addiert. Das konkrete Vorgehen ist von Linkdesign und Linkmethode abhängig. Zu den zweistufigen Linkmethoden, welche sich in 2PLM und 3PLM bewährt haben, zählen beispielsweise die mean/mean Linkmethode (Loyd & Hoover, 1980), die weighted mean/mean Linkmethode (van der Linden & Barrett, 2016), die mean/sigma Linkmethode (Marco, 1977) und die Characteristic Curve Methoden (Haebara, 1980; Stocking & Lord, 1983). Die folgende detaillierte Beschreibung der Linkmethoden gilt für ihre Anwendung im 1PLM.

Mean/mean Linkmtehode. Die Berechnung der additiven Komponente B ist abhängig vom Linkdesign. Bei einem Ankeritem-Design berechnet sich B wie folgt (Kolen & Brennan, 2014):

$$\mathrm{B} = M\left(b_{j,t_1}\right) - M\left(b_{j,t_2}\right), \tag{5}$$

wobei $M\left(b_{j,t_1}\right)$ und $M\left(b_{j,t_2}\right)$ die Mittelwerte der Ankeritemparameter der Messzeitpunkte t_1 und t_2 darstellen. Die verlinkten Itemparameter zum zweiten Messzeitpunkt ergeben sich schließlich aus

$$b^*_{j,t_2} = b_{j,t_2} + \mathrm{B}. \tag{6}$$

Bei einem Ankergruppe-Design müssen sowohl die Ankergruppe als auch der zweite Messzeitpunkt an die Skala des ersten Messzeitpunktes angeglichen werden. Wir nehmen an, dass $b_{j,t_1,ls}$ und $b_{k,t_2,ls}$ die Schwierigkeitsparameter der *j* bzw. *k* Items repräsentieren, die der Längsschnittstichprobe zum ersten Messzeitpunkt t_1 bzw. zum zweiten Messzeitpunkt t_2, vorgegeben wurden. Ebenso seien $b_{j,t_0,ag}$ und $b_{k,t_0,ag}$ die Schwierigkeitsparameter der *j* und *k* Items, welche der Ankergruppe zu einem gemeinsamen Messzeitpunkt vorgegeben wurden. Das Verlinken kann in zwei Schritte gegliedert werden.

1. Um die Ankergruppe an die Skala des ersten Messzeitpunktes der Längsschnittstudie anzugleichen, dienen die *j* Items als Ankeritems. Orientiert an Gleichung (5) ergibt sich
   $\mathrm{B}_1 = M\left(b_{j,t_1,ls}\right) - M\left(b_{j,t_0,ag}\right)$.
2. Um den zweiten Messzeitpunkt t_2 der Längsschnittstichprobe mit der (zuvor verlinkten) Ankergruppe zu verlinken, dienen die k Items, welche zu t_2 vorgegeben wurden, als Ankeritems. Abermals orientiert an Gleichung (5), ergibt sich
   $\mathrm{B}_2 = M\left(b^*_{k,t_0,ag}\right) - M\left(b_{k,t_2,ls}\right)$.

Fischer et al. (2019) folgend, finden also sowohl B_1 als auch B_2 Beachtung in der schlussgültigen Gleichung

$$\mathrm{B} = M\left(b^*_{k,t_0,ag}\right) + M\left(b_{j,t_1,ls}\right) - M\left(b_{j,t_0,ag}\right) - M\left(b_{k,t_2,ls}\right). \quad (7)$$

Die verlinkten Itemparameter in der Längsschnittstichprobe ergeben sich somit aus $b^*_{k,t_2,ls} = b_{k,t_2,ls} + \mathrm{B}$.

Eine erneute Skalierung des zweiten Messzeitpunktes mit den fixierten transformierten Werten b^*_{j,t_2} (im Ankeritem-Design) und $b^*_{k,t_2,ls}$ (im Ankergruppe-Design) ergibt die verlinkten Personenparameter θ^*_{i,t_2}. Merke, dass das Verhältnis der Parameter zueinander erhalten bleibt und somit der Modellfit unbeeinflusst ist.

Weighted mean/mean Linkmetode. Diese neuere Methode bezieht den Standardfehler der Itemschwierigkeitsparameter als Gewichte mit ein. Damit soll Unterschieden im item targeting bei Ankeritems begegnet werden. Gemeint sind also die Itemschwierigkeitsparameter von Ankeritems, die sich in ihrer Passung zum Stichprobenmittelwert unterscheiden.

Bezüglich mean/sigma Linkmethode und characteristic curve Methoden sei kurz erwähnt, dass beide im 1PLM kaum praktikabel anwendbar sind. Die mean/sigma Linkmethode würde den im 1PLM fixierten Parameter *a* manipulieren, weshalb die verlinkten Messzeitpunkte inhaltlich nicht mehr miteinander vergleichbar wären. Die characteristic curve Methoden würden hingegen ihren Vorteil im 1PLM verlieren, da die Modellpassung – wie bereits oben erwähnt – bei Anwendung der mean/mean Linkmethode unverändert bleibt. Interessierte Lesende finden hierzu eine kurze Ausführung in Anhang A.

Datenanforderungen

Items sind die kleinsten Bausteine in der Operationalisierung eines latenten psychologischen Konstruktes, weshalb ihnen besonderes Augenmerk in der psychometrischen Qualitätskontrolle zukommt. Eine solche Qualitätskontrolle auf Itemebene – und im Weiteren auch auf Testebene – gelingt mit Hilfe unterschiedlicher statistischer Tests (Maydeu-Olivares, 2013). Die Ergebnisse dieser helfen bei der Bewertung der Modellpassung und der Auswahl eines IRT Modells. Spezifische Anforderungen, nämlich die der Eindimensionalität als auch die der Messstabilität, werden an Items gestellt, welche zur Berechnung der Linkinformation verwendet werden.

Eindimensionalität. Um aussagekräftige Ergebnisse aus Veränderungsmessungen zu erhalten, muss das latente Konstrukt sowie dessen Operationalisierung über die Messzeitpunkte inhaltlich unverändert bleiben. Dies ist gegeben, wenn unterschiedliche Testformen aus verschiedenen Messzeitpunkten das Kriterium der Eindimensionalität erfüllen. Im Ankeritem-Design werden hierfür für jeden Messzeitpunkt zwei Modelle geschätzt und miteinander verglichen: Jeweils ein ein- und ein zweidimensionales Modell. Im zweidimensionalen Modell laden die Ankeritems auf der ersten Dimension und die testformspezifischen Items laden auf der zweiten Dimension. Analog dazu, werden im Ankergruppe-Design die beiden Dimensionen durch die beiden Testformen definiert. Vergleiche der Modellpassung zwischen einem ein- und zweidimensionalem Modell können dann z. B. anhand von Akaike's (1974) Informationskriterium oder dem Bayesianischen Informationskriterium (BIC; Schwarz, 1978) vorgenommen werden.

Stabilität psychometrischer Eigenschaften. Die Messung von Veränderung über die Zeit verlangt nach zeitstabilen Messinstrumenten. Nur wenn Items ihre relative Position auf der Logit Skala beibehalten, können Unterschiede aus verschiedenen Messungen auf die Personen rückgeführt werden. Um also das Ausmaß der kognitiven Veränderung über die Zeit abzubilden, müssen die psychometrischen Merkmale wiederholt vorgegebener/in verschiedenen Gruppen eingesetzter Items unveränderlich sein. Dies wird auch als Subgruppeninvarianz oder engl. differential item functioning (DIF; Holland & Wainer, 2012) bezeichnet. Die Prüfung der Stabilität psychometrischer Eigenschaften geschieht anhand statistischer Tests, deren Teststärke bekanntermaßen abhängig ist von der Stichprobengröße (Cohen, 1994). Gleichzeitig sind psychologische Testverfahren in ihrer Messgenauigkeit nicht vergleichbar mit physikalischen Messskalen wie z. B. Temperatur und Entfernung. Aus diesen Gründen scheint es in dieser Forschungsarbeit sinnvoll, die Definition von „unveränderlich“ vielmehr inhaltlich (z. B. basierend auf früheren Forschungsergebnissen) festzumachen anstelle – wie üblich – unhinterfragt eine strikte Null-Hypothese zu verwenden. Im Gegensatz zu einer klassischen Null-Hypothese wird bei Anwenden einer solchen „Minimum Effekt Hypothese“ (Murphy & Myors, 1999)ein (inhaltlich begründetes) akzeptables Maß an Ungenauigkeit festgelegt, mit welchem Items – samt ihrer psychometrischen Eigenschaften – als hinreichend stabil und „unverändert“ anerkannt werden. Dieses Vorgehen schafft zudem eine Diskussionsgrundlage im wissenschaftlichen Austausch, welches besonders wichtig scheint vor dem Hintergrund der im letzten Jahrzehnt aufgekommenen Debatte um Replizierbarkeit in den sozialen Wissenschaften (z.B. Shrout & Rodgers, 2018).

Linkfehler

Da Messung und Schätzung der latenten Variable fehlerbehaftet sind, ist das Linkergebnis beeinflusst durch die Auswahl an Ankeritems aus einem theoretisch unerschöpflichen Pool möglicher Items. Die Auswahl anderer Ankeritems könnte zu einer anderen Linkkonstante B führen. Dieser Fehlerquelle wird im Linkfehler Rechnung getragen. Der Linkfehler kommt beim Einsatz statistischer Tests zum Tragen, welche auf verlinkten Daten basieren (d. h. zumindest zwei Messzeitpunkte umfassen). Ein größerer Linkfehler führt hierbei zu einem Verlust von Teststärke, da er in den gepoolten Standardfehler eingeht. Siehe Fischer et al. (2019) oder auch OECD (2014) für mehr Information.

Linkfehler im Ankeritem-Design. Der Linkfehler basiert auf den k Ankeritems. Im 1PLM wird die Differenz $\Delta b_{j,1PLM} = b_{j,t_1} - b^*_{j,t_2}$ zwischen dem Ankeritemparameter vom ersten Messzeitpunkt t_1, b_{j,t_1} und dem verlinkten Ankeritemparameter vom zweiten Messzeitpunkt b^*_{j,t_2} berechnet, welcher aus Gleichung (6) resultiert. Angelehnt an PISA 2012 (OECD, 2014), ergibt sich der Linkfehler aus der Standardabweichung dieser Differenzen, standardisiert an der Anzahl der Ankeritems: $Linkfehler_{\text{AID}} = \frac{SD(\Delta b_j)}{\sqrt{k}}$. Bei Verwendung der Fixed Parameter Calibration müssen vor der Berechnung der Differenz $\Delta b_{j,1PLM} = b_{j,t_1} - b_{j,t_2}$ die Mittelwerte der Ankeritemparameter beider Messzeitpunkte (t_1 sowie t_2 vor dem Verlinken) einander angeglichen werden.

Linkfehler im Ankergruppe-Design. Im Ankergruppe-Design berechnet sich der Linkfehler aus dem gepoolten Linkfehler für die k_{t1} Items von Messzeitpunkt t_1 und die k_{t2} Items von Messzeitpunkt t_2. Als solches entstehen zwei Differenzen: a) die Differenzen zwischen den

Itemparametern von t_1 der Längsschnittstichprobe und der Ankergruppe, $\Delta b_{j,t_1} = b_{j,t_1,ls} - b_{j,t_1,ag}$ sowie b) die Differenzen zwischen den verlinkten Itemparametern von t_2 der Längsschnittstichprobe und der Ankergruppe, $\Delta b_{j,t_2} = b^*_{j,t_2,ls} - b_{j,t_2,ag}$. Der Linkfehler wird dann berechnet als: $Linkfehler_{\text{AGD}} = \sqrt{\left(\frac{SD(\Delta b_{j,t_1})}{\sqrt{k_{t_1}}}\right)^2 + \left(\frac{SD(\Delta b_{j,t_2})}{\sqrt{k_{t_2}}}\right)^2}$. Bei Verwendung der Fixed Parameter Calibration berechnet sich die Differenz aus den Itemparametern der unverlinkten Saklierung sowie den verlinkten Itemparametern (d.h. den Itemeparametern, welche von der Ankergruppe stammen). Auch hier müssen davor die Mittelwerte der Ankeritemparameter einander angeglichen werden.

Aktueller Forschungsstand

Da das Verknüpfen von Datenerhebungen traditionell aus dem Bereich angloamerikanischer LSAs stammt, basiert die meiste Forschung zum Leistungsvergleich zwischen Linkmethoden auf (in den USA üblicheren) komplexeren statistischen Modellen (wie dem 2PLM und dem 3PLM), Querschnittdaten, Ankeritem-Design, großen Stichproben und relativ gering zu erwartenden Unterschieden in den Mittelwerten der latenten Variable zwischen Gruppen. Die Ergebnisse aus der Vielzahl an vorliegenden Studien sind zumeist heterogen, was an der großen Zahl experimentell variierter Faktoren liegen mag. Unabhängig von der zu Grunde liegenden Linkmethode findet sich jedoch Übereinstimmung darin, dass die Reliabilität des Linkergebnisses positiv mit der Stichprobengröße sowie der Ankeritemzahl korreliert. Weitestgehend unerforscht ist die Performanz verschiedener Linkmethoden im Kontext starker Fähigkeitsveränderungen zwischen Messzeitpunkten, kurzer Testlänge, relativ wie absolut geringer Anzahl an Ankeritems sowie moderat eingeschränkter Modellgeltung, wie sie

üblicherweise in Längsschnitterhebungen gegeben sind, was Ausschlag für die vorliegende Forschungsarbeit war.

Offene Forschungsfragen

Wie oben dargestellt, findet die IRT immer häufiger Eingang in die breite Forschung und Anwendung, sodass sich ihr Einsatzgebiet stetig erweitert und auch Forschungsfragen bedienen soll, die mit kleineren Stichprobenzahlen und kürzeren Tests auskommen (müssen). Diese Einschränkungen begrenzen aus methodischer Sicht maßgeblich das Angebot statistischer Modelle auf parametersparsame Versionen. Diesem Umstand ist im aktuellen Forschungsstand bei der Erforschung von kognitiven Entwicklungsverläufen jedoch nicht ausreichend Rechnung getragen. Es gibt eine Forschungslücke betreffend des Verlinkens längsschnittlicher, Rasch skalierter Daten. Wenn auch es keine methodischen Unterschiede zwischen dem Verlinken von querschnittlich erhobenen – zumeist stichprobenstarken – Datensätzen aus Kohortenvergleichen sowie Daten zum längsschnittlichen Entwicklungsverlauf gibt, so gilt es doch bei einer Stichprobe mehrfach befragter Personen einige Besonderheiten zu beachten, die zum einen das Testdesign betreffen und zum anderen mit zusätzlichen Anforderungen an die erhobenen Daten einhergehen. Bezüglich ersterem muss beispielsweise die Möglichkeit von Erinnerungseffekten bedacht werden. Konsequenzen auftretender Erinnerungseffekte gefährden potentiell die Stabilität von Itemschwierigkeiten bei Ankeritems. Zusätzlich muss versucht werden, die Motivation der Testpersonen langfristig aufrecht zu erhalten, weshalb die Testbelastung niedrig gehalten werden sollte. Dies verlangt einerseits möglichst kurze Testbatterien als auch eine besonders gute Passung zwischen dem Schwierigkeitsniveau eines Tests und der mittleren Personenfähigkeit zu einem Messzeitpunkt. Daraus folgt zum einen, dass die absolute Anzahl zu

verwendender Ankeritems pro Messzeitpunkt niedriger ist als bei Querschnittdesigns und zum anderen sollen die ausgewählten Ankeritems ein Höchstmaß an Information zum mittleren Fähigkeitsniveau zu allen der zu verknüpfenden Messzeitpunkte liefern.

Basierend auf dem hier skizzierten status quo leiten sich folgende Forschungsfragen ab, die nachfolgend – gegliedert in die drei vorliegenden Beiträge – dargestellt werden.

Beitrag 1. Anhand empirischer Daten untersucht Beitrag 1 (Fischer et al., 2019), ob Linkmethoden, welche üblicherweise in komplexeren Messmodellen angewandt werden, in das 1PLM übertragbar sind. Im Weiteren werden Kriterien identifiziert, welche möglicherweise vorliegende Unterschiede sichtbar machen und diese in einen interpretativen Rahmen setzen. Unterschiede im Ergebnis zwischen verschiedenen Linkdesigns und Linkmethoden sind naturgemäß zu erwarten, da die für den Linkprozess verwendete Information eine jeweils andere ist. Die einzigartige Studienkonstellation in Beitrag 1 ermöglicht es, drei Linkmethoden (Concurrent Calibration, Fixed Parameter Calibration und mean/mean Linkmethode) sowie zwei Linkdesigns (Ankeritem, Ankergruppe) miteinander zu vergleichen. Die Unterschiede werden anhand der drei Kriterien Linkfehler, gefundene mittlere Veränderung zwischen den Messzeitpunkten und Modellpassung verglichen. Die Ergebnisse dieser Studie, welche auf empirischen Daten beruhen, haben eine hohe ökologische Validität.

Beitrag 2. Um den Einfluss verschiedener Faktoren auf das Linkergebnis in einem experimentellen Design zu untersuchen, bedient sich Beitrag 2 (Fischer et al., 2021) einer Simulationsstudie. Bei dieser wird eine Längsschnittmessung mit vier Messzeitpunkten ($t_1 – t_4$) und einem Ankeritem-Design simuliert. Das fiktive Wachstum orientiert sich hierbei an einer abflachenden Wachstumskurve aus dem NEPS, um eine Generalisierbarkeit der Ergebnisse trotz

des artifiziellen Studiencharakters zu erlauben. Das hierbei simulierte Wachstum ist zum Teil beträchtlich (bis zu 0.7 logits) und stellt damit die vier eingesetzten Linkmethoden (Concurrent Calibration (CC), Fixed Parameter Calibration (FPC), mean/mean Linkmethode sowie weighted mean/mean (wm/m) Linkmethode) auf einen aus Kohortenvergleichen kaum gekannten Prüfstand. Die folgenden Variablen werden experimentell variiert: Ankeritemzahl: Die bei einer Testlänge von 25 Items eingesetzte Zahl an Ankeritems entspricht in ihrem relativen Anteil zwar den Empfehlungen der Literatur, fällt aber in absoluten Zahlen gesehen sehr klein aus (3 (12%), 5 (20%), 7 (28%), 9 (36%)) und findet ihre Motivation in der Frage nach der Stabilität der Linkmethoden. Stichprobengröße: Abhängig von der Stichprobengröße ist die Genauigkeit der Parameterschätzung, welche einen direkten Einfluss auf die Linkinformation hat. Modellpassung: Um die Robustheit gegenüber Modellverletzungen zu prüfen, werden von 1 abweichende Diskriminationsparameter simuliert.

Die resultierenden Unterschiede werden anhand von Konvergenzrate, Bias, relativem Bias und root mean square error (RMSE) des geschätzten Mittelwertes und der geschätzten Varianz des latenten Merkmals untersucht um Antwort auf die Frage zu finden, wie sehr die geschätzte von der wahren Veränderung abweicht. Der Bias wurde berechnet als $\hat{\tau}_d - \tau$, wobei $\hat{\tau}_d$ der geschätzte Parameter der *k*-ten Replikation von Bedingung *d* darstellt, sowie τ dem wahren Wert entspricht. Nachfolgend wurde der Bias über alle *k* Replikationen jeder Bedingung gemittelt. Der relative Bias dient als Effektstärke und wird berechnet als Quotient $(\bar{\tau}_d - \tau)/\tau$, wobei $\bar{\tau}_d$ der gemittelte geschätzte Parameter über alle *k* Replikationen ist. Der RMSE hingegen lässt Rückschlüsse über die Präzision der Parameterschätzung zu und wurde berechnet mittels

$\sqrt{\frac{1}{c}\sum_{k=1}^{c}(\bar{\tau}_k - \tau)^2}$. Als solches ist der RMSE definiert als Wurzel aus dem Mittelwert des quadrierten Bias.

Beitrag 3. Die im Rahmen dieser Forschungsarbeit gewonnenen Erkenntnisse finden ihre praktische Anwendung im NEPS. In Beitrag 3 (Fischer, Rohm, Gnambs, & Carstensen, 2016) wird der Prozess des Verlinkens anhand von zwei empirischen Beispielen auf detaillierte Art und Weise beschrieben, um das Vorgehen nachvollziehbar und anschaulich zu machen. Ziel der Forschungsarbeit ist unter anderem, die aus der Forschung abgeleiteten Schlussfolgerungen einer breiteren und methodisch weniger geübten Masse zugänglich zu machen und so die Kompetenzentwicklung im breiten Feld der sozialen Wissenschaften weiter zu stärken.

Diskussion

Die vorliegende Forschungsarbeit legt den Fokus auf die bis jetzt wenig erforschte Thematik der Modellierung kognitiver Entwicklungsverläufe und die damit einhergehenden Herausforderungen wie z.B. starke Fähigkeitsveränderung zwischen Messzeitpunkten, kurze Testlänge, limitierte Stichprobengröße und potentiell eingeschränkte Modellgeltung. Das Verlinken angrenzender Messzeitpunkte erfolgt im Rahmen von Item Response Theory und dem Rasch Modell.

Überraschenderweise zeigen sich sowohl im ersten Beitrag (Fischer et al., 2019) als auch im zweiten Beitrag (Fischer et al., 2021) nur marginale Unterschiede im Linkergebnis zwischen den verschiedenen IRT-Linkmethoden. Die Linkmethoden scheinen also gleichermaßen unempfindlich gegenüber den Faktoren Modellpassung, Stichprobengröße und Ankeritemzahl. Auch große Veränderungen in der latenten Variable können zufriedenstellend

abgebildet werden. Insgesamt weicht dieses Ergebnis ab von eher inkohärenten Resultaten in der bestehenden Forschungsliteratur. In der Praxis scheint es dennoch ratsam, die Ergebnisse verschiedener Linkmethoden als Kontrollmaßnahme miteinander zu vergleichen, da sie auf unterschiedlichen Linkinformationen beruhen.

Hervorhebenswert ist der Unterschied im Linkergebnis zwischen den beiden angewandten Linkdesigns (Fischer et al., 2019). Es zeigt sich ein signifikant geringeres Wachstum unter Anwendung eines Ankergruppe-Designs verglichen mit einem Ankeritem-Design. Da die Linkinformation bei ersterem alleinig aus der Ankergruppe stammt, scheint dieses Ergebnis gut erklärbar mit dem in der Literatur beschriebenen Effekt eines systematischen Stichprobenausfalls in Längsschnitterhebungen (z.B. Beaver, 2013), zumal hier die Motivation zur Teilnahme durch die Probanden mehrfach aufgebracht werden muss. Tatsächlich ergab sich bei weiteren Analysen, dass die mittlere Fähigkeit von Personen, die zu beiden Messzeitpunkten (Klasse 5 [Duchhardt & Gerdes, 2012] und Klasse 7 [Schnittjer & Gerken, 2017]) an der Erhebung mathematischer Kompetenz teilgenommen hatten (N = 3 833, M = 0.11 Logits, SD = 1.15), in Klasse 5 um 0.45 logits höher war, als die mathematische Kompetenz von Personen, die nachfolgend nicht an der Erhebung in Klasse 7 teilgenommen haben (N = 1 360, M = -0.34 Logits, SD = 1.16. Es scheint also denkbar, dass ein systematischer Stichprobenausfall die Bandbreite von Entwicklungsverläufen einschränkt und somit die Generalisierbarkeit des Linkergebnisses reduziert. Während das Linkergebnis im Ankergruppe-Design potentiell durch einen größeren (weil doppelten) Linkfehler samt Fehler in der Parameterschätzung beeinflusst wird, ist bei der Verwendung eines Ankeritem-Designs also möglicherweise mit einer (positiven) Verzerrung der mittleren kognitiven Fähigkeitsentwicklung zu rechnen. Tiefer gehende Analysen

bezüglich eines möglichen systematischen Stichprobenausfalls in Längsschnitterhebungen samt dessen Auswirkungen auf das Linkergebnis scheinen daher ratsam.

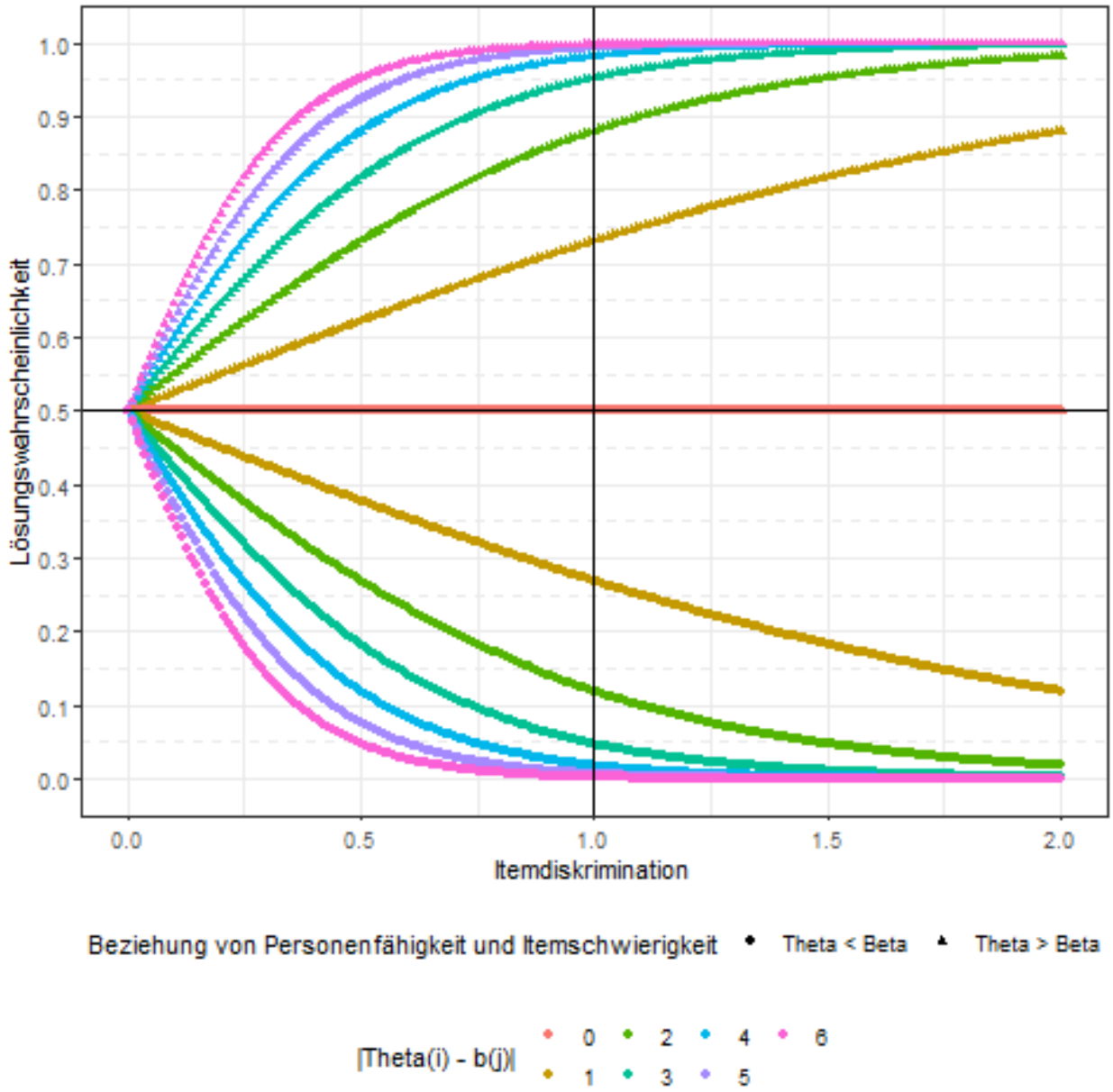
Betreffend die Frage nach der Stabilität von Linkmethoden, zeigt die Concurrent Calibration deutliche Auffälligkeiten: Bis zu 79% der Modelle, welche die Concurrent Calibration anwendeten, erreichten das Konvergenzkriterium nicht. Während es kaum einen Zusammenhang mit den Faktoren Modellpassung und Stichprobengröße zu haben schien, gab es deutliche Unterschiede betreffend der Ankeritemzahl: Die Konvergenzraten stiegen bei größeren Ankeritemzahlen auf bis zu 83%. Jedoch wiesen auch die geschätzten Parameter aus konvergierten Modellen starke Abweichungen von den wahren Werten auf, weshalb die Methode der Concurrent Calibration in diesem Zusammenhang nicht empfehlenswert scheint. Keine der anderen Linkmethoden führt zu Konvergenzproblemen. Da bei Anwendung der Concurrent Calibration die Daten aller Messzeitpunkte gemeinsam geschätzt werden, lassen sich keine Aussagen dazu treffen, inwiefern die anfänglich große mittlere Veränderung des latenten Merkmals bei den Konvergenzproblemen eine Rolle gespielt hat.

In Fischer et al. (2021) wird deutlich, dass eine eingeschränkte Modellpassung durch moderat von $a = 1$ abweichende Diskriminationsparameter von Ankeritems zu einer leichten Verschätzung von Mittelwert und Standardabweichung der latenten Variable führt. Dieser Effekt wird kleiner bei einer größeren Anzahl von Ankeritems und tritt nicht auf bei guter Modellpassung. Dies legt kompensatorische Effekte nahe, die durch das Zusammenspiel von Schwierigkeit, Diskrimination und Fähigkeit entstehen. Basierend auf dem 2PLM (wie dargestellt in Gleichung (2)), gibt Abbildung 4 Aufschluss über den Zusammenhang von Itemdiskrimination, Itemschwierigkeit und Personenfähigkeit im 2PLM: Unabhängig von der

absoluten Lokalisation auf der logit-Skala, gewichtet die Itemdiskrimination a_j die Differenz von Personenfähigkeit Θ_i und Itemschwierigkeit b_j. Mit zunehmender Größe dieser Differenz steigt der Einfluss der Itemdiskrimination a_j auf die resultierende Lösungswahrscheinlichkeit p_{ij}. Bei $\Theta_i = b_j$ nivelliert sich der Einfluss von a_j, sodass Rasch Modell und 2PLM in der Schätzung von p_{ij} übereinstimmen.

Abbildung 4

Darstellung der Beziehung zwischen Personenfähigkeit Θ_i, Itemschwierigkeit b_j und Itemdiskrimination a_j im 2PLM



Anmerkung. Die Kurven repräsentieren die Lösungswahrscheinlichkeit p_{ij} für den Betrag von $|\Theta_i - b_j|$ in ganzzahligen Schritten zwischen 0 und 6 in Abhängigkeit von a_j. Die durchgezogene vertikale Linie bei $a_j = 1$ repräsentiert p_{ij} bei Gültigkeit des Rasch Modells (also $a_j = 1$). Bei

Modellierung mit dem Rasch Modell von Daten, die dem 2PLM folgen, lässt sich die Richtung der zu erwartenden Verschätzung von p_{ij} erkennen.

Würde nun ein Test mit dem 1PLM modelliert, in dem einige Items dem 2PLM folgen, so fände der Diskriminationsparameter bei ebendiesen Items seinen Niederschlag in einem Schwierigkeitsparameter b_j^*, der vom wahren Wert b_j abweicht (Humphrey, 2018). Dieser Umstand wird besonders dann relevant, wenn dasselbe Item von zwei Gruppen mit unterschiedlicher Fähigkeit bearbeitet wird und gleichzeitig – bedingt durch $a_j \neq 1$ – die Schwierigkeitsschätzung nicht mehr unabhängig ist von der zu Grunde liegenden Stichprobe. Theoretisch lässt sich dies ableiten durch Gleichsetzen der Terme aus dem Zähler von (3) und (2):

$$a_j(\theta_i - b_j) = \theta_i - b_j^*. \tag{8}$$

Umformen nach b_j^* ergibt

$$b_j^* = \theta_i - a_j(\theta_i - b_j) \tag{9}$$

und weiter

$$b_j^* = \theta_i - a_j\theta_i + a_j b_j. \tag{10}$$

(9) verdeutlicht, warum große Mittelwertunterschiede in der latenten Variable von zwei zu verlinkenden Gruppen besonders problematisch sind bei der Verwendung des Rasch Modells bei Ankeritems mit $a_j \neq 1$: Abhängig von der Differenz zwischen Populationsfähigkeit und Itemschwierigkeit zu den Messzeitpunkten t_1 und t_2, würden die Schwierigkeitsparameter $b_{j,t1}^*$ und $b_{j,t2}^*$ resultieren und somit zu einem systematischen Fehler in der Linkinformation führen,

die sich in einem verzerrten Linkergebnis niederschlägt. Die Richtung dieses Fehlers lässt sich aus Abbildung 4 ableiten (siehe auch Tabelle 1).

Tabelle 1

Verschätzung der Lösungswahrscheinlichkeiten bei Rasch Modellierung von Daten, die dem 2PLM folgen

	$\boldsymbol{\theta_i < b_j}$	$\boldsymbol{\theta_i > b_j}$
$\boldsymbol{a_j < 1}$	$p_{ij,\mathrm{RM}} < p_{ij,2\mathrm{PLM}}$	$p_{ij,\mathrm{RM}} > p_{ij,2\mathrm{PLM}}$
$\boldsymbol{a_j > 1}$	$p_{ij,\mathrm{RM}} > p_{ij,2\mathrm{PLM}}$	$p_{ij,\mathrm{RM}} < p_{ij,2\mathrm{PLM}}$

Anmerkung. a_j = Itemdiskriminationsparameter von item j, b_j = Itemschwierigkeitsparameter von item j, Θ_i = Fähigkeit von Person i, $p_{ij,\mathrm{RM}}$ = Lösungswahrscheinlichkeit bei Rasch Modellierung von Daten, die dem 2PLM folgen, $p_{ij,2\mathrm{PLM}}$ = Lösungswahrscheinlichkeit bei der Modellierung mittels 2PLM von Daten, die dem 2PLM folgen.

Durch Einsetzen von (10) bei (2) ergibt sich

$$p_{ij}^{*}\left(X = 1 \middle| \theta_i, a_j, b_j\right) = \frac{\exp[\theta_i - (\theta_i - a_j\theta_i + a_j b_j)]}{1 + \exp[\theta_i - (\theta_i - a_j\theta_i + a_j b_j)]}, \tag{11}$$

was sich durch Kürzen in Gleichung (2) umformen lässt. Merke, dass (11) ≡ (3) bei $a_j = 1$ und/oder $\Theta_i = b_j$. Dies zeigt, dass bei Verletzung der Annahmen des Rasch Modells der Einfluss von $a_j \neq 1$ bei gutem test targeting abgefangen oder zumindest abgemildert werden kann und begründet somit auch statistisch die Wichtigkeit einer gelungenen Passung zwischen Itemschwierigkeit und Personenfähigkeit.

Abschließend sei erwähnt, dass die Autorin es als wichtig ansieht, dass die Methodik der Praxis zuarbeitet, um den Spalt zwischen Forschung und Anwendung durch zügigen Wissenstransfer so gering wie möglich zu halten. Denn, wie anfänglich erwähnt, möchte die Psychologie als Wissenschaft beschreiben und erklären – aber erst eine Übertragung ebendieser

Erkenntnisse in die Praxis kommt auch dem psychologischen Forschungsobjekt – dem Menschen – im Alltag zu Gute.

Einschränkung der Ergebnisgültigkeit

Die vorliegende Forschungsarbeit basiert auf empirischen wie simulierten Daten. Der Erkenntnisgewinn aus den Studien ist deshalb aus unterschiedlichen Gründen eingeschränkt. So führt die Verwendung artifizieller Daten zu ebenso artifiziellen Ergebnissen, die sich nur eingeschränkt in die Praxis übertragen lassen. Diesem Argument seien jedoch zwei Punkte entgegengesetzt: Zum einen haben sich die generierten Daten stark an in der Praxis gewonnen Daten orientiert. Zum anderen sollte die Verwendung der Simulationsstudie den Effekt verschiedener Faktoren auf das Linkergebnis sichtbar machen, um den Mechanismus des Verlinkens besser verstehen zu können.

Weiterführende Forschungsfragen

Eine Auswahl möglicher, sich anschließender Forschungsfragen sei zum Ende dieser Arbeit kurz erwähnt:

a) Dem test targeting kommt eine mediierende Rolle beim Verlinken von Messzeitpunkten zu. Es hat direkten Einfluss auf die Genauigkeit der Parameterschätzung und somit auf die resultierende Linkinformation. Weiter hat es einen indirekten Einfluss auf die Auswirkung einer Modellverletzung bei $a_j \neq 1$. Obwohl die vorliegenden Daten teilweise große Fähigkeitsveränderungen (von bis zu 0.7 logits) zwischen den Messzeitpunkten beinhalten, so war das test targeting stets optimiert. Weitere Studien sind notwendig, um

den Mediator test targeting und seine Bedeutsamkeit im Linkprozess weiter zu untersuchen.

b) Bei eingeschränkter Modellpassung haben sich unter Anwendung des 1PLM kompensatorische Effekte bei zunehmender Ankeritemzahl gezeigt. Weiter zu untersuchen ist, in welcher Art und Weise Itemdiskrimination, Itemschwierigkeit und Personenfähigkeit detailliert zusammenwirken und wie sie das Linkergebnis beeinflussen. Lässt sich dieser Einfluss in der Praxis messen, unterscheiden sich ein- und zweistufige Linkmethoden in ihrer Robustheit diesbezüglich und welche Bedeutung ergibt sich für die Auswahl von Ankeritems, sind Fragen, die weiterführend zu untersuchen sind.

c) Weiterführende Analysen werden benötigt, um die unterschiedlichen Ergebnisse bisheriger Forschungsarbeiten und der vorliegenden Forschungsarbeit bezüglich des Leistungsvergleiches von Linkmethoden zu untersuchen. Wie erklärt sich das Auftreten von Unterschieden zwischen Linkmethoden bei bisherigen Forschungsergebnissen und wie das Fehlen dieser in der vorliegenden Forschungsarbeit?

d) Das Ankergruppe-Design ist nicht von einem potentiellen systematischen Fehler bei der Parameterschätzung wiederholt vorgegebener Items betroffen und sieht sich auch nicht mit einem systematischen Stichprobenausfall konfrontiert. Die Reliabilität des Linkergebnisses ist hierbei allein durch die Repräsentativität und die Größe der Ankergruppe beeinflusst und deren Anwendung scheint somit eher eine ökonomische Frage zu sein. Unter welchen Umständen lässt sich dieser ökonomische Mehraufwand rechtfertigen?

Literaturverzeichnis

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. Retrieved from https://doi.org/10.1109/TAC.1974.1100705

Beaver, K. M. (2013). Intelligence and selective attrition in a nationally representative and longitudinal sample of Americans. *Personality and Individual Differences*, *55*(2), 157–161. https://doi.org/10.1016/j.paid.2013.02.015

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord, and M. R. Novick (Eds.), *Statistical Theories of Mental Test Scores* (pp. 397-479). Addison-Wesley: Reading.

Blossfeld, H.-P. (Ed.) (2011). *Zeitschrift für Erziehungswissenschaft Sonderheft: Vol. 14*. *Education as a lifelong process: The German National Educational Panel Study (NEPS)*. Wiesbaden: VS-Verl.

Cohen, J. (1994). The Earth Is Round (p <.05). *American Psychologist*, *49*(12), 997–1003. https://doi.org/10.4324/9781315629049-14

De Ayala, R. J. (Ed.) (2022). *Methodology in Social Sciences*. *The theory and practice of item response theory* (Second edition). New York, London: Guilford Press.

Duchhardt, C., & Gerdes, A. (2012). *NEPS Technical Report for mathematics – scaling results of Starting Cohort 3 in fifth grade* (NEPS Working Paper). Bamberg.

Fischer, G. H., & Molenaar, I. W. (2012). *Rasch Models: Foundations, Recent Developments, and Applications*. Springer Science & Business Media.

Fischer, L., Gnambs, T., Rohm, T., & Carstensen, C. H. (2019). Longitudinal linking of Rasch-model-scaled competence tests in large-scale assessments: A comparison and evaluation of different linking methods and anchoring designs based on two tests on mathematical competence administered in grades 5 and 7. *Psychological Test and Assessment Modeling*, *61*, 34–67.

Fischer, L., Rohm, T., Gnambs, T., & Carstensen, C. (2016). Linking the Data of the Competence Tests (NEPS Survey Paper No. 1). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

Fischer, L., Rohm, T., Carstensen, C. H., & Gnambs, T. (2021). Linking of Rasch-Scaled Tests: Consequences of Limited Item Pools and Model Misfit. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.633896

Haebara, T. (1980). Equating Logistic Ability Scales by a Weighted Least Squares Method. *Japanese Psychological Research*, *22*(3), 144–149. https://doi.org/10.4992/psycholres1954.22.144

Holland, P. W., & Wainer, H. (2012). *Differential Item Functioning*. Routledge. https://doi.org/10.4324/9780203357811

Humphrey, S. N. (2018). The Impact of Levels of Discrimination on Vertical Equating in the Rasch Model. *Journal of Applied Measurement*, *19*(3), 216–228.

Kim, S.-H., Kwak, M., Bian, M., Feldberg, Z., Henry, T., Lee, J., . . . Cohen, A. S. (2020). Item Response Models in Psychometrika and Psychometric Textbooks. *Frontiers in Education*, *5*. https://doi.org/10.3389/feduc.2020.00063

Kim, S. (2006). A Comparative Study of IRT Fixed Parameter Calibration Methods. *Journal of Educational Measurement*, *43*(4), 355–381. https://doi.org/10.1111/j.1745-3984.2006.00021.x

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (Third Edition). *Statistics for Social and Behavioral Sciences*. New York: Springer.

Lienert, G. A. (Ed.) (1998). *Testaufbau und Testanalyse* (6. Auflage, Studienausgabe). Weinheim: Beltz Verlagsgruppe. http://www.content-select.com/index.php?id=bib_view&ean=9783621278454

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement*, *17*(3), 179–193. http://www.jstor.org/stable/1434833

Marco, G. L. (1977). Item Characteristic Curve Solutions To Three Intractable Testing Problems. *Journal of Educational Measurement*, *14*(2), 139–160. https://doi.org/10.1111/j.1745-3984.1977.tb00033.x

Maydeu-Olivares, A. (2013). Goodness-of-Fit Assessment of Item Response Theory Models. *Measurement: Interdisciplinary Research & Perspective*, *11*(3), 71–101. https://doi.org/10.1080/15366367.2013.831680

Murphy, K. R., & Myors, B. (1999). Testing the hypothesis that treatments have negligible effects: Minimum-effect tests in the general linear model. *Journal of Applied Psychology*, *84*(2), 234–248. https://doi.org/10.1037/0021-9010.84.2.234

NEPS-Netzwerk (2022). *Nationales Bildungspanel, Scientific Use File der Startkohorte Klasse 5*. Bamberg. Retrieved from https://doi.org/10.5157/NEPS:SC3:12.0.0

OECD (2014). *PISA 2012 Technical Report.* Paris, Frankreich.

Pohl, S., & Carstensen, C. H. (2012). *NEPS Technical Report - scaling the data of the competence tests* (No. 14). https://www.neps-data.de/Portals/0/Working%20Papers/WP_XIV.pdf

Pohl, S., & Carstensen, C. H. (2013). Scaling of competence tests in the National Educational Panel Study - many questions, some answers, and further challenges. *Journal for Educational Research Online*, *5*, 189–216. https://doi.org/10.25656/01:8430

Pohl, S., Haberkorn, K., & Carstensen, C. H. (2015). Measuring Competencies across the Lifespan - Challenges of Linking Test Scores. In M. Stemmler, A. von Eye, & W. Wiedermann (Eds.), *Springer Proceedings in Mathematics & Statistics. Dependent Data in Social Sciences Research* (Vol. 145, pp. 281–308). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-20585-4_12

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago: MESA Press.

Schnittjer, I., & Gerken, A.-L. (2017). *NEPS Technical Report for mathematics - scaling results of Starting Cohort 3 in seventh grade* (NEPS Survey Papers No. 16). Bamberg. https://www.neps-data.de/Portals/0/Survey%20Papers/SP_XVI.pdf

Schwarz, G. (1978). Estimating the Dimensions of a Model. *The Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Shrout, P. E., & Rodgers, J. L. (2018). Psychology, Science, and Knowledge Construction: Broadening Perspectives from the Replication Crisis. *Annual Review of Psychology*, *69*(1), 487–510. https://doi.org/10.1146/annurev-psych-122216-011845

Stocking, M. L., & Lord, F. M. (1983). Developing a Common Metric in Item Response Theory. *Applied Psychological Measurement*, *7*(2), 201–210. https://doi.org/10.1177/014662168300700208

Svetina, D., Crawford, A. V., Levy, R., Green, S. B., Scott, L., Thompson, M., . . . Kunze, K. L. (2013). Designing small-scale tests: A simulation study of parameter recovery with the 1-PL. *Psychological Test and Assessment Modeling*, *55*(4), 335–360.

Vale, C. D. (1986). Linking item parameters onto a common scale. *Applied Psychological Measurement*, *10*(4), 333–344. https://doi.org/10.1177/014662168601000402

Van der Linden, W. J., & Barrett, M. D. (2016). Linking item response model parameters. *Psychometrika*, *81*(3), 650–673. https://doi.org/10.1007/s11336-015-9469-6

Von Davier, A. A., Carstensen, C. H., & von Davier, M. (2006). Linking competencies in educational settings and measuring growth. *ETS Research Report Series*, *2006*(1), 36. https://doi.org/10.1002/j.2333-8504.2006.tb02018.x

Anhang A

Im Folgenden werden zwei Linkmethoden dargestellt, deren Einsatz in 2PLM und 3PLM etabliert ist, sich aber aus verschiedenen Gründen nicht im 1PLM anbietet.

Mean/sigma Linkmethode. Die weiter oben in Gleichung (4) dargestellte lineare Transformation ist eine Vereinfachung der Gleichung

$$\theta^*_{i,t_2} = \mathrm{A}\theta_{i,t_2} + \mathrm{B}, \tag{12}$$

welche im 2PLM und 3PLM gilt (Kolen & Brennan, 2014). A stellt hierbei eine multiplikative Konstante dar. Da sich A im Falle der mean/mean Linkmethode aus dem Quotienten der Mittelwerte der Itemdiskriminationsparameter zweier zu verlinkender Messzeitpunkte berechnet und deshalb wegen der Fixierung von $a = 1$ im 1PLM immer in eins resultiert, ergibt sich Gleichung (4) bei der Anwendung der mean/mean Linkmethode im 1PLM. In der mean/sigma Linkmethode berechnet sich A (hier dargestellt für das Ankeritem-Design) jedoch aus

$$\mathrm{A} = \frac{SD\left(b_{j,t_1}\right)}{SD\left(b_{j,t_2}\right)}, \tag{13}$$

also aus den Standardabweichungen der Ankeritemschwierigkeitsparameter der Messzeitpunkte t_1 und t_2 und würde zumeist in $a \neq 1$ resultieren. Da die lineare Transformation auf alle Parameter (also auch auf den üblicherweise auf 1 fixierten Parameter a) angewandt werden muss, ergibt sich

$$a^*_{j,t_2} = \mathrm{A}a_{j,t_2} + \mathrm{B}. \tag{14}$$

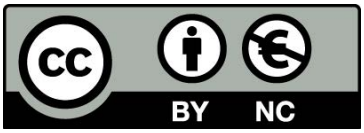
a^*_{j,t_2} wäre also identisch für jedes Item, würde jedoch von 1 abweichen. Somit wären die beiden verlinkten Skalen von t_1 und t_2 inhaltlich nicht mehr direkt vergleichbar.

Characteristic Curve Methoden. Diese zwei – fast identischen - schätzintensiven Methoden nach Haebara (1980) und Stocking & Lord (1983), versuchen (eingesetzt im 1PLM) jene additive Konstante B zu finden, welche die Differenz zwischen den Lösungswahrscheinlichkeiten p_{ij}, welche aus der unabhängig skalierten und der verlinkten Testform stammen, minimiert. Als Ausgangswert für den Schätzprozess kann z. B. die additive Konstante B aus der mean/mean Linkmethode dienen.

Anhang B

Copyright for manuscripts

Fischer, L., Gnambs, T., Rohm, T., & Carstensen, C. H. (2019). Longitudinal linking of Rasch-model-scaled competence tests in large-scale assessments: A comparison and evaluation of different linking methods and anchoring designs based on two tests on mathematical competence administered in grades 5 and 7. *Psychological Test and Assessment Modeling*, *61*(1), 37-64. URL: https://psycnet.apa.org/record/2019-21145-003



Fischer, L., Rohm, T., Carstensen, C. H., & Gnambs, T. (2021). Linking of Rasch-scaled tests: Consequences of limited item pools and model misfit. *Frontiers in psychology*, *12*, 633896. https://doi.org/10.3389/fpsyg.2021.633896

Fischer, L., Rohm, T., Gnambs, T., & Carstensen, C. H. (2016). Linking the data of the competence tests. *NEPS Survey Paper*, *1*.

Psychological Test and Assessment Modeling, Volume 61, 2019 (1), 37-64

Longitudinal linking of Rasch-model-scaled competence tests in large-scale assessments: A comparison and evaluation of different linking methods and anchoring designs based on two tests on mathematical competence administered in grades 5 and 7

Luise Fischer[1,2], *Timo Gnambs*[1,3], *Theresa Rohm*[1,2] & *Claus H. Carstensen*[2]

Abstract

Measuring growth in an item response theory framework requires aligning two tests on a common scale known as longitudinal linking. So far, no consensus exists regarding the appropriate method for the linking of longitudinal data scaled according to the Rasch model in large-scale assessments. Therefore, an empirical study was conducted within the German National Educational Panel Study to identify appropriate linking methods for the comparison of competencies across time. The study examined two anchoring designs based either on anchor-items or an anchor-group and three linking methods (mean/mean linking, fixed parameters calibration, and concurrent calibration). Two tests on mathematical competence were administered to a sample of n = 3,833 German students (48 % girls) in Grades 5 and 7. An independent link sample (n = 581, 53 % girls) drawn from the same population was administered both tests at the same time. The assumptions of unidimensionality were confirmed; differential item functioning was examined using effect-based hypotheses tests. Anchoring designs and linking methods were compared and evaluated using diverse criteria such as link error, mean growth rate estimation, and model fit. Overall, little differences among the linking methods and anchoring designs were found. However, mean growth was found to be significantly smaller in the anchor-group design.

Keywords: linking, item response theory, longitudinal, effect based hypotheses testing, competences

---

[1] *Correspondence concerning this article should be addressed to:* Luise Fischer, Educational Measurement, Leibniz Institute for Educational Trajectories, Wilhelmsplatz 3, 96047 Bamberg, Germany; email: luise.fischer@uni-bamberg.de

[2] Department of Psychology and Methods of Educational Research, University of Bamberg, Bamberg, Germany

[3] Institute for Education and Psychology, Johannes Kepler University Linz, Austria

Introduction

The measurement of an individual's growth in an item response theory (IRT) framework requires placing two tests on a common scale. This is referred to as longitudinal linking (A. von Davier, Carstensen, & M. von Davier, 2006). Therefore, linking data is an essential prerequisite for investigating educational trajectories. Most large-scale assessments (LSA) focus on differences between age cohorts such as the *Programme of International Student Assessment* (PISA), the *Trends in International Mathematics and Science Study* (TIMMS), the *Progress in International Reading Literacy Study* (PIRLS), or the American *National Assessment of Educational Progress* (NAEP). Only few LSAs allow for the study of an individual's change over time, for example, the German *National Educational Panel Study* (NEPS; Blossfeld, Roßbach, & von Maurice, 2011), the American *Early Childhood Longitudinal Program* (ECLS), or longitudinal extensions of PISA (e.g., Prenzel, Carstensen, Schöps, & Maurischat, 2006). Furthermore, unidimensional Rasch-type models as well as the more complex two-parametric and three-parametric logistic models (2PL and 3PL; Birnbaum, 1968) are used in the practice of vertical scaling of educational assessments (A. von Davier et al., 2006). However, the latter models that additionally include discrimination and guessing parameters are clearly more popular; other model parameterizations such as the difficulty-plus-guessing model (Kubinger & Draxler, 2006) have also been introduced but, as of yet, have not been frequently used in LSAs. So far, no consensus exists regarding the appropriate method for the linking of longitudinal data scaled according to the Rasch model in large-scale assessments. This study investigated whether certain linking methods, usually applied in 2PL or 3PL modeled cross-sectional data, can be transferred to Rasch-model-scaled longitudinal data. Moreover, these linking methods were compared and evaluated in different anchoring designs (anchor-items design and anchor-group design) using data from the NEPS.

Linking of Rasch-type models

In Rasch-type models it is assumed that the probability P of person n to correctly answer item i is conditioned on the interaction of two parameters (both of them being necessary and sufficient), that is, a person's ability β (which is not directly observable and, therefore, latent) and an item's difficulty parameter δ. In order to model ordered response categories in polytomous items, Masters (1982) developed a partial credit model (PCM):

$$P\left(X_{kni}=1|\beta_n,\delta_{ik}\right)=\frac{\exp\left(\beta_n-\delta_{ik}\right)}{1+\exp\left(\beta_n-\delta_{ik}\right)}, \tag{1}$$

where δ_{ik} is the difficulty of the k^{th} step in item i. In the special case of dichotomous data, the PCM reduces to the well-known Rasch model (Rasch, 1980):

$$P\left(X_{ni}=1|\beta_n,\delta_i\right)=\frac{\exp\left(\beta_n-\delta_i\right)}{1+\exp\left(\beta_n-\delta_i\right)}. \tag{2}$$

In Rasch-type models the person ability parameter β_n and the item difficulty parameter δ_i are both localized on a common "ability" scale. As the zero in this ability scale is set arbitrarily (i.e., depending on the parameter constraints), any statement on the change of a person's ability over a period of time needs to be based on data that is longitudinally linked (van der Linden & Barrett, 2016).

Anchoring designs

Due to a time-lag between test administrations in longitudinal educational assessments accompanied by a corresponding ability development, ability distributions will most likely differ when assessing the same sample repeatedly. Therefore, participants are regarded as non-equivalent groups in repeated measurements (A. von Davier et al., 2006). Thus, the procedure of linking longitudinal data requires an overlap of information between the two tests (Pohl, Haberkorn, & Carstensen, 2015). This information overlap is either achieved by identical items (i.e., common items) administered at both measurement points (anchor-items design) or by persons who answer items from both tests at the same measurement point (anchor-group design; Vale, 1986; see Figure 1). If a common item in an anchor-items design meets several conditions (see below), it can serve as a link item. A. von Davier and colleagues (2006) refer to the same design in a cross-sectional context as a design for non-equivalent groups with anchor test (NEAT). Linking in the NEAT design is generally referred to as vertical linking (A. von Davier et al., 2006). Though longitudinal linking and vertical linking differ in name depending on the adopted data collection design (longitudinal versus cross-sectional), they do not differ conceptually, when the samples are non-equivalent groups. As such, both approaches are concerned with practical issues when it comes to linking (e.g., Seock-Ho Kim & A. Cohen, 1992; Kolen & Brennan, 2014). However, linking based on longitudinal designs

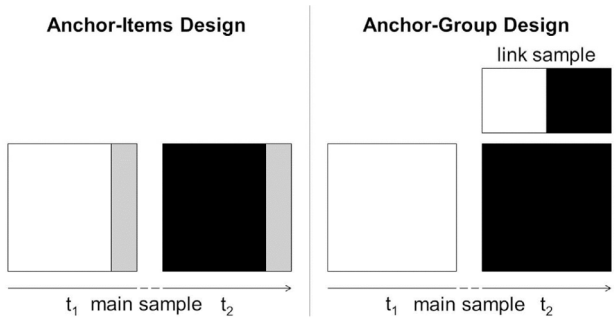


Figure 1:
Anchoring Designs for Longitudinal Linking. Each rectangle constitutes one measurement point. While white and black rectangles represent test specific items, grey rectangles represent common items between the tests. t1 = first measurement point; t2 = second measurement point

faces additional challenges due to participants' motivations and panel dropout that require reduced test lengths, potentially leading to a decreased accuracy in parameter estimation. As such, the absolute number in link items as well as the link items' estimation accuracy may (drastically) differ among longitudinal and vertical linking.

For the anchor-group design an overlap of information is achieved through an independent link sample. Participants of the link sample need to be sampled from the same population as the main sample (i.e., the sample a researcher is primarily interested in). The mean age of the link sample should correspond to the age of the main sample at t_1 or t_2 or should fall somewhere between the age groups of the two measurement occasions (Pohl et al., 2015). Thus, the link sample is administered both tests at the same measurement point. Therefore, the participant's answers on the items are not influenced by longitudinal ability development. Thus, the item difficulty parameters represent an unaltered relationship of the two tests. In this anchoring design, no common items are necessary.

With respect to our question concerning longitudinal designs where the same participants are assessed repeatedly, the choice of an anchoring design depends, amongst others, on test length, potential memory effects as well as repetition effects in link items and the expected amount of change in the latent construct between two measurements affecting item difficulty. Domains using content-based items (e.g., reading literacy) are more prone to memory and repetition effects than domains using number- and operation-based items (e.g., mathematical literacy). Although an anchor-group design causes additional costs and the resulting link information is afflicted with an additional sampling error, it may still be the preferable choice depending on the measured construct. In terms of the validity of a link (i.e., the extent to which the link information represents the various facets of the underlying construct) the anchor-group design is more comprehensive than the anchor-items design since all items (i.e., both tests completely) contribute to the link information. Also, when test length is limited and change in the latent ability is expected to be large, an anchor-group design may be preferable regarding the reliability of each measurement point. As reliability is increasing the more closely the test difficulty and the person ability distribution match, no item position is occupied by a common item that potentially provides only little information due to fitting the ability of the sample poorly at the second measurement point.

IRT linking methods

A linking method translates the link information in order to put the parameters from two (or more) tests on a common scale (Vale, 1986). The selection of a linking method is codetermined by the anchoring design: While an anchor-group design is less common in practice and thus, only a small number of corresponding linking methods have been suggested, a wide selection of established linking methods is available for the anchor-items design / NEAT design (see M. von Davier and Carstensen, 2006 for an overview and Kolen & Brennan, 2014, for an elaborated introduction). Some of the most popular IRT linking methods are the mean/mean method (Loyd & Hoover, 1980), mean/sigma method (Marco, 1977), the characteristic curve methods (Haebara, 1980; Stocking & Lord, 1983), fixed parameters calibration, and concurrent calibration. Also hybrid approaches such as combin-

ing concurrent calibration with fixed parameters calibration (e.g., PISA cycle of 2015; Organisation for Economic Co-operation and Development (OECD), 2017) or characteristic curve methods (Briggs & Weeks, 2009) have been used in LSAs. All but the concurrent calibration use separate calibrations in each sample before transforming the item and person parameters. Thus, an already established scale (e.g., the scale from the first measurement point) serves as a reference scale. This may be attractive for example in longitudinal designs where the focus lies on measuring change and a reference scale has already been implemented due to sequentially published data. The following section describes three IRT linking methods: mean/mean linking, fixed parameters calibration, and concurrent calibration that are applicable in Rasch-type models (i.e., discrimination parameters retain their fixed value of one) and thus, were examined in this study in more detail.

Mean/mean method based on the anchor-items design (m/m$_{AID}$)

In this method the item difficulty parameter estimates δ_i (Rasch, 1980) of the link items are used for computing two scaling constants, slope *A* and intercept *B,* to shift scale *Y* to the reference scale *X* (Loyd & Hoover, 1980):

$$\delta_Y^* = A\delta_Y + B \tag{3}$$

The scaling constants based on the anchor-items design (AID) are computed from the link items as

$$A_{\text{AID}} = M(\alpha_{Y\text{link}}) / M(\alpha_{X\text{link}}) \tag{4}$$

with $M(\alpha_{Y\text{link}})$ and $M(\alpha_{X\text{link}})$ being the means of the link item discrimination parameters from scales *Y* and *X* and as

$$B_{\text{AID}} = M(\delta_{X link}) - A^* M(\delta_{Y\text{link}}) \tag{5}$$

with $\delta_{X link}$ = difficulty estimates of the link items of scale *X* and $\delta_{Y link}$ = difficulty estimates of the link items of scale *Y*. The discrimination parameter of the linked scale is then obtained by $\alpha_Y^* = \alpha_Y / A$. In Rasch models, mean/mean linking always results in $A = 1$ and therefore, the scale is shifted without changing the distribution of the item difficulty estimates[4]. Therefore, (5) reduces to

$$B_{\text{AID}} = M(\delta_{X\text{link}}) - M(\delta_{Y\text{link}}). \tag{6}$$

[4] In the mean/sigma method the slope of the linear scale transformation is computed as $A = SD(\xi_{Ylink}) / SD(\xi_{Xlink})$ using the standard deviations of the link item difficulty parameters from scales *Y* and *X,* typically resulting in $A \neq 1$. Consequently, the discrimination parameter $\alpha_Y^* = \alpha_Y / A$ is changed which would violate the basic assumption of the Rasch model that assumes constant discriminations of 1. Consequently, this linking method was excluded from further investigation in the present study.

To establish the link, *B* is added to each item difficulty parameter of the scale intended to link. In doing so, the item difficulty parameters of scale *Y* are shifted on the logit scale in such a way that the mean difficulty of the link items of both scales are equal. This procedure has no influence on the relation of item difficulty parameters within scale *Y*.

As the characteristic curve methods only differ with regard to the estimation of the scaling constants *A* and *B* the basic concept of the linking approach is identical to the moments approach (i.e., mean/mean linking). Previous studies in cross-sectional contexts found rather small differences in parameter accuracy for the two estimation approaches (e.g., Seonghoon Kim & Kolen, 2006). Consequently, the characteristic curve methods were excluded from further investigation in the present study.

Mean/mean method based on the anchor-group design (m/m$_{AGD}$)

To link scale *Y* to the reference scale *X* using an anchor-group design (AGD), the principle of the mean/mean method using anchor-items can be adapted by including the information of the link sample, which provides the unaltered relationship of the scales *X* and *Y* (as was described in the previous section). In contrast to the anchor-items design the link information is based on the entire test including all items. As in the anchor-items design the computation of $A_{\text{AGD}} = M(\alpha_Y) / M(\alpha_X)$ always results in 1. Adapting (6) for the anchor-group design results in

$$B_{\text{AGD}} = M\left(\delta_X\right) - M\left(\delta_Y\right) + \left(M\left(\delta_{Y,LS}\right) - M\left(\delta_{X,LS}\right)\right) \tag{7}$$

with $M(\delta_X)$ and $M(\delta_Y)$ being the mean item difficulties of the scales *X* and *Y* for the main sample and $M(\delta_{X,LS})$ and $M(\delta_{Y,LS})$ being the respective means of the link sample. To establish the link, *B* is added to each item difficulty parameter of the scale intended to be linked. As in the anchor-items design, this procedure has no influence on the relation of item difficulty parameters within scale *Y*.

Because the mean/mean method (regardless of the underlying anchoring design) is based on a linear transformation, all difficulty estimates are shifted equally on the logit scale. Strictly speaking, the mean/mean method has no additional constraints, because the logit scale is invariant to linear transformations (Rasch, 1980). Thus, model fit is not influenced by the mean/mean method. As van der Linden and Barrett (2016) correctly point out, this shifting constant is always based on an arbitrarily chosen constraint that may (or may not) approximate the true parameters. In any case, no verification of this constraint is possible when using empirical data.

Fixed parameters calibration (FPC) based on the anchor-items design

The item difficulty parameters of the link items from the reference scale are fixed in the separate calibration of the scale intended to be linked. Thus, identical difficulty parameters of link items (anchored to the reference scale) result in both scales. This strict constraint may lead to a decrease in model fit in the linked scale due to possible differential item functioning (DIF) and due to sampling error whereas the model fit of the reference scale is not affected. A. von Davier and colleagues (2006) point out that this procedure is not advisable if two populations significantly differ in ability when taking two test forms. Transferring this thought to a longitudinal measurement would lead to the conclusion that the method of fixed parameters would not be advisable when large cognitive development is expected.

Concurrent calibration (CC) based on the anchor-items design

Both tests are scaled jointly in a concurrent analysis where each measurement loads on a single dimension. Items included in both tests are constrained to have identical item parameters in both samples. This quite strict one-step procedure strives for the golden mean between the two tests. Compared to calibrating both tests separately, some limitations in model fit have to be accepted on both tests due to possible DIF. Still, calibrating both tests concurrently seems a promising approach with regard to estimation efficiency (Jodoin, Keller, & Swaminathan, 2003) as well as the reduction of sampling error (Hanson & Beguin, 2002).

Previous findings on vertical linking

Though a lot of research has been done comparing IRT linking methods in the field of vertical linking the findings provide only little clarity on the suitability of linking methods (Arai & Mayekawa, 2011; Jodoin et al., 2003; Seock-Ho Kim & A. Cohen, 1992, 1992; Lei & Zhao, 2012; Tong & Kolen, 2007). This may be due to the vast variety of manipulated factors examined in studies that potentially influence the link outcome such as a) linking methods, b) anchoring design, c) type of data (empirical versus simulated), d) characteristics of common items (proportion within a test, range of item difficulties, dichotomous or polytomous items, DIF), e) test length, f) sample characteristics (size, motivation to participate such as high- versus low-stakes tests), g) test targeting, h) number of measurement points i) time gap/developmental progress between measurements, j) underlying IRT models, and k) violation of model assumptions (e.g., unidimensionality assumption). However, with the huge number of experimental conditions and comparisons drawn from different evaluation criteria, it seems rather difficult to disentangle the prior findings and to rank order the linking methods. Nevertheless, one effect that is consistently reported in literature is that increasing the sample size and the number of anchor items improves the link performance – regardless of the linking method. Some authors (Hanson & Beguin, 2002; Lei & Zhao,

2012) found that concurrent calibration resulted in smaller error than separate calibration. This effect was explained by Hanson and Beguin (2002) with the increased sample size in concurrent calibration compared to separate calibration. While Arai and Mayekawa (2011) suggested that the ratio of common items should exceed 10 % in order to not worsen the performance of concurrent calibration and fixed parameters calibration, Kolen and Brennan (2014) recommended a share of (at least) 20 %. Using empirical data Jodoin et al. (2003) found that separate calibration (mean/sigma method) resulted in less mean growth than concurrent calibration and fixed parameters calibration.

However, empirical studies comparing IRT linking methods on longitudinal data scaled with the Rasch model in LSAs are still missing. Therefore, the present empirical study aims at comparing and evaluating linking methods that fit the assumptions of Rasch-type models (i.e., mean/mean, fixed parameters calibration, concurrent calibration). Moreover, it aims at comparing linking methods based on two different anchoring designs (i.e., anchor-items design, anchor-group design) on the same data. As such, conclusions on the comparability of the link process and link outcome of linking methods based on the two anchoring designs could be drawn. In particular, the results of the present study will extend previous findings on vertical linking to the longitudinal context and complement research on two-parametric and three-parametric models by focusing on Rasch-type measurement models.

Method

Sample

We selected a panel sample (i.e., main sample) from the NEPS (Blossfeld, Roßbach, & von Maurice, 2011), which is a LSA based on a longitudinal design conducted in Germany. In the NEPS, participants from different age cohorts are followed up and are periodically administered low-stakes competence tests in various domains in order to measure competence development over the life span. In the present study, a total of n = 3,833 participants (48 % girls, 95 % born in Germany, 51 % attending high school), sampled representatively from schools across all 16 federal states, received a mathematics competence test in Grade 5 (age: M = 10.91, SD = .52) and Grade 7 (age: M = 12.91, SD = 0.52)[5]. Moreover, from the same population (but sampled from different schools) as the panel sample n = 581 participants (53 % girls, 93 % born in Germany, 44 % attending high school) attending Grade 7 (age: M = 13.08, SD = 0.59) were additionally sampled as independent link sample.

The study was approved by the Federal Ministries of Education in Germany and the data protection board of the National Educational Panel Study. Informed consent was given by parents, students, and educational institutions to take part in the study. Data from the

[5] Note that 1,360 of the initially 5,193 participants in Grade 5 did not take part in the measurement in Grade 7 and were thus, excluded from the analyses in the present study.

panel sample are available from http://www.neps-data.de for researchers who meet the criteria for access to confidential data. Data from the independent link sample are not accessible due to legal reasons.

Instruments

The conceptual framework underlying the mathematics tests administered in the NEPS is described in Neumann et al. (2013). Prior to test administration several pilot studies were conducted to guarantee that the final test form reflected the intended conceptual framework. As such, test development for the respective math tests in Grades 5 and 7 was theory driven, based on a Rasch-model-conforming unidimensional mathematical literacy concept. Additionally, the psychometric quality and fit to the Rasch model were empirically checked throughout test construction as well as for the final test forms, which were administered to the panel sample and the independent link sample.

The mathematics tests administered in Grades 5 and 7 included 24 items (marginal reliability = .80; Adams, 2005) and 23 items (marginal reliability = .76), respectively. In each test one item was polytomous, whereas the rest were dichotomous. Six dichotomous items were common to both tests and served as potential link items. As such, the number of common items corresponded to the recommended share of 20 % (Kolen & Brennan, 2014) in the literature. These common items were selected by educational experts on mathematics for broadly covering the underlying conceptual framework. Furthermore, these six items were expected to fit the anticipated change in ability between Grades 5 and 7 well. In order to prevent position effects (e.g., Hohensinn, Kubinger, Reif, Holocher-Ertl, Khorramdel, & Frebort, 2008; Trendtel & Robitzsch, 2018), all six common items retained their original position (see Tables 2 and 3) within each test from Grade 5 to Grade 7. Additionally, violation of local independence was checked to detect possible interaction effects with measurement point-unique items. To minimize the risk of memory effects the items reflected typical tasks administered in math classes at school. Thus, it was unlikely that students were able to remember correct solutions for these items across a time span of two years.

As the mathematics tests were not administered in a high-stakes setting, missing values were not handled as incorrect responses (Pohl & Carstensen, 2013). Consequently, if a participant gave no response, the answer was treated as missing (and not as incorrect). On average, participants had $M = 1.8$ ($SD = 2.4$) missing values in Grade 5 and $M = 0.7$ ($SD = 1.4$) missing values in Grade 7. The participants were tested at school in a group setting with a limited test time of 30 minutes per measurement occasion. For a detailed description of the scaling results see Duchhardt and Gerdes (2012) as well as Schnittjer and Gerken (2017).

Study Design

Both, anchor-items design and anchor-group design (see Figure 1) were combined in this study: While participants from the panel sample took the two mathematics tests with a time-lag of two years between Grades 5 and 7, participants of the link sample took both tests at one measurement point in Grade 7. To avoid memory and other effects in the link sample the six common items were included only in the Grade 7 test. In order to account for item position and test length effects the common items were replaced by new items of similar content and difficulty in Grade 5.

Statistical Analyses

All data were scaled using the PCM (Masters, 1982), which is an extension of the Rasch model to polytomous items applying item-specific rating scales. For linking methods based on separate calibration (i.e., mean/mean linking based on anchor-items and anchor-group design as well as FPC) each measurement occasion was scaled separately constraining the mean ability to zero while the linking was conducted afterwards. Applying the concurrent calibration we modelled our data using a two-dimensional PCM, setting the mean ability to zero at Grade 5 (dimension 1) and estimating the mean ability of Grade 7 (dimension 2). In line with Andersen (1985), we assumed that the difference in mean ability between Grades 5 and 7 represented the change of ability in the longitudinal panel sample. The software used was ACER ConQuest 4 (Adams, Wu, & Wilson, 2016) based on a marginal maximum likelihood estimation (Bock & Aitkin, 1981), in order to accommodate the partially missing responses. Note, that contemporary IRT software is unable to handle the present data when based on a conditional maximum likelihood estimation (Fischer & Molenaar, 2012).

As any empirical data can never fully meet the strict assumptions of a theoretical model such as the Rasch model, statistical tests will always discard a model if only the sample size is big enough. As a consequence, we assessed model fit using the weighted mean square (WMNSQ; Wright & Masters, 1982), its respective *t*-value and the corrected item-total correlation. The WMNSQ is a quantitative measure of fit discrepancy. It is based on the weighted deviation of an actual person's response from Rasch model expectation. Being distributed as mean squares, the expected value is 1 (Bond & Fox, 2015). In assessing model fit we adopted rules of thumb proposed in the literature (Pohl & Carstensen, 2012) and viewed a WMNSQ > 1.2 and a respective *t*-value > |8| as considerable item misfit. Note, that a well item fit according to the WMNSQs indicates that items of a test discriminate sufficiently at the various person ability levels, thus, meeting the respective specification in the Rasch model. For the corrected item-total correlation a value greater than .2 was deemed acceptable. Local independence on the item level was evaluated based on Yen's Q_3 (1993) statistic, indicating no substantial violation for values < |.20|. Moreover, visual comparisons of the observed and model-implied item characteristic curves were conducted to identify potentially misfitting items.

Examination of assumptions for longitudinal linking

In order to link adjacent measurement points, several assumptions have to be met. Tests and link items have to meet the assumptions of unidimensionality and must not show DIF (Pohl et al., 2015; A. von Davier et al., 2006). Common items that do not meet these assumptions should not be used as link items and may be modelled as group specific (unique) item parameters (e.g., Oliveri & M. von Davier, 2011).

Unidimensionality

To measure competence development within a domain over a period of time, the underlying theoretical construct must not change between time points. The unidimensionality assumption was examined twofold. First, a test can be considered essentially unidimensional when the standardized residuals of a one-dimensional model exhibit approximately zero-order correlations. While in case of an anchor-items design the residuals were derived from a one-dimensional model of the two separately scaled tests, in case of an anchor-group design the residuals were derived from a one-dimensional model of the two concurrently scaled tests that were administered in the link sample. Second, further evidence of a unidimensional scale is given if the ratio of the first two eigenvalues derived from the standardized residuals does not exceed 1.5 (Smith Jr, 2002).

Differential item functioning

The localization of the linked scale is determined by the resulting link information. Consequently, the person ability estimation (and as such the magnitude of the participant's ability change between two measurement points) is influenced by the link information. In order to not mix up change in person ability and drift in item difficulty, the link item parameters $\hat{\delta}_{Xi\text{link}}$ and $\hat{\delta}_{Yi\text{link}}$ must not change (i.e., retain their relative position on the logit scale) between two test administrations. DIF was examined using a Wald test that compares the estimated item difficulties resulting from a maximum likelihood estimation (Draba, 1977):

$$t_{XY} = \frac{\hat{\delta}_{Xi\text{link}} - \hat{\delta}_{Yi\text{link}}}{\sqrt{SE\left(\hat{\delta}_{Xi\text{link}}\right)2 + SE\left(\hat{\delta}_{Yi\text{link}}\right)2}}. \tag{8}$$

The resulting test statistic is t distributed. LSAs often have to deal with excessive test power due to large sample sizes. Consequently, the result of statistical tests becomes less meaningful. Instead of a classical null hypothesis (Cohen, 1994), Murphy and Myors (1999) suggested using a minimum effect hypothesis. Here, the critical value is not defined by an assumed difference of zero but by a proportion of variance accounted for. We followed the Educational Testing Service determining the critical value (Zieky,

1993) as 1.54 % variance accounted for to identify relevant deviations of item difficulty parameters between two groups.

When using an anchor-group design, DIF is examined among the two groups of main sample and link sample, applying the same procedure as for the anchor-items design.

Evaluation of linking methods

The three linking methods and two anchoring designs were evaluated with regard to three criteria:

Link error

The link error becomes relevant when comparisons are made between ability estimates of different measurement points. It is conceptualized as standard error (*SE*) of differences between the separately scaled and linked item difficulty parameters of the link items from the test intended to be linked:

$$SE = SD_{Y,Y^*} / \sqrt{k} \tag{9}$$

with $SD_{Y,Y}$* = standard deviation of the link item parameter differences from the separately scaled scale *Y* and the linked scale *Y*,* and *k* = the number of link items (adapted from PISA 2009 Technical Report; OECD, 2012 and PISA 2012 Technical Report; OECD, 2014). When an anchor-group design is used all items are handled as link items. The standard error of differences is then calculated as standard error of differences between the main sample and the link sample for each test and is pooled afterwards. An adapted approach is necessary for the computation of the link error emerging from a CC based on an anchor-items design. In contrast to m/m and FPC where the link item estimates are only changed in the latter measurement point, the link item estimates are changed in both measurement points when using a concurrent calibration. Therefore, the amount of change in link item estimates is split among the two measurement points by leaving the number of link items unchanged. In order to avoid counting the number of link items double, *k* was halved. To account for the standard deviation of differences in link items twice (once for each measurement point X and Y), the link error had to be pooled. For the concurrent calibration the link error was then computed as

$$SE = \sqrt{\left(\frac{SD_{X,X^*}}{\sqrt{\frac{k}{2}}}\right)^2 + \left(\frac{SD_{Y,Y^*}}{\sqrt{\frac{k}{2}}}\right)^2} \tag{10}$$

As such, when analysing mean differences of a group including at least two time points, the link error has to be considered by including it into the pooled *SE* (for further details

see Organisation for Economic Co-operation and Development, 2014). Consequently, a larger link error contributes to a reduced test power. Furthermore, the link error can be understood as bias, concerning every participant equally. Therefore, the standard deviation of ability scores is not affected by the link error.

Mean growth rate estimation

Since the linking methods are based on different link information, they vary in their estimation of the mean growth rate which reflects the estimated mean change in participant's ability between the two test administrations. However, because our research was based on empirical data where the true change is unknown, a potential bias in the link results cannot be further investigated. For the separately scaled models (based on m/m_{AID}, m/m_{AGD} and FPC) the mean growth rate estimate was obtained by a "post hoc" two-dimensional analysis where each test administration (i.e., Grades 5 and 7) loaded on a single dimension. The mean ability of the first test administration served as a reference category (i.e., it was fixed to zero). Due to the preceding link procedure each difficulty parameter was estimated in prior analyses and, thus, fixed to these values. The mean growth rate was computed as the mean change in the weighted maximum likelihood ability estimate (WLE; Warm, 1989) using the examinee response vector and the item parameters.

Model fit

After linking the two measurement points, we fitted a two-dimensional model for each of the linked data that constrained the item parameters to the previously derived and linked values (see above). This intermediate step was necessary to make the model fits and information criteria (Akaike information criterion (AIC); Akaike, 1974 and Bayesian information criterion (BIC); Schwarz, 1978) of the separately scaled models (based on m/m_{AID}, m/m_{AGD} and FPC) and the concurrently scaled model (using concurrent calibration) comparable in order to evaluate how the different restrictions inherent to the different linking methods effected the model fit.

Results

A PCM was used to analyze the panel sample and link sample. Model identification was obtained by constraining the mean ability to zero. For the panel sample the mean item difficulty estimates of the separately scaled mathematics tests applied in Grades 5 and 7 were M = -0.63 (SD = 1.11, Min = -2.74, Max = 1.44) and M = -0.58 (SD = 1.01, Min = -3.13, Max = 1.19), respectively. The latent correlation of the Mathematical competences was r = .93 (p = .00) across the two measurement points. For the concurrently scaled link sample the mean item difficulty estimates of Grades 5 and 7 were M = -1.16 (SD = 0.90, Min = -2.58, Max = 1.03) and M = -0.35 (SD = 0.99, Min = -2.93, Max = 1.50). Overall,

Table 1:
Item Fit of Separately Scaled Mathematics Tests in Grades 5 and 7 for Panel Sample and Link Sample

		Percentage correct	Item difficulty	*SE*	WMNSQ	*t*	r_{it}	Yen's Q_3
panel sample grade 5	*M (SD)*	60.66 (19.11)	-0.63 (1.11)	0.05 (0.01)	1.00 (0.05)	-0.07 (2.90)	0.37 (0.07)	0.00 (0.03)
	Min/Max	23.34/91.30	-2.74/1.44	0.04/0.07	0.92/1.14	-5.80/9.30	0.25/0.47	-0.06/0.37
panel sample grade 7	*M (SD)*	59.39 (16.94)	-0.58 (1.01)	0.05 (0.01)	1.00 (0.07)	-0.06 (4.17)	0.40 (0.08)	0.00 (0.02)
	Min/Max	27.68/93.25	-3.13/1.19	0.04/0.07	0.88/1.16	-8.4/10.10	0.24/0.55	-0.06/0.12
link sample grade 5	*M (SD)*	70.54 (15.61)	-1.16 (0.90)	0.12 (0.01)	1.01 (0.07)	0.20 (1.42)	0.37 (0.08)	0.00 (0.05)
	Min/Max	30.89/89.86	-2.58/1.03	0.11/0.15	0.89/1.14	-2.10/3.30	0.24/0.53	-0.14/0.33
link sample grade 7	*M (SD)*	55.62 (17.17)	-0.35 (0.99)	0.11 (0.02)	0.99 (0.07)	-0.14 (1.60)	0.40 (0.07)	0.00 (0.05)
	Min/Max	22.80/92.39	-2.93/1.50	0.11/0.17	0.88/1.15	-2.60/3.60	0.26/0.51	-0.14/0.14

Note. n = 3,833 (panel sample) and n = 581 (link sample); item difficulty = location parameter; *SE* = Standard error of difficulty / location parameter; WMNSQ = Weighted mean square; t = t-value for WMNSQ: comparing *Min/Max* among panel sample and link sample leads to the conclusion that mere sample size rather than actual item misfit was responsible for the difference in t-values between the two samples (Smith, Rush, Fallowfield, Velikova, & Sharpe, 2008); r_{it} = Corrected item-total correlation; Yen's Q_3 = Yen's (1993) corrected Q_3: statistic tests for violation of local independence assumption if Q_3 values > |.20|;

Table 2:
Difficulty Estimates of the Separately, Concurrently and Linked Scaled Grade 5-Test

No.	Item	Panel sample	Link sample	Linked Estimates			
				m/m		FPC	CC
				m/m_{AGD}	m/m_{AID}		
1	Item 1	-0.51	-1.06	≙ PS	≙ PS	≙ PS	-0.51
2	Item 2	-1.15	-1.32	≙ PS	≙ PS	≙ PS	-1.15
3	Item 3	-0.92	-1.42	≙ PS	≙ PS	≙ PS	-0.92
4	Item 4	0.86	0.24	≙ PS	≙ PS	≙ PS	0.86
5	Item 5	-0.17	-1.74	≙ PS	≙ PS	≙ PS	-0.17
6	Item 6	0.38	-0.03	≙ PS	≙ PS	≙ PS	0.38
7[a]	Item 7	0.49	-	≙ PS	≙ PS	≙ PS	0.49*
8	Item 8	-1.98	-1.58	≙ PS	≙ PS	≙ PS	-1.98
9[a]	Item 9	-2.72	-	≙ PS	≙ PS	≙ PS	-2.58*
10[a]	Item 10	-0.69	-	≙ PS	≙ PS	≙ PS	-0.85*
11	Item 11	-0.86	-1.42	≙ PS	≙ PS	≙ PS	-0.86
12	Item 12	1.44	1.03	≙ PS	≙ PS	≙ PS	1.44
13	Item 13	-0.22	-0.78	≙ PS	≙ PS	≙ PS	-0.22
14	Item 14	-1.33	-2.10	≙ PS	≙ PS	≙ PS	-1.33
15	Item 15	-1.55	-1.40	≙ PS	≙ PS	≙ PS	-1.55
16	Item 16	-2.19	-2.29	≙ PS	≙ PS	≙ PS	-2.19
17	Item 17	-0.53	-1.28	≙ PS	≙ PS	≙ PS	-0.53
18	Item 18	-0.23	-1.11	≙ PS	≙ PS	≙ PS	-0.23
19[a]	Item 19	0.11	-	≙ PS	≙ PS	≙ PS	0.18*
20	Item 20	-1.27	-1.65	≙ PS	≙ PS	≙ PS	-1.27
21[a]	Item 21	0.92	-	≙ PS	≙ PS	≙ PS	0.98*
22	Item 22	-2.74	-2.58	≙ PS	≙ PS	≙ PS	-2.74
23[a]	Item 23	-0.41	-	≙ PS	≙ PS	≙ PS	-0.43*
24	Item 24	0.22	-0.46	≙ PS	≙ PS	≙ PS	0.21
$M_{\text{all items}}$		-0.63 (1.11)	-	≙ PS	≙ PS	≙ PS	-0.62 (1.10)
$M_{\text{link items excluded}}$		-0.71 (1.07)	-1.16 (0.90)	≙ PS	≙ PS	≙ PS	-0.71 (1.07)
$M_{\text{link items}}$		-0.38 (1.28)	-	≙ PS	≙ PS	≙ PS	-0.37 (1.26)

Note. n = 3833 (panel sample) and n = 581 (link sample). The difficulty estimates for the linking methods m/m (based on anchor-items as well as on an anchor-group) and FPC match the independently scaled Grade 5 test, because they do not change the reference scale. The CC changes the reference scale and therefore, the resulting difficulty estimates differ from those of Grade 5. The six common items in the link sample were replaced by new items which were excluded from analyses. The mean difficulty estimates include the respective standard deviation in parentheses M (SD). Grades 5 and 7 were modelled unidimensional in the link sample. No. = item order in test administration; PS = panel sample; m/m_{AID} = mean/mean method based on anchor-items design; m/m_{AGD} = mean/mean method based on anchor-group design; FPC = fixed parameters calibration; CC = concurrent calibration.
[a]link item; *parameter was constrained

Table 3:
Difficulty Estimates of the Separately, Concurrently and Linked Scaled Grade 7-Test

				Linked Estimates			
				m/m			
No.	Item	Panel sample	Link sample	m/m_{AGD}	m/m_{AID}	FPC	CC
1	Item 25	-0.36	-0.07	0.32	0.36	0.39	0.39
2	Item 26	0.50	0.62	1.19	1.22	1.25	1.25
3	Item 27	0.21	0.36	0.89	0.93	0.96	0.96
4	Item 28	0.29	0.37	0.97	1.01	1.04	1.04
5[a]	Item 7	-0.27	-0.04	0.42	0.46	0.49*	0.49*
6	Item 29	-1.36	-0.98	-0.67	-0.63	-0.61	-0.60
7[a]	Item 9	-3.13	-2.93	-2.44	-2.40	-2.72*	-2.58*
8[a]	Item 10	-1.83	-1.41	-1.14	-1.10	-0.69*	-0.85*
9	Item 30	-0.52	-0.46	0.16	0.20	0.23	0.23
10	Item 31	0.24	0.39	0.93	0.97	0.99	1.00
11	Item 32	-0.65	-0.50	0.04	0.08	0.10	0.11
12	Item 33	-1.83	-1.42	-1.14	-1.10	-1.08	-1.07
13	Item 34	-0.12	-0.03	0.57	0.61	0.63	0.63
14	Item 35	-1.82	-1.57	-1.13	-1.09	-1.07	-1.06
15	Item 36	-1.35	-1.29	-0.66	-0.62	-0.60	-0.59
16	Item 37	-0.15	0.26	0.54	0.58	0.60	0.60
17	Item 38	-0.39	-0.38	0.29	0.33	0.36	0.36
18[a]	Item 19	-0.50	-0.08	0.19	0.22	0.11*	0.18*
19	Item 39	0.66	1.05	1.35	1.39	1.41	1.41
20[a]	Item 21	0.28	0.38	0.97	1.01	0.92*	0.98*
21	Item 40	1.19	1.50	1.88	1.91	1.94	1.94
22[a]	Item 23	-1.22	-0.71	-0.53	-0.49	-0.41*	-0.43*
23	Item 41	-1.26	-1.09	-0.57	-0.53	-0.51	-0.50
$M_{all\ items}$		-0.58 (1.01)	-0.35 (0.99)	0.11 (1.01)	0.14 (1.01)	0.16 (1.03)	0.17 (1.02)
$M_{link\ items}$		-1.11 (1.24)	-0.80 (1.22)	-0.42 (1.24)	-0.38 (1.24)	-0.38 (1.28)	-0.37 (1.26)

Note. n = 3833 (panel sample) and n = 581 (link sample). Linking with m/m_{AGD}: constant $B_{anchor\text{-}group}$ = .681 (see (7)) is added to each parameter of PS. Linking with m/m_{AID}: constant $B_{anchor\text{-}items}$ = .726 (see (6)) is added to each parameter of PS. Using FPC or CC: constraints are set by anchoring or equalizing parameters (indicated by *). The mean difficulty estimates include the respective standard deviation in parentheses M (SD). Grades 5 and 7 were modelled unidimensional in the link sample. No. = item order in test administration; m/m_{AGD} = mean/mean method based on anchor-group design; m/m_{AID} = mean/mean method based on anchor-items design; FPC = fixed parameters calibration; CC = concurrent calibration.
[a]link item; *parameter was constrained

item fit was satisfactory (see Table 1; Pohl & Carstensen, 2012) as indicated by corrected item-total correlations (r_{it}) exceeding .23 and WMNSQ falling between 0.88 and 1.16 (Pohl & Carstensen, 2012). With Smith Jr's (2002) ratio test not exceeding 1.5 and the $Q3$ statistics falling between Min = -.14 and Max = .37, the unidimensionality assumption for both tests in the panel sample and link sample was supported and for all but two items[6] no violation of local independence was detected.

The item difficulty estimates of the panel sample and link sample of the separately, concurrently and linked scaled tests of Grades 5 and 7 are summarized in Tables 2 and 3.

Differential item functioning

DIF was examined between the common items (in the anchor-items design) and between the panel sample and link sample (in the anchor-group design). The difference in item difficulties between the six items administered at both measurement occasions and the results of the respective minimum effect hypotheses tests are summarized in Table 4.

Table 4:
Examination of Differential Item Functioning for the Six Common Items in the Anchor-Items Design

Link item	δ_{G5}	δ_{G7}	$\Delta\delta$	$SE_{\Delta\delta}$	t	F
Item 7	0.87	0.84	-0.03	0.06	-0.48	0.23
Item 9	-2.34	-2.02	0.32	0.09	3.33	11.09
Item 10	-0.30	-0.72	-0.42	0.06	-6.44	41.52
Item 19	0.49	0.61	0.12	0.06	2.03	4.12
Item 21	1.30	1.39	0.09	0.06	1.56	2.42
Item 23	-0.02	-0.11	-0.09	0.06	-1.39	1.93

Note. Item difficulty estimates (with their means set to zero) based on participants that took part in Grades 5 and 7 (N = 3,833). The t and F statistics resulted from a Wald test (see (8)). None of the common items exceeded the critical value of $F_{0154}(1, 3{,}831) = 88.3$ for a p = .05. Therefore, all items met the assumption of showing no substantial DIF and qualified as link items. $\Delta\delta$ = difference in item difficulty parameters between Grades 7 and 5 (positive values indicate easier items in Grade 5); $SE_{\Delta\delta}$ = pooled standard error; t = t statistic; F = F statistic.

[6] The residuals of the Grade 5 test-unique items 2 and 3 correlated at .37 in the panel sample and at .33 in the link sample. An inspection of the item content revealed that both items were somewhat similarly phrased (i.e., some words overlapped) and were presented one after another. However, visual checks of the respective item characteristic curves and an evaluation of the item level fit statistics for these items in the panel sample (WMNSQ: 0.98, 1.03; t-value: -0.9, 1.9; corrected item-total correlation: .43, .38) and the link sample (WMNSQ: 1.1, 1.02; t-value: 1.9, 0.4; corrected item-total correlation: .29, .38) did not identify a severe misfit. Therefore, both items were included in the final scaling procedure as intended by the test developers.

None of the resulting *F* statistics was significant (all *p*s > .05) and, as such, indicated no pronounced DIF qualifying the six common items as link items. Also, no relevant DIF was found between the panel sample and link sample (see Table 5). None of the *F* statistics indicated significantly ($p < .05$) different item parameters between the two samples.

Table 5:
Examination of Differential Item Functioning between Panel Sample and Link Sample in Grades 5 and 7

Grade 5					Grade 7				
Test	**item**	**$\Delta\delta$**	**$SE_{\Delta\delta}$**	***F***	**Test**	**item**	**$\Delta\delta$**	**$SE_{\Delta\delta}$**	***F***
G5	Item 1	0.09	0.12	0.64	G7	Item 25	-0.06	0.11	0.28
G5	Item 2	-0.29	0.13	5.39	G7	Item 26	0.12	0.11	1.02
G5	Item 3	0.04	0.13	0.11	G7	Item 27	0.08	0.11	0.51
G5	Item 4	0.16	0.11	2.06	G7	Item 28	0.16	0.11	1.86
G5	Item 5	1.12	0.13	72.42	G7	Item 7	0.00	0.11	0.00
G5	Item 6	-0.04	0.12	0.12	G7	Item 29	-0.14	0.12	1.45
G5	Item 7	-	-	-	G7	Item 9	0.03	0.19	0.03
G5	Item 8	-0.86	0.13	41.36	G7	Item 10	-0.18	0.13	2.06
G5	Item 9	-	-	-	G7	Item 30	0.17	0.11	2.16
G5	Item 10	-	-	-	G7	Item 31	0.08	0.11	0.54
G5	Item 11	0.11	0.13	0.65	G7	Item 32	0.08	0.11	0.53
G5	Item 12	-0.04	0.12	0.13	G7	Item 33	-0.18	0.13	1.96
G5	Item 13	0.11	0.12	0.90	G7	Item 34	0.14	0.11	1.52
G5	Item 14	0.31	0.14	4.77	G7	Item 35	0.02	0.13	0.02
G5	Item 15	-0.60	0.13	21.94	G7	Item 36	0.17	0.12	1.96
G5	Item 16	-0.36	0.15	5.35	G7	Item 37	-0.18	0.11	2.49
G5	Item 17	0.30	0.12	5.82	G7	Item 38	0.22	0.11	3.71
G5	Item 18	0.42	0.12	11.74	G7	Item 19	-0.19	0.11	2.80
G5	Item 19	-	-	-	G7	Item 39	-0.16	0.12	1.78
G5	Item 20	-0.08	0.14	0.31	G7	Item 21	0.14	0.12	1.48
G5	Item 21	-	-	-	G7	Item 40	-0.07	0.13	0.33
G5	Item 22	-0.62	0.17	13.67	G7	Item 23	-0.27	0.12	5.28
G5	Item 23	-	-	-	G7	Item 41	0.07	0.17	0.17
G5	Item 24	0.22	0.11	3.59					

Note. Item difficulty estimates based on the panel sample and the link sample. The *F* statistics resulted from the squared *t*-value of a Wald test (see (8)). None of the items exceeded the critical value of $F_{0154}(1, 4{,}412) = 99.2$ for $p = .05$. Therefore, no DIF was found. G5 = mathematics test in Grade 5; G7 = mathematics test in Grade 7; $\Delta\delta$ = difference in item difficulty parameters between panel sample and link sample (positive values indicate easier items in the link sample); $SE_{\Delta\delta}$ = pooled standard error; *t* = *t* statistic; *F* = *F* statistic.

Evaluation of linking methods

The results of the evaluation criteria for the three linking methods mean/mean (either based on an anchor-group or an anchor-items design), FPC and CC are summarized in Table 6.

Table 6:
Results of Linking Method Evaluation

Linking Method	Link Error	$\Delta\beta$	$Var(\Delta\beta)$	AIC	BIC	Parameters
m/m_{AGD}	0.11	0.68 (0.78)	0.76	190,964	191,295	53
m/m_{AID}	0.10	0.72 (0.83)	0.76	190,964	191,295	53
FPC	0.10	0.74 (0.85)	0.76	191,079	191,373	47
CC	0.10	0.75 (0.86)	0.76	191,023	191,317	47

N = 3,833. The link error was calculated using (9) and (10). m/m = mean/mean method (based either on the AID or AGD). Δβ = mean growth estimation in person ability parameters between Grades 5 and 7 in logits: positive values indicate a gain of ability between the measurement points (in parentheses: Cohen's d for repeated measures ANOVA; Morris & DeShon, 2002); $Var(\Delta\beta)$ = Variance of Change between Grades 5 and 7; AIC = Akaike's information criterion; BIC = Bayesian information criterion; Parameters = number of estimated parameters during scaling; m/m_{AGD} = mean/mean method based on anchor-group design; m/m_{AID} = mean/mean method based on anchor-items design; FPC = fixed parameters calibration; CC = concurrent calibration.

Link error

While the link errors for the methods m/m_{AID}, FPC and CC were equivalent (i.e., 0.10), the link error of m/m_{AGD} was slightly larger, i.e., 0.11. Note, that m/m_{AID} and FPC always result in perfectly matching link errors due to the fact that the calculations were based on the same difficulty estimates.

Mean growth rate

For Grades 5 and 7 the *SD* of ability estimates was identical among all linking procedures (SD_{G5} = 1.16, SD_{G7} = 1.24). Taking into account the high latent correlation in ability ($r_{G5,G7}$ = .93) between the Grades 5 and 7 it was not surprising that no substantial differences were found. As such, no evidence for the phenomenon of scale shrinkage (i.e., a reduction of sample variance induced trough IRT linking methods; see Briggs & Weeks, 2009) was found among the linking methods. The differences in the mean growth rates (m/m_{AGD}: $\Delta\beta$ = 0.68, m/m_{AID}: $\Delta\beta$ = 0.72, FPC: $\Delta\beta$ = 0.74, CC: $\Delta\beta$ = 0.75) were analyzed using a one-way repeated-measures ANOVA. Mauchly's test indicated that the assumption of sphericity had been violated ($\chi^2(5)$ = 46,30, p = .00); therefore, degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity (ε = .46). The results showed that the amount of growth was significantly effected by the

applied linking procedure $F(1.37, 5{,}266) = 180{,}393$, $p < .001$. As such, effect sizes for the differences in mean growth between Grades 5 and 7 (d_{RM}) among the linking procedures were calculated (see Morris & DeShon, 2002). Effect sizes resulted in m/m_{AGD}: $d_{RM} = 0.78$, m/m_{AID}: $d_{RM} = 0.83$, FPC: $d_{RM} = 0.85$ and CC: $d_{RM} = 0.86$. With the differences in effect sizes having a range of 0.08 (between m/m_{AGD} and CC), the difference in mean growth among the linking procedures was considered rather small. Still, while m/m_{AID}, FPC and CC form a homogenous group, m/m_{AGD} seemed a bit trailed off. For these differences directly trace back to the differences in difficulty estimates resulting from the different linking procedures, additional analyses were calculated. No significant difference between the mean difficulty estimates of the separately scaled Grade 5 test ($M = 0.63$, $SD = 1.11$; which equally represented the estimates of m/m_{AGD}, m/m_{AID} as well as FPC) and the concurrent calibration ($M = 0.62$, $SD = 1.10$, $t(23) = -0.34$, $p = .74$, $d_{RM} = 0.01$) was found (see Figure 2). Furthermore, difficulty estimates of Grade 7 (see Fig 3) were analyzed using a one-way repeated-measures ANOVA. Again, Mauchly's test indicated that the assumption of sphericity had been violated ($\chi^2(5) = 220.21$, $p = .00$); therefore, degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity ($\varepsilon = .34$). The results showed that the difficulty estimates were significantly effected by the linking procedure $F(1.00, 22.09) = 5.69$, $p < .03$, $\eta_p^2 = .21$. Post hoc tests using the Bonferroni correction revealed that m/m_{AGD} ($M = 0.11$, $SD = 0.21$) was signifi-

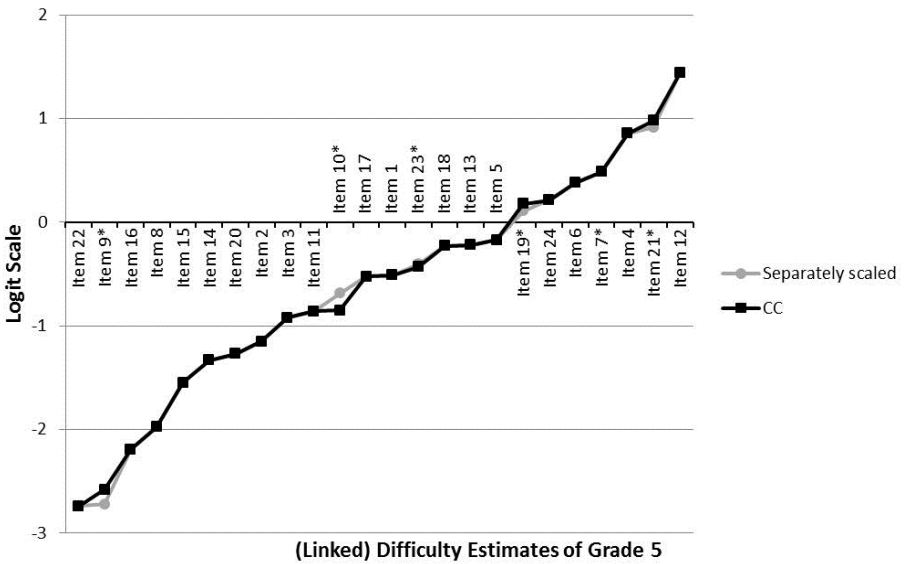


Figure 2:
Difficulty Estimates of Grade 5. Separately scaled = the separately scaled Grade 5 test equals the estimates of m/mAGD, m/mAID and FPC; CC = concurrent calibration; item difficulties are in ascending order; link items are denoted by *

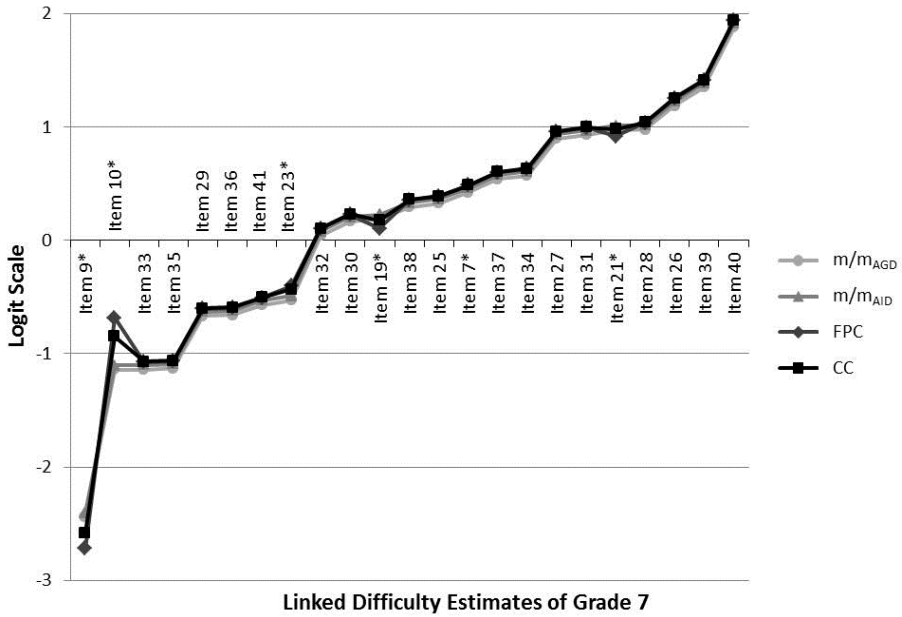


Figure 3:
Linked Difficulty Estimates of Grade 7. m/m(AGD) = mean/mean linking based on anchor-group design; m/m(AID) = mean/mean linking based on anchor-items design; FPC = fixed parameters calibration; CC = concurrent calibration; item difficulties are in ascending order; link items are denoted by *

cantly different from m/m$_{AID}$ (M = 0.14, SD = 0.21, p = .00, Cohen's d = -0.38) and CC (M = 0.17, SD = 0.21, p = .00, Cohen's d = -0.63) but not significantly different from FPC (M = 0.16, SD = 0.21, p = .18, Cohen's d = -0.55).

Model fit

The identical model fit of m/m$_{AGD}$ and m/m$_{AID}$ (Deviance = 190,857, number of parameters = 53, AIC = 190,964, BIC = 191,295) originated from their shared principle of linking methods. As mentioned above, model fit is not influenced when scaling using the mean/mean method (regardless of the anchoring design). Therefore, there is no point in comparing model fit among the mean/mean method and the other two linking methods in terms of building an evaluative rank order between them. Still, as the resulting model fit of the mean/mean method represents the cumulated model fit of the two separately scaled measurement points of Grades 5 and 7, it may serve as a general reference value for FPC (Deviance = 190,985, number of parameters = 47, AIC = 191,079, BIC = 191,373) and CC (Deviance = 190,928, number of parameters = 47, AIC = 191,023, BIC = 191,317).

However, it is no surprise that the information criteria both favored the CC over the FPC, given the method's constraints. Clearly, equalizing parameters leaves a model more space to fit the data than anchoring parameters to a fixed value.

Discussion

In this study we compared and evaluated the methods mean/mean linking (based on an anchor-items design (m/m_{AID}) as well as on an anchor-group design (m/m_{AGD}), fixed parameters calibration (FPC) and concurrent calibration (CC) on their performance to align two tests on a common scale. We applied the criteria of link error, mean growth rate estimation and model fit to evaluate the linking performance. The empirical data used in this study are based on participants that were administered two tests on mathematical literacy in Grades 5 and 7. In practice, the link information in LSA is typically either based on an anchor-group design or an anchor-items design. In contrast, the design of this study allowed a simultaneous comparison of both, anchoring designs as well as linking methods.

Overall, little differences among the linking methods were found. With the linking based either on a rather small absolute number of six link items (representing a proportion of 25 %) or a small link sample, measurement error and sampling error were less likely to cancel out. Of all evaluation criteria, differences among the linking methods and anchoring designs were most explicitly reflected by the mean growth. Though we found rather small differences in mean growth among the linking methods (in ascending order: m/m_{AID} < FPC < CC) this trend supported the findings of Jodoin et al. (2003) who reported less mean growth for methods based on a linear transformation (i.e., mean/sigma method) compared to FPC and CC using empirical data. Concluding from the findings of our more in-depths analysis it seems plausible to expect increasing differences among the linking methods the more subsequent measurement points are added. A bigger difference in mean growth was found between the anchoring designs. The significant differences in difficulty estimates of medium effect size between m/m_{AGD} and m/m_{AID} as well as m/m_{AGD} and CC probably resulted from the different sources of link information (i.e., either link sample or anchor items). Though the anchor-group design should result in a more valid link due to the bigger number of link items, it was also based on a smaller sample size and thus, more prone to sampling error. However, as little research exists in evaluating linking methods based on the anchor-group design more research is necessary to further investigate effects of age, sample size, characteristics of domain-specific development and the amount of time between measurement points.

Consistent with Hanson and Béguin (2002) and Lei and Zhao (2012) we found no difference in link error among the linking methods nor the two anchoring designs. Rather small differences in model fit criteria reflected the model's constraints as expected. Furthermore, the linking methods showed no substantial influence on the sample variance.

As no DIF in link items was found among Grades 5 and 7 in the panel sample, as well as among the panel sample and the link sample in Grade 7 we concluded that there was no substantial memory effect in repeatedly administered link items. As such, no effect was

found on the response behavior of the students to answer a math item they had already worked on two years ago.

In contrast to the suggestion of A. von Davier and colleagues (2006) no evidence was found that FPC was not advisable when two populations significantly differed in (mean) ability when taking two test forms. However, as intraindividual change over time was very homogenous in our sample (with barely any change in rank order), there were only little differences in ability distributions among Grades 5 and 7.

Overall, the present case study found few differences between the examined linking methods. This suggests that the estimation of competence development is not profoundly effected by the methodological choices adopted for scaling the results. However, for the interpretation of these results one needs to keep in mind that they are based on a rather specific setting: longitudinal comparisons between Grades 5 and 7 for mathematical competencies among German students. It is unclear to what degree these findings extend to, for example, other populations, content domains, or age groups. Therefore, the generalizability of the presented results needs to be explored in further research that evaluates the robustness of linking methods applied to Rasch-model-scaled longitudinal data in different settings.

Limitations of the study

Since our analyses are based only on two measurement points, effects may accumulate over measurement points when adding subsequent measurements. This urges the necessity of sticking to an already applied linking method when linking data of more than two measurement occasions to avoid change in competence development being influenced by a change of linking methods. Furthermore, dropout rate is an issue in longitudinal designs (see Zinn & Gnambs, 2018). For various reasons participants drop out of the sample and cannot be reached anymore. Therefore, refreshing the sample periodically is necessary to perpetuate a proper sample size. Especially in the context of institutional education (e.g., school, university) the remaining sample after dropout typically represents a positive selection of participants. As a consequence, DIF in item parameters has to be examined between both groups of participants defined by taking part in one or two measurement points. In this study, 1,360 from 5,193 participants took only the test in Grade 5 and did not take part in Grade 7. The mean ability of these 1,360 participants is 0.47 logits lower than the mean of the participants that were to stay in the sample. Hence, future research is challenged with the question if and how linking methods differ in their ability estimation when applied in extended samples (i.e. samples including also 'cross-sectional' participants).

Though various procedures are discussed in the literature to provide statistically guided help to identify reasonable link items (e.g., Bechger & Maris, 2015) the authors were not aware of a solution to overcome the more general issue of identification in Rasch-type models. We therefore opted for a construct-driven decision procedure for selecting suitable link items from the common items.

Conclusion

As memory effects in items become more likely with repeated administration the anchor-group design seems a conceivable alternative to the anchor-items design in longitudinal measurements. Despite the overall small effects found, we join Hanson and Béguin (2002) in their advice to compare various linking methods as this enables the researcher to examine differences in linking methods (albeit the true parameter is never known when analyzing empirical data) but also serves as a reminder that the link result is based on an arbitrarily chosen link information (e.g., van der Linden & Barrett, 2016).

Author Note

This research was supported in part by grants of the Deutsche Forschungsgemeinschaft awarded to Claus H. Carstensen.

This paper uses data from the National Educational Panel Study (NEPS): Starting Cohort 3–Grade 5 and Grade 7, NEPS:SC3:5.0.0. From 2008 to 2013, NEPS data were collected as part of the Framework Programme for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research (BMBF; https://www.bmbf.de). As of 2014, the NEPS survey is carried out by the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg in cooperation with a nationwide network. The authors declare that they have no competing interests.

References

Adams, R. J. (2005). Reliability as a measurement design effect. *Studies in Educational Evaluation*, *31*, 162–172. https://doi.org/10.1016/j.stueduc.2005.05.008

Adams, R. J., Wu, M. L., & Wilson, M. R. (2016). *ConQuest*. Camberwell: ACER. Retrieved from https://www.acer.edu.au/conquest/acer-conquest1

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723. https://doi.org/10.1109/TAC.1974.1100705

Andersen, E. B. (1985). Estimating latent correlations between repeated testings. *Psychometrika*, *50*, 3–16. https://doi.org/10.1007/BF02294143

Arai, S., & Mayekawa, S.-i. (2011). A comparison of equating methods and linking designs for developing an item pool under item response theory. *Behaviormetrika*, *38*, 1–16. https://doi.org/10.2333/bhmk.38.1

Bechger, T. M., & Maris, G. (2015). A statistical test for differential item pair functioning. *Psychometrika*, *80*, 317–340. https://doi.org/10.1007/s11336-014-9408-y

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability.In F. M. Lord, & M. R. Novick (Eds.), *Statistical Theories of Mental Test Scores* (pp. 397-479). Reading, England: Addison-Wesley.

Blossfeld, H.-P., Roßbach, H.-G., & von Maurice, J. (Eds.). (2011). *Zeitschrift für Erziehungswissenschaft Sonderheft: Vol. 14*. *Education as a lifelong process: The German National Educational Panel Study (NEPS)*. Wiesbaden, Germany. VS Verlag für Sozialwissenschaften.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*, 443–459. https://doi.org/10.1007/BF02293801

Bond, T., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ:Routledge.

Briggs, D. C., & Weeks, J. P. (2009). The impact of vertical scaling decisions on growth interpretations. *Educational Measurement: Issues and Practice*, *28*, 3–14. https://doi.org/10.1111/j.1745-3992.2009.00158.x

Cohen. (1994). The earth is round (p <.05). *American Psychologist*, *49*, 997–1003. https://doi.org/10.1037/0003-066X.49.12.997

Draba, R. E. (1977). The identification and interpretation of item bias. *Research Memorandum*, *25*.

Duchhardt, C., & Gerdes, A. (2012). *NEPS Technical Report for mathematics – scaling results of Starting Cohort 3 in fifth grade* (NEPS Working Paper). Bamberg, Germany: University of Bamberg, National Educational Panel Study.

Fischer, G. H., & Molenaar, I. W. (2012). *Rasch models: Foundations, recent developments, and applications*: Springer.

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research*, *22*, 144-149;. https://doi.org/10.4992/psycholres1954.22.144

Hanson, B. A., & Beguin, A. A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement*, *26*, 3–24. https://doi.org/10.1177/0146621602026001001

Hohensinn, C., Kubinger, K. D., Reif, M., Holocher-Ertl, S., Khorramdel, L., & Frebort, M. (2008). Examining item-position effects in large-scale assessment using the Linear Logistic Test Model. *Psychology Science Quarterly*, *50*, 391.

Jodoin, M. G., Keller, L. A., & Swaminathan, H. (2003). A comparison of linear, fixed common item, and concurrent parameter estimation equating procedures in capturing academic growth. *The Journal of Experimental Education*, *71*, 229–250. https://doi.org/10.1080/00220970309602064

Kim, Seock-Ho, & Cohen, A. (1992). Effects of linking methods on detection of DIF. *Journal of Educational Measurement*, *29*, 51–66.

Kim, Seonghoon, & Kolen, M. J. (2006). Robustness to format effects of IRT linking methods for mixed-format tests. *Applied Measurement in Education*, *19*, 357–381.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (Third Edition). *Statistics for Social and Behavioral Sciences*. New York, NY: Springer.

Kubinger, K. D., & Draxler, C. (2006). A comparison of the Rasch model and constrained item response theory models for pertinent psychological test data. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models: Extensions and applications* (pp. 295–312). New York, NY: Springer.

Lei, P.-W., & Zhao, Y. (2012). Effects of vertical scaling methods on linear growth estimation. *Applied Psychological Measurement*, *36*, 21–39. https://doi.org/10.1177/0146621611425171

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement*, *17*, 179–193. https://doi.org/10.1111/j.17453984.1980.tb00825.x

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement*, *14*, 139–160. https://doi.org/10.1111/j.1745-3984.1977.tb00033.x

Masters. (1982). A rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174. https://doi.org/10.1007/BF02296272

Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods*, *7*, 105–125. https://doi.org/10.1037/1082-989X.7.1.105

Murphy, K. R., & Myors, B. (1999). Testing the hypothesis that treatments have negligible effects: Minimum-effect tests in the general linear model. *Journal of Applied Psychology*, *84*, 234–248. https://doi.org/10.1037/0021-9010.84.2.234

Neumann, I., Duchhardt, Christoph, Grüßing, M., Heinze, A., Knopp, E., & Ehmke, T. (2013). Modeling and assessing mathematical competence over the lifespan. *Journal for Educational Research Online / Journal Für Bildungsforschung Online*, *5*, 80–109.

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.

Organisation for Economic Co-operation and Development. (2012). *PISA 2009 Technical Report.* Paris, France: Author. Retrieved from http://site.ebrary.com/lib/alltitles/docDetail.action?docID=10595644

Organisation for Economic Co-operation and Development. (2014). *PISA 2012 Technical Report*. Paris, France: Author.

Organisation for Economic Co-operation and Development. (2017). *PISA 2015 Technical Report*. Paris, France: Author.

Pohl, S., & Carstensen, C. H. (2012). *Scaling the data of the competence tests (NEPS technical report 14).* Bamberg, Germany: University of Bamberg, National Educational Panel Study.

Pohl, S., & Carstensen, C. H. (2013). Scaling of competence tests in the National Educational Panel Study-Many questions, some answers, and further challenges/Skalierung der Kompetenztests im Nationalen Bildungspanel-Viele Fragen, einige Antworten und weitere Herausforderungen. *Journal for Educational Research Online*, *5*, 189.

Pohl, S., Haberkorn, K., & Carstensen, C. H. (2015). *Measuring competencies across the lifespan - challenges of linking test scores*. In M. Stemmler, A. von Eye, & W. Wiedermann (Eds.), Dependent Data in Social Sciences Research (pp. 281-308). Springer.

Prenzel, M., Carstensen, C. H., Schöps, K., & Maurischat, C. (2006). Die Anlage des Längsschnitts bei PISA 2003. In *PISA Konsortium Deutschland (Ed.). PISA 2003: Untersuchungen zur Kompetenzentwicklung im Verlauf eines Schuljahres* (pp. 29–62). Münster, Germany: Waxmann.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: Mesa Press.

Schnittjer, I., & Gerken, A.-L. (2017). *NEPS Technical Report for mathematics - scaling results of Starting Cohort 3 in seventh grade* (NEPS Survey Papers No. 16). Bamberg, Germany: University of Bamberg, National Educational Panel Study.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461–464. https://doi.org/10.1214/aos/1176344136

Smith, A. B., Rush, R., Fallowfield, L. J., Velikova, G., & Sharpe, M. (2008). Rasch fit statistics and sample size considerations for polytomous data. *BMC Medical Research Methodology*, *8*, 33. https://doi.org/10.1186/1471-2288-8-33

Smith Jr, E. V. (2002). Understanding Rasch measurement: Detecting and evaluating the impact of multidimenstionality using item fit statistics and principal component analysis of residuals. *Journal of Applied Measurement, 3,* 205-231.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7,* 201–210. https://doi.org/10.1177/014662168300700208

Tong, Y., & Kolen, M. J. (2007). Comparisons of methodologies and results in vertical scaling for educational achievement tests. *Applied Measurement in Education*, *20*, 227–253.

Trendtel, M., & Robitzsch, A. (2018). Modeling item position effects with a Bayesian item response model applied to PISA 2009–2015 data. *Psychological Test and Assessment Modeling*, *60*, 241–263.

Vale, C. D. (1986). Linking item parameters onto a common scale. *Applied Psychological Measurement*, *10*, 333–344. https://doi.org/10.1177/014662168601000402

Van der Linden, W., & Barrett, M.-D. (2016). Linking item response model parameters. *Psychometrika*, *81*, 650–673. https://doi.org/10.1007/s11336-015-9469-6

von Davier, A., Carstensen, C. H., & von Davier, M. (2006). Linking competencies in educational settings and measuring growth. *ETS Research Report Series*, *2006*, 36. https://doi.org/10.1002/j.2333-8504.2006.tb02018.x

von Davier, M., & Carstensen, C. H. (Eds.). (2006). *Multivariate and mixture distribution Rasch models: Extensions and applications*. New York: Springer.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*, 427–450. https://doi.org/10.1007/BF02294627

Wright, B. D., & Masters. (1982). *Rating scale analysis*. Chicago, Il: Mesa Press.

Yen, W. M. (1993). Scaling performance assessments: strategies for managing local item dependence. *Journal of Educational Measurement*, *30*, 187–213. https://doi.org/10.1111/j.1745-3984.1993.tb00423.x

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. W. H. H. Wainer (Ed.), *Differential item functioning* (pp. 337–347). Hillsdale, NJ: Lawrence Erlbaum.

Zinn, S., & Gnambs, T. (2018). Modeling competence development in the presence of selection bias. *Behavior Research Methods*, *50*, 2426–2441. https://doi.org/10.3758/s13428-018-1021-z

frontiers in Psychology

ORIGINAL RESEARCH
published: 06 July 2021
doi: 10.3389/fpsyg.2021.633896

Linking of Rasch-Scaled Tests: Consequences of Limited Item Pools and Model Misfit

*Luise Fischer[1,2], Theresa Rohm[1,2], Claus H. Carstensen[2] and Timo Gnambs[1]**

[1] Leibniz Institute for Educational Trajectories, Bamberg, Germany, [2] Psychological Methods of Educational Research, University of Bamberg, Bamberg, Germany

In the context of item response theory (IRT), linking the scales of two measurement points is a prerequisite to examine a change in competence over time. In educational large-scale assessments, non-identical test forms sharing a number of anchor-items are frequently scaled and linked using two– or three-parametric item response models. However, if item pools are limited and/or sample sizes are small to medium, the sparser Rasch model is a suitable alternative regarding the precision of parameter estimation. As the Rasch model implies stricter assumptions about the response process, a violation of these assumptions may manifest as model misfit in form of item discrimination parameters empirically deviating from their fixed value of one. The present simulation study investigated the performance of four IRT linking methods—fixed parameter calibration, mean/mean linking, weighted mean/mean linking, and concurrent calibration—applied to Rasch-scaled data with a small item pool. Moreover, the number of anchor items required in the absence/presence of moderate model misfit was investigated in small to medium sample sizes. Effects on the link outcome were operationalized as bias, relative bias, and root mean square error of the estimated sample mean and variance of the latent variable. In the light of this limited context, concurrent calibration had substantial convergence issues, while the other methods resulted in an overall satisfying and similar parameter recovery—even in the presence of moderate model misfit. Our findings suggest that in case of model misfit, the share of anchor items should exceed 20% as is currently proposed in the literature. Future studies should further investigate the effects of anchor item composition regarding unbalanced model misfit.

Keywords: Rasch model, item response theory, linking methods, model misfit, anchor- items design, limited item pools

OPEN ACCESS

Edited by:
Pei Sun,
Tsinghua University, China

Reviewed by:
Ze Lu,
McMaster University, Canada
Jorge N. Tendeiro,
Hiroshima University, Japan

***Correspondence:**
Timo Gnambs
timo.gnambs@lifbi.de

Specialty section:
This article was submitted to Educational Psychology, a section of the journal Frontiers in Psychology

Received: *26 November 2020*
Accepted: *14 June 2021*
Published: *06 July 2021*

Citation:
Fischer L, Rohm T, Carstensen CH and Gnambs T (2021) Linking of Rasch-Scaled Tests: Consequences of Limited Item Pools and Model Misfit. Front. Psychol. 12:633896. doi: 10.3389/fpsyg.2021.633896

INTRODUCTION

Investigating differences between groups that were administered non-identical test forms in an item response theory (IRT) framework requires aligning two (or more) test forms onto a common scale, which is known as linking (Kolen and Brennan, 2014). As the process of linking requires an overlap of information among scales, this is frequently achieved by using an anchor-items design (Vale, 1986, p. 333–344), where test forms share a number of common items. Linking is a common procedure in the context of large-scale assessments (LSA) in educational measurement such as the

Programme of International Student Assessment (PISA) or the *American National Assessment of Educational Progress* (NAEP), which are characterized by large item pools and sample sizes. As such, LSAs provide an appropriate field for the application of 2-parameter logistic (2PL) and 3-parameter logistic (3PL) models (Birnbaum, 1968, p. 397–472) as a basis for scaling and linking the data. In contrast, in contexts which are characterized by a limited pool of items and small to medium sample sizes (as often is the case in studies with restricted economical resources or longitudinal designs) the sparser Rasch (1960) model is a suitable alternative (Sinharay and Haberman, 2014, p. 23–35). As of yet, the linking of Rasch-scaled data in this specific context was rarely researched.

In this article, we systematically investigate the linking of Rasch-scaled data based on limited item pools and small to medium sample sizes. To mimic applied settings, the data simulation mirrored a longitudinal design similar to the German *National Educational Panel Study* (NEPS; Blossfeld et al., 2011). Although mean change in a longitudinal design is often larger than differences among groups in a cross-sectional design, the linking is conceptually equivalent (von Davier et al., 2006). More specifically, the present simulation study deals with the issues of comparing and evaluating the performance of four IRT linking methods and investigating the absolute and relative number of anchor items required in these contexts. Moreover, as strict assumptions are made on equal item slopes in the Rasch model that are hardly met in empirical data, the robustness of linking methods toward model-data misfit is investigated.

In the following sections, we describe the Rasch model, the four common IRT linking methods, as well as challenges inherent to linking with limited item pools and sample sizes. Next, we describe the set-up of the simulation study and report the present findings. Finally, we discuss implications and limitations of our results.

THE RASCH MODEL

In the Rasch (1960) model, it is assumed that the probability P of person $n \in 1 \ldots N$ to correctly answer a dichotomous item $i \in 1 \ldots I$ is conditioned on the interaction of two parameters, that is, a person's ability β_n and an item's difficulty δ_i on a latent continuum:

$$P(X_{ni} = 1 \mid \beta_n, \delta_i) = \frac{\exp(\beta_n - \delta_i)}{1 + \exp(\beta_n - \delta_i)}. \quad (1)$$

Compared to 2PL and 3PL models, no parameter for item discrimination α_i is directly incorporated. Therefore, a higher precision in (anchor) item difficulties can be obtained at smaller sample sizes (Thissen and Wainer, 1982, p. 397–412) in the Rasch (1960) model.

Every item i, belonging to a test form fitting a Rasch model, measures the same latent construct with equal item discriminations α_i at all levels of β. Stated differently, items are not allowed to differ in their power to discriminate among persons (Wright, 1977, p. 97–116) and, thus, an irrevocable rank order among individuals $\beta_1 \ldots < \beta_n < \ldots \beta_N$ is determined based on the sufficient statistics of the person sum scores. As it can be challenging for empirical data to fully meet this strict specification, the question is *not* whether the data does or does not fit to a model, but is rather a "matter of degree" (Meijer and Tendeiro, 2015). As the weighting by α_i of person sum scores is ignored in case of Rasch model-data misfit (i.e., $\alpha_i \neq 1$), sample mean and variance estimates of the latent variable might be biased (Humphry, 2018, 216–228) as they are based on (1). Additionally, the precision of (anchor) item difficulties decreases (Thissen and Wainer, 1982, p. 397–412).

IRT LINKING METHODS

In IRT, only individual proficiencies and item difficulties located on equally defined scales are directly comparable over different measurement occasions (Kolen and Brennan, 2014). As such, prior to investigating proficiency development or group differences in an IRT framework, it is required to align two (or more) test forms onto a common scale (e.g., using an anchor-items design). As anchor item parameters are assumed to be measurement invariant and, thus, to maintain their difficulty over time, they allow for displaying an individual's change in proficiency. Several IRT linking methods exist, differentially "translating" the linking information during the linking process. The present study focuses on IRT linking methods compatible with Rasch-type models (van der Linden and Hambleton, 2013) that preserve uniform item discrimination parameters across the linked scales (Fischer et al., 2019, p. 37–64). The different linking methods scale the different test forms either separately or concurrently. In separate calibration methods, anchor item difficulty parameters of each test form are estimated prior to the linking process. This subsequently extracted link information is then implemented uniquely by each linking method. Hence, a once established reference scale remains unchanged throughout the course of measurement. In the present section, the three different calibration methods (1) fixed parameter calibration (Kim, 2006, p. 355–381), (2) mean/mean linking (Loyd and Hoover, 1980, p. 179–193), and (3) weighted mean/mean linking (van der Linden and Barrett, 2016, p. 650–673) are shortly described. Additionally, (4) a one-step approach of simultaneously calibrating and concurrently linking all test forms (e.g., Kim and Cohen, 1998, p. 131–143) is presented.

Fixed Parameter Calibration (FPC)

The parameter of anchor item $l \in 1L$ with $L \subseteq I$ of test form A intended to link are fixed using the estimated item parameters of the referencing test form B:

$$\delta_{Al} = \delta_{Bl}, \quad (2)$$

leaving no possibility for differences in anchor item parameters. Test forms based on a longitudinal design that vary in their sets of anchor items are linked sequentially (i.e., after test form t_2 is linked to t_1, t_3 is linked to t_2 and so on).

Mean/Mean Linking (m/m)

To link test form *A* to test form *B* and, therefore, obtain the linked item difficulty parameters δ^*_{Ai}, the linking constant *v* is added to each item δ_{Ai}:

$$\delta^*_{Ai} = \delta_{Ai} + v; \quad (3)$$

with *v* being the difference of the means of the *anchor item* difficulty parameters δ_{AL} and δ_{BL}:

$$v = M(\delta_{BL}) - M(\delta_{AL}). \quad (4)$$

After the linking results that $M(\delta^*_{AL}) = M(\delta_{BL})$.

Weighted Mean/Mean Linking (wm/m)

This approach incorporates estimation precision in weighting the anchor item difficulty parameter estimates by the inverse of their squared standard errors, $SE^{-2}_{\delta_{Al}}$ and $SE^{-2}_{\delta_{Bl}}$, prior to conducting a mean/mean linking, replacing *v* with

$$v' = \frac{\left(\sum_{l=1}^{L} \delta_{Bl} SE^{-2}_{\delta_{Bl}}\right)}{\left(\sum_{l=1}^{L} SE^{-2}_{\delta_{Bl}}\right)} - \frac{\left(\sum_{l=1}^{L} \delta_{Al} SE^{-2}_{\delta_{Al}}\right)}{\left(\sum_{l=1}^{L} SE^{-2}_{\delta_{Al}}\right)}. \quad (5)$$

As such, the precision of the anchor item difficulty estimates of test forms *A* and *B* is taken into account, aiming at reducing the link error (i.e., a reflection of the uncertainty introduced to the link due to the selection of link items). In other words, *v'* is identical to *v* when the anchor item difficulty parameter estimates have equal standard errors within a test form. Hence, weighted mean/mean linking is expected to outperform mean/mean linking when anchor items differ in precision.

Concurrent Calibration (CC)

All test forms are scaled concurrently in a one-step estimation procedure, constraining the anchor item difficulties across time points. As such, anchor item difficulties are simultaneously fitted to best meet the characteristics of all measurement points interacting with the samples' proficiency distributions.

Imprecision of (anchor) item difficulty estimates is reflected in their increased standard error (*SE*). In order to minimize estimation imprecision in item and person parameter estimates at *each time point*, a sample's proficiency and a test's difficulty should considerably overlap (i.e., also known as test targeting). In other words, the mean and variance of some test items' difficulty should closely fit the proficiency distribution of a respective sample. Of course, this claim is also true for sets of anchor items. Since sets of anchor items are administered repeatedly, they are expected to fit *several* proficiency distributions simultaneously. Consequently, the more diverging these proficiency distributions are, the more wide-spread a section of the latent scale needs to be covered by the sets of anchor items. It is to be noted that anchor items located at the outer edges of these joint ability distributions are prone to an increased *SE*. Svetina et al. (2013, p. 335–360) reported that a mismatch between item and person parameter distributions (i.e., if the item difficulties are, on average, too easy or too difficult as compared to the average proficiency distribution of the sample) impacted the recovery of item difficulty parameters more than the person parameter estimates. As such, linking methods that do not derive the linking information from the item level may be more "forgiving" with respect to imprecise estimates, as they are more likely to cancel out. As was shown by van der Linden and Barrett (2016, 650–673), the linking result of wm/m was superior to m/m in situations when anchor items did not perfectly display the samples' ability distribution. Therefore, the estimated amount of change is expected to be closer to its true value, compared to a result that is based on linking methods that link on the item level. Consequently, the method of weighted mean/mean linking that accounts for possible imprecisions in difficulty estimates by weighting anchor items by their *SE*s is expected to outperform the linking methods mean/mean linking, concurrent calibration and fixed parameter calibration (in the given order).

CHALLENGES FOR THE LINKING OF RASCH-SCALED DATA

Model-Data Misfit

There is a rather limited body of research examining the influence of Rasch model-data misfit on linking results. For example, Zhao and Hambleton (2017, p. 484) showed that in an LSA context with large sample sizes (N = 50,000) and long tests (78 items) with many anchor items (k = 39) fixed parameter calibration was more sensitive to model misfit and more robust against sizable ability shifts (up to 0.5 logits) as compared to linking methods that preserve the relation between item difficulty parameters during linking (i.e., mean/sigma method; Marco, 1977, and the characteristic curve methods; e.g., Stocking and Lord, 1983). As such, model fit was crucial to the appropriate use of FPC. So far, no research investigated the sensitivity and reactivity of IRT linking methods toward model misfit under more realistic conditions with smaller samples and shorter tests. Following Zhao and Hambleton (2017, p. 484), we hypothesized that FPC would be more sensitive toward model misfit as compared to CC, whereas m/m and wm/m would be least affected.

Number of Anchor Items

Kolen and Brennan (2014) formulated a rule of thumb for large item pools, proposing that the number of anchor items should make up about 20%. Nothing was stated for item pools consisting of less than 200 items. If a single anchor item would fully reflect the latent construct and was free of differential item functioning (DIF), this item would be sufficient for aligning two tests on a common scale. As this hardly is the case in practice, several anchor items are typically used in operational tests. Generally, a larger number of anchor items is assumed to reduce random link error and, thus, is expected to more precisely recover the true value of mean change. Moreover, a larger number of anchor items increase the content validity of the link. However, when test length is rather short (i.e., 25 items) and changes in proficiency between measurement points of a longitudinal sample are expected to be sizable (i.e., $\geq$0.25 logits; Zhao and Hambleton, 2017, p. 484), one repeatedly administered identical test form (i.e., 100% anchor items) would potentially affect test targeting and test reliability. In other words, when samples differ substantially in their mean proficiencies, the number of anchor

items in a short test form becomes a question of measurement precision at each measurement point. More precisely: An item's difficulty that matches a sample's mean ability well at the first measurement point t_1 cannot match a sample's mean ability well at the second measurement point t_2 when there was a significant change in the sample's ability between t_1 and t_2. Here is a demonstrative example: We assume that there is a significant change in ability of a sample that is administered two test forms with a length of 15 items sharing a number of 10 anchor items. We further assume that these 10 anchor items have a very good test targeting at t_1. From that follows that the test targeting of these 10 anchor items would have to be worse at t_2, affecting test reliability. Furthermore, administering items repeatedly may provoke memory effects that become more probable to emerge with an increasing number of anchor items. This leads to the question which proportion of anchor items can optimally balance measurement precision and linking information. Is the advice of a 20% anchor items share transferable to (rather) short test forms? In addition, questions about the minimum number of anchor items necessary to accurately display growth, and how model-data misfit interacts with the number of anchor items, remain.

To sum up, the present study aimed at comparing the performance of four common IRT linking methods (fixed parameter calibration, mean/mean linking, weighted mean/mean linking and concurrent calibration) based on Rasch-scaled simulated data. Particularly, we examined to what degree the number of anchor items and the degree of Rasch model-data misfit affected the linking for the different approaches.

METHODS

Data Generation

We simulated data for four time points (t_1–t_4) to measure within-individual growth in an anchor-items design (Vale, 1986, p. 333–344). The simulation was modeled after empirical data from the German National Educational Panel Study (NEPS; Blossfeld et al., 2011). The NEPS aims at measuring competence development over the life span. Therefore, respondents from different age cohorts (e.g., 10- or 15 years old) are followed and receive repeated competences tests at different ages in their lives. Thus, the measured competences of these respondents are characterized by large changes across childhood and adolescence. As such, the NEPS is confronted with various methodological issues such as linking test forms administered at different ages that vary significantly in their average difficulty. Nonetheless, these tests were intended to measure the same underlying construct. To gain deeper insight in the linking process under these conditions the setup of the present simulation study was oriented on reading tests, that were administered in grades 5, 7, 9, and 12 of the NEPS (Pohl et al., 2012; Krannich et al., 2017; Scharl et al., 2017). The observed mean proficiencies (in logits) were 0.0, 0.7, 1.2, and 1.5, respectively. Similar, we randomly drew proficiencies from normal distributions with these means and unit variances. We simulated responses to four test forms each including 25 items. The true item difficulties were generated in R 3.5.2 (R Core Team, 2018) from multivariate normal distributions matching the proficiency distributions (see **Table 1**), thus, resulting in a good test targeting. As the anchor items had to fit two distributions simultaneously ($t_{1/2}$, $t_{2/3}$, $t_{3/4}$), they were set to fall between two distributions (see **Tables 1, 2**). Anchor items maintained their difficulty parameters over time and as such met the assumption of measurement invariance. The item response models were estimated using the R-package TAM 3.1-26 (Kiefer et al., 2018) that iteratively updated the prior ability distribution using the EM algorithm (Bock and Aitkin, 1981, p. 443–459) during MML estimation (Kang and Petersen, 2012, p. 311–321). Due to the need of extensive computational power for the concurrent calibration, the quasi Monte Carlo estimation algorithm (based on 1,000 nodes) was used, whereas the Gauss-Hermite quadrature was used for the other linking methods. The original code for data generation is provided at https://osf.io/7vta8/.

Experimental Factors

For each simulated sample the four test forms (t_1–t_4) were linked based on the four linking methods of fixed parameter

TABLE 1 | True item difficulty and item discrimination parameters of the four test forms (t_1–t_4).

	Difficulty				Discrimination			
Position	t_1	t_2	t_3	t_4	t_1	t_2	t_3	t_4
1	$t_{1/2\_1}$: −1.255		$t_{3/4\_1}$: −0.272		$t_{1/2\_1}$: 0.804		$t_{3/4\_1}$: 1.267	
2	$t_{1/2\_2}$: −0.755		$t_{3/4\_2}$: 0.154		$t_{1/2\_2}$: 1.068		$t_{3/4\_2}$: 1.026	
3	$t_{1/2\_3}$: −0.415		$t_{3/4\_3}$: 0.576		$t_{1/2\_3}$: 1.266		$t_{3/4\_3}$: 1.237	
4	$t_{1/2\_4}$: 0.170		$t_{3/4\_4}$: 1.015		$t_{1/2\_4}$: 0.935		$t_{3/4\_4}$: 0.949	
5	$t_{1/2\_5}$: 0.534		$t_{3/4\_5}$: 1.493		$t_{1/2\_5}$: 0.737		$t_{3/4\_5}$: 0.789	
6	$t_{1/2\_6}$: 0.766		$t_{3/4\_6}$: 1.615		$t_{1/2\_6}$: 0.862		$t_{3/4\_6}$: 0.923	
7	$t_{1/2\_7}$: 0.966		$t_{3/4\_7}$: 1.889		$t_{1/2\_7}$: 1.270		$t_{3/4\_7}$: 1.023	
8	$t_{1/2\_8}$: 1.328		$t_{3/4\_8}$: 2.533		$t_{1/2\_8}$: 1.240		$t_{3/4\_8}$: 1.022	
9	$t_{1/2\_9}$: 1.900		$t_{3/4\_9}$: 3.218		$t_{1/2\_9}$: 0.935		$t_{3/4\_9}$: 1.038	
10	−2.537	$t_{2/3\_1}$: −1.048		0.149	0.767	$t_{2/3\_1}$: 1.040		0.808
11	−1.328	$t_{2/3\_2}$: 0.148		0.229	1.029	$t_{2/3\_2}$: 0.930		0.926
12	−0.998	$t_{2/3\_3}$: 0.578		0.270	0.940	$t_{2/3\_3}$: 1.010		1.134
13	−0.832	$t_{2/3\_4}$: 0.723		0.277	0.832	$t_{2/3\_4}$: 1.130		1.164
14	−0.664	$t_{2/3\_5}$: 0.925		0.342	0.973	$t_{2/3\_5}$: 0.930		0.884
15	−0.459	$t_{2/3\_6}$: 1.061		0.567	0.782	$t_{2/3\_6}$: 1.040		1.048
16	−0.360	$t_{2/3\_7}$: 1.570		0.957	0.808	$t_{2/3\_7}$: 0.960		0.860
17	−0.210	$t_{2/3\_8}$: 1.855		1.476	1.132	$t_{2/3\_8}$: 0.920		0.849
18	0.032	$t_{2/3\_9}$: 2.737		1.549	0.887	$t_{2/3\_9}$: 1.090		1.226
19	0.182	−0.485	−0.068	2.017	1.202	0.850	0.969	0.969
20	0.214	−0.258	0.166	2.266	1.147	1.100	0.971	1.110
21	0.300	0.187	0.312	2.529	0.987	1.040	1.136	0.823
22	0.602	0.864	1.620	2.995	1.165	0.880	0.821	0.966
23	0.769	1.365	1.921	3.094	0.860	1.100	0.941	1.012
24	0.879	1.738	2.434	3.170	1.321	0.850	1.096	0.993
25	1.498	2.489	2.961	3.393	0.928	1.170	0.935	1.188
M	0.013	0.707	1.205	1.500	0.995	1.006	1.008	1.009
SD	1.008	1.051	1.096	1.159	0.178	0.144	0.111	0.138

Framed parameters represent anchor items linking adjacent measurement points. Position = item position in each test form; $t_{1/2}$, $t_{2/3}$, $t_{3/4}$ = true anchor item parameters linking measurement points t_1, t_2, t_3, t_4; M = mean of 25 true item parameters; SD = standard deviation of 25 true item parameters.

TABLE 2 | Descriptive statistics of the true anchor item parameters split by the experimental factor number of anchor items.

	$t_{1/2}$			$t_{2/3}$			$t_{3/4}$		
	Anchor item difficulty parameters								
Anchor	**Position**	**M**	**SD**	**Position**	**M**	**SD**	**Position**	**M**	**SD**
3	2,5,8	0.369	1.051	2,5,8	0.976	0.855	2,5,8	1.393	1.193
5	2,3,4,6,9	0.333	1.050	1,5,6,7,8	0.873	1.138	2,3,4,6,9	1.316	1.193
7	1,2,4,5,6,7,9	0.332	1.066	1,3,4,5,6,8,9	0.976	1.169	1,3,4,5,6,7,9	1.362	1.096
9	1–9	0.360	1.022	1–9	0.950	1.074	1–9	1.358	1.120
	Anchor item discrimination parameters								
	Position	**M**	**SD**	**Position**	**M**	**SD**	**Position**	**M**	**SD**
3	2,5,8	1.015	0.256	2,5,8	0.927	0.006	2,5,8	0.946	0.136
5	2,3,4,6,9	1.013	0.160	1,5,6,7,8	0.978	0.058	2,3,4,6,9	1.035	0.123
7	1,2,4,5,6,7,9	0.944	0.178	1,3,4,5,6,8,9	1.023	0.077	1,3,4,5,6,7,9	1.032	0.171
9	1–9	1.013	0.206	1–9	1.006	0.076	1–9	1.030	0.148

*Anchor = Number of anchor items used for linking; t1/2, t2/3, t3/4 = true anchor item parameters linking adjacent measurement points; Position = selected anchor items out of anchor set (see **Table 1** for anchor item identification); M = mean of true anchor item parameters; SD = standard deviation of true anchor item parameters.*

calibration, mean/mean linking, weighted mean/mean linking, and concurrent calibration. Model fit was varied in two ways by either meeting the Rasch model assumptions of constant item discriminations ($\alpha_i = 1$) or modeling slight deviations (see **Table 1**) by drawing them from $N(1, 0.14^2)$. The resulting item discrimination parameters mirrored empirical results from a 2PL scaling of the tests (Krannich et al., 2017) mentioned above and, thus, were assumed to reflect a moderate degree of misfit within the range of operational proficiency test forms. Linking was based on a number of 3 (12%), 5 (20%), 7 (28%), or 9 (36%) common items among adjacent test forms (see **Table 1**). While 5 anchor items fell in line with recommendations in the literature (Kolen and Brennan, 2014), the other conditions evaluated the consequence of using more anchor items (7 or 9) or relying on a very restricted set of anchor items. The sample size condition was varied twofold ($N = 500$, $N = 3{,}000$). Overall, in addition to the within-subject experimental factor (four IRT-linking methods), three between-variable experimental factors—model fit (2), number of anchor items (4) and sample size (2)—were manipulated resulting in $4 \times 2 \times 4 \times 2 = 64$ conditions. Each within-subject experimental condition was simulated 100 times, to control for random sampling error.

Outcome Variables

We examined (a) the convergence rate of models as well as calculated (b) bias, (c) relative bias, and (d) root mean square error (RMSE) for sample mean and variance of the latent variable. The bias was calculated as $\widehat{\tau}_d - \tau$, with $\widehat{\tau}_d$ denoted as parameter estimate of the kth replication of condition d and τ denoting the true parameter value. The bias was then averaged over all k replications of each condition. Serving as an effect size, the relative bias was calculated as a proportion of $(\overline{\tau}_d - \tau)/\tau$, with $\overline{\tau}_d$ being the averaged parameter estimate over all k replications. Following Forero et al. (2009, p. 625–641), we considered a relative bias below 10% as acceptable. The RMSE gives the precision of a parameter estimate and was calculated as $\sqrt{\frac{1}{c}\sum_{k=1}^{c}(\widehat{\tau}_k - \tau)^2}$. As such the RMSE was defined as the square root of the mean of the squared bias.

RESULTS

Only negligible differences among the three linking methods of fixed parameter calibration, mean/mean and weighted mean/mean linking were found with regard to the outcome variables bias, relative bias and RMSE. Results are, therefore, reported combined. Descriptive statistics split by linking methods and experimental factors of the respective outcome variables are reported in **Supplementary Tables 2–5**.

Convergence Rates

Only 50.8% (i.e., 813 of 1,600 samples) of the models calibrated concurrently converged. Non-convergence was split about evenly among the experimental factors of sample size and model-data misfit, but varied substantially among different numbers of anchor items (see **Supplementary Table 1**). Moreover, in-depth analyses (not reported in this manuscript) of successfully converged concurrently calibrated models revealed that smaller numbers of iteration steps did not necessarily lead to a more precise parameter estimation. As these findings were questioning the applicability of concurrent calibration in settings based on small absolute numbers of anchor items, it was excluded from further analyses. In contrast, all models that were calibrated separately (fixed parameter calibration, mean/mean linking and weighted mean/mean linking) converged.

Sample Mean

Bias

Overall, there was no (change in) bias over the three time points ($M_{t2-t4} = 0.00$; t_1 was constrained to 0 due for model

identification) in the absence of model misfit. Neither sample size nor the number of anchor items had a substantial effect on the consistency of the bias of sample mean in the absence of model misfit (see **Figure 1**); although the bias was marginally smaller when sample size was N = 3,000 compared to N = 500. However, the sample mean was less well recovered in case of moderate model misfit (see **Figure 1** and **Supplementary Table 2**). Rather consistently, the sample mean was underestimated over the three time points, t_2–t_4, in all conditions but the conditions based on linking using 9 (36%) anchor items. The amount and pattern of the bias of sample mean emerged in a rather heterogeneous picture among time points and the number of anchor items. Overall, we found that the bias of sample mean rather decreased with an increasing number of anchor items.

Relative Bias

The relative bias was always explicitly below 10% and only rose above 5% in 2 conditions (see **Supplementary Table 2**) and was, thus, considered acceptable.

RMSE

The RMSE of sample mean linearly increased from t_2 to t_4 (see **Figure 2**). Sample size influenced the amount of RMSE as expected: smaller sample size led to a bigger RMSE with marginally steeper slope over time (N = 500: t_2 = 0.06 (SD = 0.04), t_3 = 0.08 (SD = 0.06), t_4 = 0.10 (SD = 0.08) compared to a larger sample size (N = 3,000: t_2 = 0.03 (SD = 0.02), t_3 and t_4 = 0.04 ($SD_{t3,t4}$ = 0.03). Additionally, the RMSE of sample mean was in general smaller when linking based on a larger number of anchor items. More precisely, a larger number of anchor items seemed more beneficial for a smaller sample size (N = 500). It has to be noted that a moderate Rasch model-data misfit did not necessarily lead to a decreased estimation precision of the sample mean. Rather the effect of model misfit on the RMSE of sample mean seemed to depend on the number of anchor items and was intercepted when the linking was based on at least 5 (20%) anchor items.

Sample Variance

Bias

Overall, there was no change in bias or its SD over the four time points (M_{t1-t4} = 0.00, SD_{t1-t4} = 0.06) in the absence of model misfit. Neither sample size nor the number of anchor items had a substantial effect on the consistency of the bias of sample variance in the absence of model misfit (see **Figure 3**). In case of moderate Rasch model-data misfit, the sample variance was marginally underestimated at t_1 and almost rose back to its true value with measurement progressing. This finding was similarly observed for different number of anchor items and sample size.

Relative Bias

The relative bias was considered acceptable in all conditions as it was always below 5% (see **Supplementary Table 4**).

RMSE

The RMSE of sample variance did not change from t_1 to t_4 (see **Figure 4**). Sample size influenced the amount of RMSE as expected: smaller sample size led to a larger RMSE [N = 500: t_1–t_4 = 0.07 (SD_{t1-t4} = 0.05)] compared to a larger sample size [N = 3,000: t_1–t_4 = 0.03 (SD_{t1-t4} = 0.02)]. No effect was found on the precision of the sample variance estimate due to the number of anchor items or a moderate Rasch model-data misfit.

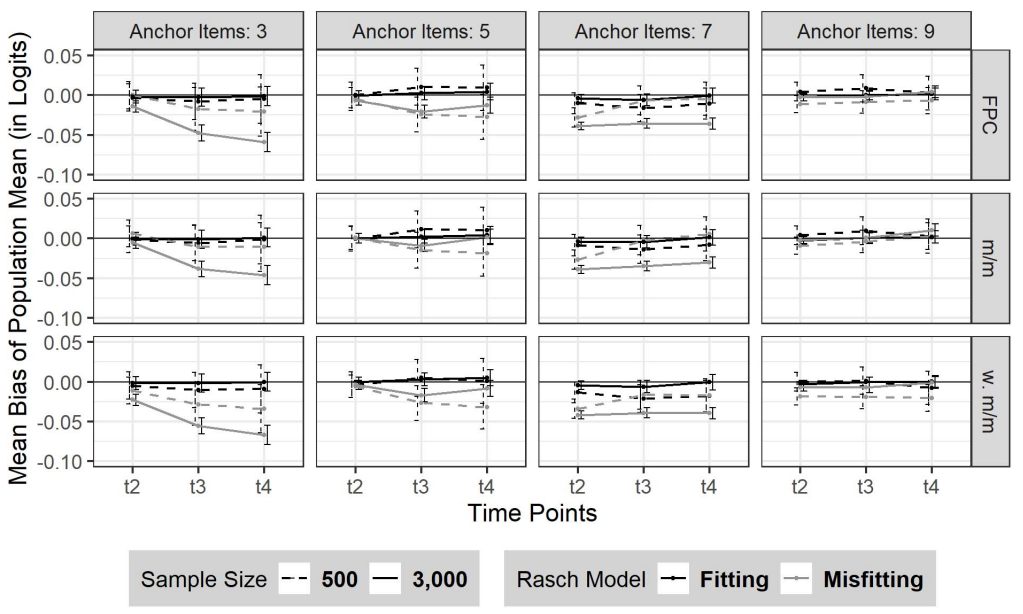


FIGURE 1 | Bias of sample mean over three time points (t_2–t_4). The figure is split by three linking methods and the experimental factors number of anchor items, sample size and Rasch model-data fit. FPC = fixed parameter calibration, m/m = mean/mean linking, w. m/m = weighted mean/mean linking. 95% confidence intervals are depicted.

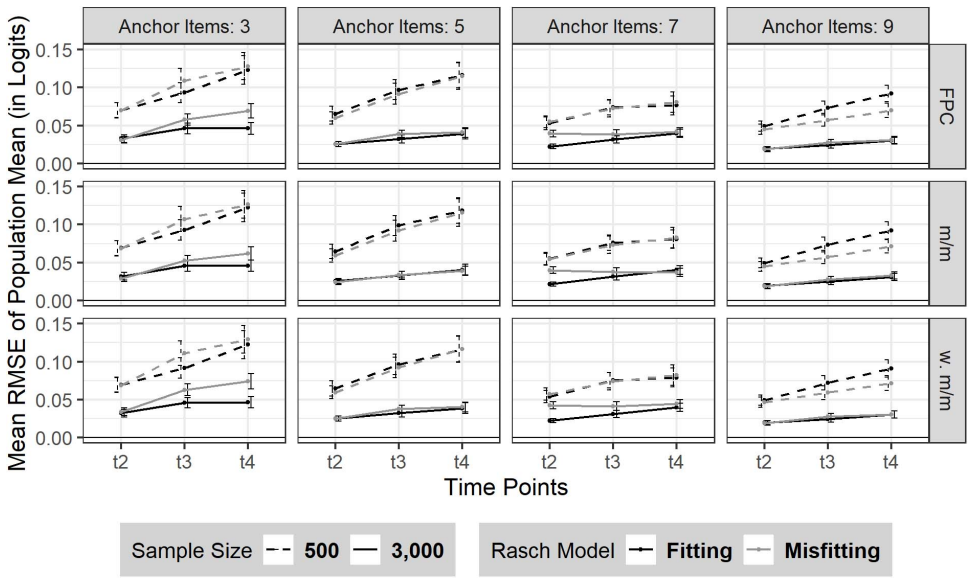


FIGURE 2 | RMSE of sample mean over three time points (t_2–t_4). The figure is split by the three linking methods and the experimental factors number of anchor items, sample size and Rasch model-data fit. FPC = fixed parameter calibration, Mean/Mean = mean/mean linking, w. Mean/Mean = weighted Mean/Mean. 95% confidence intervals are depicted.

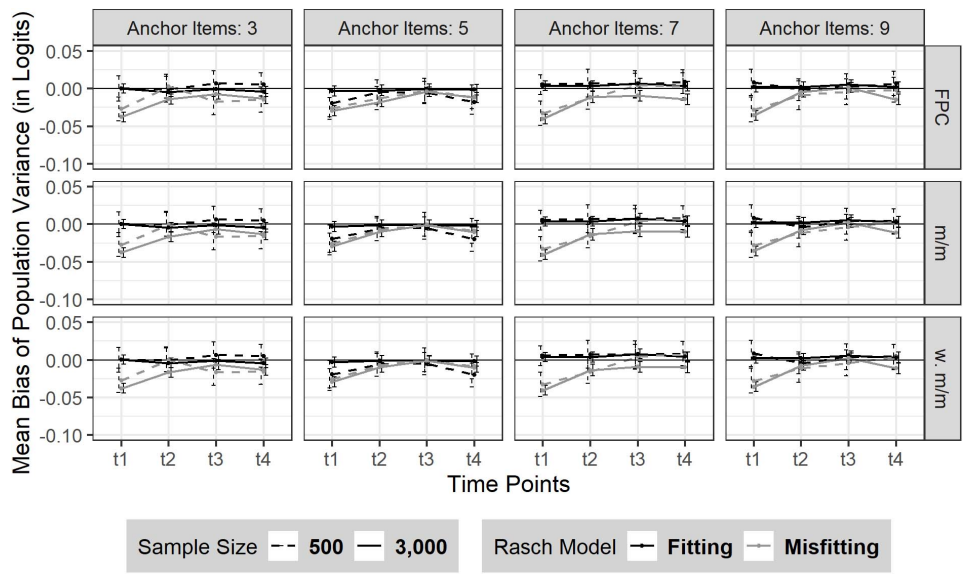


FIGURE 3 | Bias of sample variance over four time points (t_1–t_4). The figure is split by three linking methods and the experimental factors number of anchor items, sample size and Rasch model-data fit. FPC = fixed parameter calibration, m/m = mean/mean linking, w. m/m = weighted mean/mean linking. 95% confidence intervals are depicted.

DISCUSSION

The present simulation study focused on the comparison of four common IRT-linking methods (fixed parameter calibration, mean/mean linking, weighted mean/mean linking and concurrent calibration) within three experimental conditions (number of anchor items, sample size and model-data fit). Due to convergence issues, the application of concurrent calibration is not advisable for Rasch-scaled data when linking is based on a small absolute number of anchor items. The separate calibration linking methods somewhat unexpectedly resulted in negligible differences in the outcome variables of bias, relative bias and RMSE of sample mean and variance of the latent variable. Hence, the choice of linking method had no effect on the link outcome. This finding may result from the well fitted test targeting at each measurement point in the present

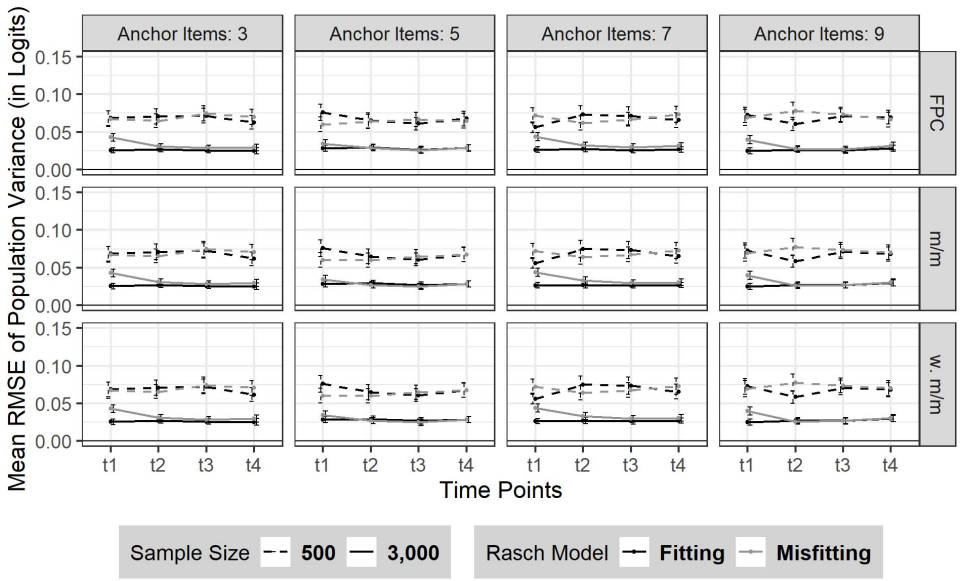


FIGURE 4 | RMSE of sample variance over four time points (t_1–t_4). The figure is split by three linking methods and the experimental factors number of anchor items, sample size and Rasch model-data fit. FPC = fixed parameter calibration, Mean/Mean = mean/mean linking, w. Mean/Mean = weighted Mean/Mean. 95% confidence intervals are depicted.

study. Thus, even though mean change between time points was substantial (up to 0.7 logits), there were only small differences in measurement precision within each set of anchor items, potentially depriving the method of weighted mean/mean linking of its unique strength in adjusting for differences in anchor item's *SE*s. Moreover, different amounts of mean change in proficiency over time were handled equally well by the three separate calibration methods. It is to be noted that no differences were found among the three linking methods in sensitivity and reactivity regarding moderate Rasch model-data misfit in the context of longitudinal linking.

In the absence of model misfit, the mean recovery of sample mean and variance was very good, regardless of the sample size or the number of anchor items used. However, in case of moderate Rasch model-data misfit, the parameters of sample mean and variance were generally slightly underestimated, suggesting an influence of the empirical relationship of anchor item difficulty parameters δ_i and anchor item discrimination parameters α_i. In contrast to prior findings reported in the literature (Zhao and Hambleton, 2017, p. 484), no substantial differences in performance were found between linking methods that based the linking on the anchor item level (e.g., FPC) or the anchor set level (e.g., m/m, wm/m). More specific, a certain composition of δ_i and α_i in the anchor items seemed to substantially influence the estimation of sample parameters. Factors characterizing this certain composition may include a deviation of item discrimination from 1 on the anchor item and/or anchor set level (i.e., whether misfit is balanced or not), the correlation's amount and/or direction of δ_i and α_i as well as person-item fit. Additionally, further investigating the consequences of Rasch model-data misfit seems a promising approach in detangling the compositional effects of anchor items. As the degree of model misfit was assumed to reflect a moderate degree of misfit within the range of operational proficiency test forms, we would furthermore deduce that an increasing degree of model misfit leads to an increasing deviation of parameter estimates from their true parameter.

In the present simulation study, change in proficiency was modeled as decelerating growth in steps of 0.7, 0.5, and 0.3 logits. Nevertheless, the amount of change between two time points seemed independent from the number of anchor items advisable to sufficiently map the change in proficiency distributions of the latent variable. This may suggest a transferability of the present findings to situations in that differences among groups are less pronounced.

It is to be noted, that the consistency of sample mean and variance estimation differed in their sensitivity to the number of anchor items in the case of moderate Rasch model-data misfit. However, accumulating effects (as reported by Keller and Keller, 2011, p. 362–379) of bias were only found when linking was based on 3 (12%) anchor items. While a number of 9 (36%) anchor items seemed sufficient to somewhat balance moderate misfit and resulted in good sample mean recovery, the recovery of sample variance seemed independent of the number of anchor items used. Similarly, for estimation precision of the sample mean, a bigger number of anchor items somewhat attenuated moderate Rasch model-data misfit, although this effect was more beneficial to a smaller sample size. Estimation precision of sample variance seemed to only depend on the sample size.

Practical Implications

As no substantial impact on parameter recovery of sample mean and variance was found due to moderate Rasch model-data misfit, the Rasch model seemed rather robust in the

present context. However, special attention should be payed to anchor items, as their characteristics critically determine sample parameter estimates. Therefore, using a 2PL model seems a practicable diagnostical tool to uncover noticeable deviations in anchor item discrimination parameters. Only marginal differences were found between the three IRT-linking methods of fixed parameter calibration, mean/mean linking and weighted mean/mean linking. More specifically, all of them were equally robust toward a moderate Rasch model-data misfit and different numbers of anchor items even when mean growth was substantial (0.7 logits). As such, the decision for a linking method could rely on more functional factors (e.g., scale preservation, practicability) in case of a well fitted test targeting. If, however, test targeting is expected to be poor, we agree with van der Linden and Barrett (2016, p. 650–673) that weighted mean/mean linking seems to be the preferable choice, as it allows for the inclusion of measurement precision as well as leaving the "pre linking" model fit unaltered. Furthermore, we would like to stress the point that defining an appropriate share of anchor items should depend on the respective Rasch model-data fit rather than following Kolen and Brennan's (2014) rule of thumb suggesting a share of 20%. In case of moderate misfit, we suggest a number of 7 (36%) anchor items, for the longitudinal linking of short (i.e., 25 items) operational test forms when a Rasch model is used for scaling. Additionally, in case of misfitting anchor items, findings hinted on a compensatory effect when the misfit present is balanced within an anchor item set.

Due to the issues of non-convergence and the disproportionate occurrence of extreme values in parameter recovery, concurrent calibration seemed less suitable for common use than separate calibration methods in longitudinal study designs using small absolute numbers of anchor items.

Limitations of the Study

The setup of the simulation study did not consider several issues relevant in empirical contexts such as missing data or differential item functioning in anchor items. Similarly, our simulated anchor items exhibited good test targeting for the two proficiency distributions intended to link, which might be hard to achieve in operational assessments. These simplifications of reality were taken into account in order to master the complexity of the central issue. As a consequence, results may be limited in their transferability to empirical data. Future research should study these aspects in more detail and, thus, could further elaborate on the conditions that allow precise linking in the context of the Rasch model. Moreover, the present study was motivated by operational LSAs which are usually characterized by relatively large sample sizes and rather short test forms. In other empirical settings that include smaller sample sizes often substantially longer test forms can be administered. Therefore, future research could address the particulars of linking in these studies. Particularly, this research could also explore whether alternative scaling approaches (e.g., the 2-parameter logistic model) might show more pronounced benefits for data exhibiting misfit to the Rasch model or whether the linking results are comparable to the findings presented in the present study.

As the mean of α_i within anchor item sets as well as the correlations of δ_i and α_i in the present simulation study were not varied systematically, the underlying mechanisms affecting the recovery of sample mean and variance in case of moderate Rasch model-data misfit was not fully traceable and, thus, limited the conclusions on certain compositional effects inherent to sets of anchor items. However, regarding longitudinal measurements, considering the empirical correlation of δ_i and α_i only, would fall short for the effect of person-item fit. As anchor item difficulties are held constant in repeated administrations to samples with variable proficiencies, person-item fit differs between time points. Therefore, differential effects of an anchor item on the estimation of sample parameters (Bolt et al., 2014, p. 141–162) are to be additionally considered between time points in case of Rasch model-data misfit (Humphry, 2018, p. 216–228).

CONCLUSION

Overall, the challenges inherent to contexts characterized by small absolute and relative numbers of anchor items due to short test length as well as small to medium sample sizes were mastered equally well by the three separate calibration methods mean/mean linking, weighted mean/mean linking and fixed parameter calibration, resulting in reliable and valid parameter recovery. However, results of the present simulation study suggested that the choice of linking method is rather secondary when linking Rasch modeled data—independent of the absence or presence of (moderate) model misfit. More important seems the awareness of the practitioner that a combination of moderate model misfit and certain factors (e.g., empirical relation of δ_i and α_i, composition of anchor items, person-item fit, sample size) may lead to a distorted parameter estimation—although at presence no applicable diagnostics nor concrete guidelines for empirical data seem at hand. As such, future research should analytically deduce and systematically investigate the consequences of an interaction between Rasch model-data misfit and certain experimental factors.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

LF conducted the literature research, drafted significant parts of the manuscript, and analyzed and interpreted the data used in this study. TG wrote the code for the simulation study. CC, TR, and TG substantively revised the manuscript and provided substantial input for the statistical analyses. All authors read and approved the final manuscript.

FUNDING

We would like to thank the Deutsche Forschungsgemeinschaft (DFG; www.dfg.de) for funding our research project within the Priority Programme 1646 entitled "Analyzing relations between latent competencies and context information in the National Educational Panel Study" under Grant No. CA 289/8-2 (awarded to CC). We furthermore thank the Leibniz Institute for Educational Trajectories (www.lifbi.de) for funding the open access publication fee.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.633896/full#supplementary-material

REFERENCES

Birnbaum, A. (1968). "Some latent trait models and their use in inferring an examinee's ability," in *Statistical Theories of Mental test Scores*, eds F. M. Lord and M. R. Novick (Reading, MA: Addison-Wesley Publishing), 397–472.

Blossfeld, H. P., Roßbach, H. G., and von Maurice, , J. (Eds.) (2011). "Zeitschrift für erziehungswissenschaft sonderheft," in *Education as a Lifelong Process: The German National Educational Panel Study (NEPS)*, Vol. 14, (Wiesbaden: VS Verlag für Sozialwissenschaften).

Bock, R. D., and Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. *Psychometrika* 46, 443–459. doi: 10.1007/bf02293801

Bolt, D. M., Deng, S., and Lee, S. (2014). IRT model misspecification and measurement of growth in vertical scaling. *J. Educ. Meas.* 51:2. doi: 10.1111/jedm.12039

Fischer, L., Gnambs, T., Rohm, T., and Carstensen, C. H. (2019). Longitudinal linking of Rasch-model-scaled competence tests in large-scale assessments: a comparison and evaluation of different linking methods and anchoring designs based on two tests on mathematical competence administered in grades 5 and 7. *Psychol. Test Assessment Model.* 61, 37–64.

Forero, C. G., Maydeu-Olivares, A., and Gallardo-Pujol, D. (2009). Factor analysis with ordinal indicators: a monte carlo study comparing DWLS and ULS estimation. *Struct. Equ. Model.* 16, 625–641. doi: 10.1080/10705510903203573

Humphry, S. M. (2018). The impact of levels of discrimination on vertical equating in the rasch model. *J. Appl. Meas.* 19, 216–228.

Kang, T., and Petersen, N. S. (2012). Linking item parameters to a base scale. *Asia Pacific Educ. Rev.* 13:2. doi: 10.1007/s12564-011-9197-2

Keller, L. A., and Keller, R. R. (2011). The long-term sustainability of different item response theory scaling methods. *Educ. Psychol. Meas.* 71, 362–379. doi: 10.1177/0013164410375111

Kiefer, T., Robitzsch, A., and Wu, M. (2018). *TAM: Test Analysis Modules. [Computer Software]*. Available online at: https://CRAN.R-project.org/package=TAM

Kim, S. (2006). A comparative study of IRT fixed parameter calibration methods. *J. Educ. Meas.* 43:4. doi: 10.1111/j.1745-3984.2006.00021.x

Kim, S., and Cohen, A. S. (1998). A comparison of linking and concurrent calibration under item response theory. *Appl. Psychol. Meas.* 22:2.

Kolen, M. J., and Brennan, R. L. (2014). *Test Equating, Scaling, and Linking: Methods and Practices. Statistics for Social and Behavioral Sciences*, 3rd Edn. New York, NY: Springer.

Krannich, M., Jost, O., Rohm, T., Koller, I., Carstensen, C. H., Fischer, L., et al. (2017). *NEPS Technical Report for Reading: Scaling results of Starting Cohort 3 for grade 7*. NEPS Survey Papers,14. Bamberg: Leibniz Institute for Educational Trajectories.

Loyd, B. H., and Hoover, H. D. (1980). Vertical equating using the Rasch model. *J. Educ. Meas.* 17, 179–193. doi: 10.1111/j.1745-3984.1980.tb00825.x

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *J. Educ. Meas.* 14:2. doi: 10.1111/j.1745-3984.1977.tb00033.x

Meijer, R. R., and Tendeiro, J. N. (2015). *The Effect of Item and Person Misfit on Selection Decisions: An Empirical Study*. LSAC Research Report Series 15:05. Newton, PA: Law School Admission Council.

Pohl, S., Haberkorn, K., Hardt, K., and Wiegand, E. (2012). *NEPS Technical Report for Reading – NEPS Technical Report for reading: Scaling results of Starting Cohort 3 in fifth grade*. NEPS Working Paper,15. Bamberg: Leibniz Institute for Educational Trajectories.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rasch, G. (1960). *Probabilistic Models For Some Intelligence And Attainment Tests: Studies In Mathematical Psychology: I*. Copenhagen: Danmarks Paedagogiske Institut.

Scharl, A., Fischer, L., Gnambs, T., and Rohm, T. (2017). *NEPS Technical Report for Reading: Scaling Results of Starting Cohort 3 for Grade 9*. NEPS Survey Papers, 20. Bamberg: Leibniz Institute for Educational Trajectories.

Sinharay, S., and Haberman, S. J. (2014). How often is the misfit of item response theory models practically significant? *Educ. Meas.* 33:1. doi: 10.1111/emip.12024

Stocking, M. L., and Lord, F. M. (1983). Developing a common metric in item response theory. *Appl. Psychol. Meas.* 7, 201–210. doi: 10.1177/014662168300700208

Svetina, D., Crawford, A. V., Levy, R., Green, S. B., Scott, L., Thompson, M., et al. (2013). Designing small-scale tests: a simulation study of parameter recovery with the 1-PL. *Psychol. Test Assessment Modeling* 55, 335–360.

Thissen, D., and Wainer, H. (1982). Some standard errors in item response theory. *Psychometrika* 47, 397–412. doi: 10.1007/BF02293705

Vale, C. D. (1986). Linking item parameters onto a common scale. *Appl. Psychol. Meas.* 10:4. doi: 10.1177/014662168601000402

van der Linden, W. J., and Barrett, M. D. (2016). Linking item response model parameters. *Psychometrika* 81:3. doi: 10.1007/s11336-015-9469-6

van der Linden, W. J., and Hambleton, R. K. (2013). *Handbook of Modern Item Response Theory*. Berlin: Springer Science & Business Media.

von Davier, A. A., Carstensen, C. H., and von Davier, M. (2006). Linking competencies in educational settings and measuring growth. *ETS Res. Rep. Ser.* 2006:1. doi: 10.1002/j.2333-8504.2006.tb02018.x

Wright, B. D. (1977). Solving measurement problems with the rasch model. *J. Educ. Meas.* 14, 97–116. doi: 10.1111/j.1745-3984.1977.tb00031.x

Zhao, Y., and Hambleton, R. K. (2017). Practical consequences of item response theory model misfit in the context of test equating with mixed-format test data. *Front. Psychol.* 8:484. doi: 10.3389/fpsyg.2017.00484

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Supplementary Material

Supplementary Table 1. Convergence Rate of Models Based on Concurrent Calibration Split by Experimental Conditions. Each condition was simulated 100 times. No. of Anchor Items = Number of anchor items; *Md* = Median of the convergence rate of the models based on concurrent calibration split by experimental conditions; Models Converged (%) = model convergence rate in percent.

Condition	No. of Anchor Items	Rasch Model	Sample Size	Models Converged (%)
1	3	Fit	500	21
2			3,000	36
3		Misfit	500	27
4			3,000	33
5	5	Fit	500	38
6			3,000	34
7		Misfit	500	43
8			3,000	44
9	7	Fit	500	59
10			3,000	51
11		Misfit	500	64
12			3,000	62
13	9	Fit	500	74
14			3,000	71
15		Misfit	500	73
16			3,000	83
Md	30/40.5/60.5/73.5	44.5/53	51/47.5	47.5

Supplementary Table 2. Descriptive Statistics for the Bias of Sample Mean Split by Experimental Conditions and Linking Methods. Anchor: = Number of anchor items used for linking; *M*(*SD*) = mean and standard deviation of the bias of sample mean; *Min/Max* = *minimum*/*maximum* of the bias of sample mean; *RB* = mean relative bias of sample mean; FPC = fixed parameter calibration; Mean/Mean = mean/mean linking; weighted Mean/Mean = weighted m/m. The bias was averaged over 100 replications.

Linking Method	*N*	Rasch model	Time Point	Anchor: 3			Anchor: 5			Anchor: 7			Anchor: 9		
				***M*(*SD*)**	***Min/Max***	***RB***	***M*(*SD*)**	***Min/Max***	***RB***	***M*(*SD*)**	***Min/Max***	***RB***	***M*(*SD*)**	***Min/Max***	***RB***
FPC	500	Fit	t_2	0.00(0.09)	-0.25/0.20	0.00	0.00(0.08)	-0.22/0.26	0.00	-0.01(0.07)	-0.22/0.16	-0.01	0.00(0.06)	-0.14/0.16	0.01
			t_3	-0.01(0.12)	-0.34/0.23	-0.01	0.01(0.12)	-0.28/0.29	0.01	-0.02(0.09)	-0.22/0.19	-0.01	0.01(0.09)	-0.17/0.19	0.01
			t_4	0.00(0.16)	-0.53/0.46	0.00	0.01(0.15)	-0.32/0.32	0.01	-0.01(0.10)	-0.24/0.27	-0.01	0.00(0.11)	-0.26/0.19	0.00
		Misfit	t_2	0.00(0.09)	-0.18/0.20	0.00	0.00(0.07)	-0.16/0.22	-0.01	-0.03(0.06)	-0.18/0.13	-0.04	-0.01(0.06)	-0.13/0.12	-0.02
			t_3	-0.02(0.14)	-0.35/0.36	-0.01	-0.02(0.11)	-0.38/0.25	-0.02	-0.01(0.09)	-0.25/0.31	-0.01	-0.01(0.07)	-0.18/0.18	-0.01
			t_4	-0.02(0.16)	-0.36/0.34	-0.01	-0.03(0.14)	-0.37/0.35	-0.02	0.00(0.11)	-0.32/0.24	0.00	-0.01(0.08)	-0.20/0.19	0.00
	3,000	Fit	t_2	0.00(0.04)	-0.12/0.11	0.00	0.00(0.03)	-0.06/0.06	0.00	0.00(0.03)	-0.06/0.05	-0.01	0.00(0.02)	-0.06/0.06	0.00
			t_3	0.00(0.06)	-0.12/0.16	0.00	0.00(0.04)	-0.09/0.15	0.00	-0.01(0.04)	-0.13/0.09	0.00	0.00(0.03)	-0.07/0.10	0.00
			t_4	0.00(0.06)	-0.14/0.17	0.00	0.00(0.05)	-0.13/0.17	0.00	0.00(0.05)	-0.12/0.09	0.00	0.00(0.04)	-0.09/0.11	0.00
		Misfit	t_2	-0.01(0.04)	-0.10/0.06	-0.02	-0.01(0.03)	-0.08/0.06	-0.01	-0.04(0.02)	-0.09/0.03	-0.06	0.00(0.02)	-0.07/0.06	0.00
			t_3	-0.05(0.05)	-0.16/0.07	-0.04	-0.02(0.04)	-0.11/0.08	-0.02	-0.04(0.03)	-0.10/0.05	-0.03	0.00(0.03)	-0.09/0.09	0.00
			t_4	-0.06(0.06)	-0.19/0.09	-0.04	-0.01(0.05)	-0.15/0.12	-0.01	-0.04(0.04)	-0.12/0.09	-0.02	0.00(0.04)	-0.09/0.11	0.00
Mean/Mean	500	Fit	t_2	0.00(0.09)	-0.22/0.19	0.00	0.00(0.08)	-0.23/0.26	0.00	-0.01(0.07)	-0.22/0.15	-0.01	0.00(0.06)	-0.14/0.14	0.01
			t_3	-0.01(0.12)	-0.34/0.22	0.00	0.01(0.12)	-0.25/0.29	0.01	-0.01(0.09)	-0.23/0.17	-0.01	0.01(0.09)	-0.20/0.20	0.01
			t_4	0.00(0.16)	-0.53/0.44	0.00	0.01(0.15)	-0.34/0.32	0.01	-0.01(0.10)	-0.25/0.30	-0.01	0.00(0.11)	-0.28/0.21	0.00
		Misfit	t_2	0.01(0.09)	-0.18/0.20	0.01	0.00(0.07)	-0.15/0.22	0.00	-0.03(0.06)	-0.19/0.14	-0.04	-0.01(0.06)	-0.14/0.12	-0.01
			t_3	-0.01(0.14)	-0.32/0.38	-0.01	-0.01(0.12)	-0.37/0.24	-0.01	0.00(0.09)	-0.23/0.32	0.00	0.00(0.07)	-0.19/0.20	0.00
			t_4	-0.01(0.16)	-0.33/0.34	-0.01	-0.02(0.15)	-0.36/0.35	-0.01	0.01(0.11)	-0.31/0.27	0.00	0.00(0.09)	-0.20/0.19	0.00
	3,000	Fit	t_2	0.00(0.04)	-0.12/0.11	0.00	0.00(0.03)	-0.06/0.06	0.00	0.00(0.03)	-0.06/0.05	-0.01	0.00(0.02)	-0.06/0.05	0.00
			t_3	0.00(0.06)	-0.12/0.16	0.00	0.00(0.04)	-0.10/0.15	0.00	0.00(0.04)	-0.11/0.09	0.00	0.00(0.03)	-0.07/0.10	0.00
			t_4	0.00(0.06)	-0.15/0.17	0.00	0.00(0.05)	-0.13/0.16	0.00	0.00(0.05)	-0.12/0.10	0.00	0.00(0.04)	-0.08/0.12	0.00
		Misfit	t_2	-0.01(0.04)	-0.08/0.07	-0.01	0.00(0.03)	-0.07/0.07	0.00	-0.04(0.02)	-0.09/0.03	-0.06	0.00(0.02)	-0.07/0.06	0.00
			t_3	-0.04(0.05)	-0.15/0.08	-0.03	-0.01(0.04)	-0.09/0.09	-0.01	-0.03(0.03)	-0.10/0.04	-0.03	0.00(0.03)	-0.10/0.09	0.00
			t_4	-0.05(0.06)	-0.18/0.11	-0.03	0.00(0.05)	-0.13/0.13	0.00	-0.03(0.03)	-0.12/0.09	-0.02	0.01(0.04)	-0.07/0.12	0.01
weighted Mean/Mean	500	Fit	t_2	0.00(0.09)	-0.25/0.17	-0.01	0.00(0.08)	-0.22/0.26	-0.01	-0.01(0.07)	-0.23/0.15	-0.02	0.00(0.06)	-0.14/0.16	0.00
			t_3	-0.01(0.11)	-0.33/0.22	-0.01	0.00(0.12)	-0.28/0.28	0.00	-0.02(0.09)	-0.22/0.17	-0.02	0.00(0.09)	-0.18/0.18	0.00
			t_4	-0.01(0.15)	-0.53/0.45	-0.01	0.00(0.15)	-0.33/0.32	0.00	-0.02(0.10)	-0.25/0.26	-0.01	-0.01(0.11)	-0.27/0.18	-0.01

Linking Method	*N*	Rasch model	Time Point	Anchor: 3 *M*(*SD*)	Anchor: 3 *Min*/*Max*	Anchor: 3 *RB*	Anchor: 5 *M*(*SD*)	Anchor: 5 *Min*/*Max*	Anchor: 5 *RB*	Anchor: 7 *M*(*SD*)	Anchor: 7 *Min*/*Max*	Anchor: 7 *RB*	Anchor: 9 *M*(*SD*)	Anchor: 9 *Min*/*Max*	Anchor: 9 *RB*
			t_2	-0.01(0.08)	-0.20/0.19	-0.02	-0.01(0.07)	-0.16/0.22	-0.01	-0.03(0.06)	-0.18/0.12	-0.05	-0.02(0.05)	-0.14/0.11	-0.03
		Misfit	t_3	-0.03(0.14)	-0.36/0.31	-0.02	-0.03(0.11)	-0.37/0.24	-0.02	-0.02(0.09)	-0.25/0.30	-0.01	-0.02(0.07)	-0.19/0.18	-0.02
			t_4	-0.03(0.16)	-0.36/0.33	-0.02	-0.03(0.14)	-0.37/0.31	-0.02	-0.02(0.11)	-0.34/0.24	-0.01	-0.02(0.09)	-0.20/0.16	-0.01
			t_2	0.00(0.04)	-0.12/0.11	0.00	0.00(0.03)	-0.06/0.06	0.00	0.00(0.03)	-0.07/0.05	-0.01	0.00(0.02)	-0.06/0.06	0.00
		Fit	t_3	0.00(0.06)	-0.12/0.16	0.00	0.00(0.04)	-0.08/0.15	0.00	-0.01(0.04)	-0.13/0.09	-0.01	0.00(0.03)	-0.07/0.10	0.00
			t_4	0.00(0.06)	-0.14/0.16	0.00	0.00(0.05)	-0.14/0.16	0.00	0.00(0.05)	-0.11/0.10	0.00	0.00(0.04)	-0.09/0.11	0.00
	3,000		t_2	-0.02(0.04)	-0.10/0.05	-0.03	0.00(0.03)	-0.07/0.06	-0.01	-0.04(0.02)	-0.10/0.02	-0.06	-0.01(0.02)	-0.07/0.06	-0.01
		Misfit	t_3	-0.06(0.05)	-0.17/0.07	-0.05	-0.02(0.04)	-0.11/0.09	-0.01	-0.04(0.03)	-0.11/0.04	-0.03	-0.01(0.03)	-0.10/0.09	-0.01
			t_4	-0.07(0.06)	-0.19/0.08	-0.04	-0.01(0.05)	-0.16/0.13	-0.01	-0.04(0.04)	-0.13/0.08	-0.03	0.00(0.04)	-0.10/0.11	0.00

Supplementary Table 3. Descriptive Statistics for the RMSE of Sample Mean Split by Experimental Conditions and Linking Methods. Anchor: = Number of anchor items used for linking; *M*(*SD*) = mean and standard deviation of the RMSE of sample mean; *Min/Max* = *minimum*/*maximum* of the RMSE of sample mean; FPC = fixed parameter calibration; Mean/Mean = mean/mean linking; weighted Mean/Mean = weighted m/m. The RMSE was averaged over 100 replications.

Linking Method	*N*	Rasch model	Time Point	Anchor: 3 *M*(*SD*)	Anchor: 3 *Min*/*Max*	Anchor: 5 *M*(*SD*)	Anchor: 5 *Min*/*Max*	Anchor: 7 *M*(*SD*)	Anchor: 7 *Min*/*Max*	Anchor: 9 *M*(*SD*)	Anchor: 9 *Min*/*Max*
			t_2	0.07(0.05)	0.00/0.25	0.07(0.05)	0.00/0.26	0.05(0.04)	0.00/0.22	0.05(0.03)	0.00/0.16
		Fit	t_3	0.09(0.07)	0.00/0.34	0.10(0.07)	0.00/0.29	0.07(0.05)	0.00/0.22	0.07(0.05)	0.00/0.19
			t_4	0.12(0.10)	0.00/0.53	0.12(0.09)	0.00/0.32	0.08(0.06)	0.00/0.27	0.09(0.06)	0.00/0.26
	500		t_2	0.07(0.05)	0.00/0.20	0.06(0.04)	0.00/0.22	0.05(0.04)	0.00/0.18	0.05(0.03)	0.00/0.13
		Misfit	t_3	0.11(0.08)	0.00/0.36	0.09(0.07)	0.00/0.38	0.07(0.06)	0.00/0.31	0.06(0.04)	0.00/0.18
FPC			t_4	0.13(0.09)	0.00/0.36	0.11(0.09)	0.00/0.37	0.08(0.07)	0.00/0.32	0.07(0.05)	0.00/0.20
			t_2	0.03(0.03)	0.00/0.12	0.03(0.02)	0.00/0.06	0.02(0.02)	0.00/0.06	0.02(0.01)	0.00/0.06
		Fit	t_3	0.05(0.04)	0.00/0.16	0.03(0.03)	0.00/0.15	0.03(0.02)	0.00/0.13	0.02(0.02)	0.00/0.10
			t_4	0.05(0.04)	0.00/0.17	0.04(0.04)	0.00/0.17	0.04(0.03)	0.00/0.12	0.03(0.02)	0.00/0.11
	3,000		t_2	0.03(0.02)	0.00/0.10	0.03(0.02)	0.00/0.08	0.04(0.02)	0.00/0.09	0.02(0.01)	0.00/0.07
		Misfit	t_3	0.06(0.04)	0.00/0.16	0.04(0.03)	0.00/0.11	0.04(0.03)	0.00/0.10	0.03(0.02)	0.00/0.09
			t_4	0.07(0.05)	0.00/0.19	0.04(0.03)	0.00/0.15	0.04(0.03)	0.00/0.12	0.03(0.02)	0.00/0.11
Mean/Mean	500	Fit	t_2	0.07(0.05)	0.00/0.22	0.06(0.05)	0.00/0.26	0.06(0.04)	0.00/0.22	0.05(0.03)	0.00/0.14

Linking Method	*N*	Rasch model	Time Point	Anchor: 3 *M*(*SD*)	Anchor: 3 *Min/Max*	Anchor: 5 *M*(*SD*)	Anchor: 5 *Min/Max*	Anchor: 7 *M*(*SD*)	Anchor: 7 *Min/Max*	Anchor: 9 *M*(*SD*)	Anchor: 9 *Min/Max*
			t_3	0.09(0.07)	0.00/0.34	0.10(0.07)	0.00/0.29	0.08(0.05)	0.00/0.23	0.07(0.05)	0.00/0.20
			t_4	0.12(0.10)	0.00/0.53	0.12(0.09)	0.00/0.34	0.08(0.06)	0.00/0.30	0.09(0.06)	0.00/0.28
			t_2	0.07(0.05)	0.00/0.20	0.06(0.04)	0.00/0.22	0.05(0.04)	0.00/0.19	0.04(0.03)	0.00/0.14
		Misfit	t_3	0.11(0.08)	0.00/0.38	0.09(0.07)	0.00/0.37	0.07(0.06)	0.00/0.32	0.06(0.04)	0.00/0.20
			t_4	0.13(0.09)	0.00/0.34	0.12(0.09)	0.00/0.36	0.08(0.07)	0.00/0.31	0.07(0.05)	0.00/0.20
			t_2	0.03(0.03)	0.00/0.12	0.03(0.02)	0.00/0.06	0.02(0.01)	0.00/0.06	0.02(0.01)	0.00/0.06
		Fit	t_3	0.05(0.04)	0.00/0.16	0.03(0.03)	0.00/0.15	0.03(0.02)	0.00/0.11	0.02(0.02)	0.00/0.10
	3,000		t_4	0.05(0.04)	0.00/0.17	0.04(0.04)	0.00/0.16	0.04(0.03)	0.00/0.12	0.03(0.02)	0.00/0.12
			t_2	0.03(0.02)	0.00/0.08	0.02(0.02)	0.00/0.07	0.04(0.02)	0.00/0.09	0.02(0.01)	0.00/0.07
		Misfit	t_3	0.05(0.04)	0.00/0.15	0.03(0.02)	0.00/0.09	0.04(0.03)	0.00/0.10	0.03(0.02)	0.00/0.10
			t_4	0.06(0.05)	0.00/0.18	0.04(0.03)	0.00/0.13	0.04(0.03)	0.00/0.12	0.03(0.02)	0.00/0.12
			t_2	0.07(0.05)	0.00/0.25	0.06(0.05)	0.00/0.26	0.05(0.04)	0.00/0.23	0.05(0.03)	0.00/0.16
		Fit	t_3	0.09(0.07)	0.00/0.33	0.10(0.07)	0.00/0.28	0.08(0.05)	0.00/0.22	0.07(0.05)	0.00/0.18
			t_4	0.12(0.09)	0.00/0.53	0.12(0.09)	0.00/0.33	0.08(0.06)	0.00/0.26	0.09(0.06)	0.00/0.27
	500		t_2	0.07(0.05)	0.00/0.20	0.06(0.04)	0.00/0.22	0.06(0.04)	0.00/0.18	0.05(0.03)	0.00/0.14
		Misfit	t_3	0.11(0.08)	0.00/0.36	0.09(0.07)	0.00/0.37	0.07(0.06)	0.00/0.30	0.06(0.05)	0.00/0.19
weighted Mean/Mean			t_4	0.13(0.09)	0.00/0.36	0.12(0.09)	0.00/0.37	0.08(0.07)	0.00/0.34	0.07(0.05)	0.00/0.20
			t_2	0.03(0.03)	0.00/0.12	0.03(0.02)	0.00/0.06	0.02(0.01)	0.00/0.07	0.02(0.01)	0.00/0.06
		Fit	t_3	0.05(0.04)	0.00/0.16	0.03(0.03)	0.00/0.15	0.03(0.02)	0.00/0.13	0.02(0.02)	0.00/0.10
			t_4	0.05(0.04)	0.00/0.16	0.04(0.04)	0.00/0.16	0.04(0.03)	0.00/0.11	0.03(0.02)	0.00/0.11
	3,000		t_2	0.03(0.02)	0.00/0.10	0.02(0.02)	0.00/0.07	0.04(0.02)	0.00/0.10	0.02(0.02)	0.00/0.07
		Misfit	t_3	0.06(0.04)	0.00/0.17	0.04(0.02)	0.00/0.11	0.04(0.03)	0.00/0.11	0.03(0.02)	0.00/0.10
			t_4	0.07(0.05)	0.00/0.19	0.04(0.03)	0.00/0.16	0.04(0.03)	0.00/0.13	0.03(0.02)	0.00/0.11

Supplementary Table 4. Descriptive Statistics for the Bias of Sample Variance Split by Experimental Conditions and Linking Methods. Anchor: = Number of anchor items used for linking; *M*(*SD*) = mean and standard deviation of the bias of sample variance; *Min/Max* = *minimum*/*maximum* of the bias of sample variance; *RB* = mean relative bias of sample variance; FPC = fixed parameter calibration; Mean/Mean = mean/mean linking; weighted Mean/Mean = weighted m/m. The bias was averaged over 100 replications.

Linking Method	*N*	Rasch model	Time Point	Anchor: 3			Anchor: 5			Anchor: 7			Anchor: 9		
				M(*SD*)	*Min*/*Max*	*RB*	*M*(*SD*)	*Min*/*Max*	*RB*	*M*(*SD*)	*Min*/*Max*	*RB*	*M*(*SD*)	*Min*/*Max*	*RB*
FPC	500	Fit	t_1	0.00(0.08)	-0.18/0.24	0.00	-0.02(0.09)	-0.18/0.34	-0.02	0.01(0.07)	-0.14/0.16	0.01	0.01(0.09)	-0.19/0.20	0.01
			t_2	0.00(0.09)	-0.21/0.23	0.00	0.00(0.08)	-0.18/0.29	0.00	0.01(0.09)	-0.20/0.32	0.01	0.00(0.07)	-0.18/0.18	0.00
			t_3	0.01(0.09)	-0.21/0.20	0.01	-0.01(0.08)	-0.17/0.19	-0.01	0.01(0.09)	-0.23/0.26	0.01	0.00(0.09)	-0.18/0.22	0.00
			t_4	0.01(0.08)	-0.18/0.17	0.01	-0.02(0.08)	-0.25/0.14	-0.02	0.01(0.08)	-0.21/0.28	0.01	0.01(0.08)	-0.17/0.18	0.01
		Misfit	t_1	-0.03(0.08)	-0.19/0.20	-0.03	-0.03(0.07)	-0.21/0.23	-0.03	-0.03(0.08)	-0.17/0.21	-0.03	-0.03(0.08)	-0.27/0.23	-0.03
			t_2	0.00(0.08)	-0.17/0.25	0.00	-0.01(0.08)	-0.20/0.18	-0.01	-0.01(0.08)	-0.15/0.26	-0.01	-0.01(0.10)	-0.21/0.26	-0.01
			t_3	-0.02(0.09)	-0.20/0.24	-0.02	0.00(0.08)	-0.19/0.19	0.00	0.00(0.08)	-0.17/0.19	0.00	0.00(0.09)	-0.21/0.19	0.00
			t_4	-0.01(0.09)	-0.21/0.19	-0.01	-0.01(0.08)	-0.21/0.17	-0.01	0.00(0.09)	-0.24/0.26	0.00	0.00(0.08)	-0.19/0.23	0.00
	3,000	Fit	t_1	0.00(0.03)	-0.07/0.07	0.00	0.00(0.03)	-0.08/0.07	0.00	0.00(0.03)	-0.06/0.08	0.00	0.00(0.03)	-0.10/0.07	0.00
			t_2	0.00(0.03)	-0.07/0.07	0.00	0.00(0.04)	-0.10/0.07	0.00	0.00(0.03)	-0.08/0.09	0.00	0.00(0.03)	-0.07/0.06	0.00
			t_3	0.00(0.03)	-0.08/0.07	0.00	0.00(0.03)	-0.09/0.09	0.00	0.01(0.03)	-0.06/0.11	0.01	0.01(0.03)	-0.06/0.07	0.01
			t_4	0.00(0.03)	-0.10/0.09	0.00	0.00(0.04)	-0.08/0.09	0.00	0.00(0.03)	-0.08/0.08	0.00	0.00(0.03)	-0.08/0.09	0.00
		Misfit	t_1	-0.04(0.03)	-0.14/0.05	-0.04	-0.03(0.03)	-0.11/0.04	-0.03	-0.04(0.03)	-0.13/0.04	-0.04	-0.04(0.03)	-0.10/0.04	-0.04
			t_2	-0.01(0.03)	-0.10/0.06	-0.01	-0.02(0.03)	-0.09/0.06	-0.02	-0.01(0.04)	-0.12/0.09	-0.01	0.00(0.03)	-0.09/0.06	0.00
			t_3	-0.01(0.03)	-0.09/0.07	-0.01	0.00(0.03)	-0.09/0.10	0.00	-0.01(0.04)	-0.08/0.09	-0.01	0.00(0.03)	-0.10/0.09	0.00
			t_4	-0.01(0.03)	-0.10/0.07	-0.01	-0.01(0.03)	-0.10/0.08	-0.01	-0.01(0.04)	-0.09/0.09	-0.01	-0.01(0.04)	-0.10/0.08	-0.01
Mean/Mean	500	Fit	t_1	0.00(0.08)	-0.18/0.24	0.00	-0.02(0.09)	-0.18/0.34	-0.02	0.01(0.07)	-0.14/0.16	0.01	0.01(0.09)	-0.19/0.20	0.01
			t_2	0.00(0.09)	-0.20/0.21	0.00	-0.01(0.08)	-0.19/0.29	-0.01	0.01(0.10)	-0.20/0.31	0.01	0.00(0.07)	-0.20/0.17	0.00
			t_3	0.01(0.09)	-0.22/0.21	0.01	-0.01(0.07)	-0.17/0.17	-0.01	0.01(0.09)	-0.21/0.24	0.01	0.00(0.09)	-0.18/0.23	0.00
			t_4	0.01(0.08)	-0.18/0.17	0.01	-0.02(0.08)	-0.27/0.15	-0.02	0.01(0.08)	-0.20/0.26	0.01	0.00(0.08)	-0.18/0.18	0.00
		Misfit	t_1	-0.03(0.08)	-0.19/0.20	-0.03	-0.03(0.07)	-0.21/0.23	-0.03	-0.03(0.08)	-0.17/0.21	-0.03	-0.03(0.08)	-0.27/0.23	-0.03
			t_2	0.00(0.08)	-0.18/0.25	0.00	-0.01(0.08)	-0.20/0.18	-0.01	-0.02(0.08)	-0.16/0.27	-0.02	-0.01(0.10)	-0.21/0.27	-0.01
			t_3	-0.02(0.09)	-0.21/0.23	-0.02	0.00(0.08)	-0.18/0.18	0.00	0.00(0.08)	-0.18/0.21	0.00	0.00(0.09)	-0.24/0.21	0.00
			t_4	-0.02(0.09)	-0.20/0.19	-0.02	-0.01(0.08)	-0.21/0.20	-0.01	0.01(0.09)	-0.23/0.28	0.01	0.00(0.09)	-0.19/0.20	0.00
	3,000	Fit	t_1	0.00(0.03)	-0.07/0.07	0.00	0.00(0.03)	-0.08/0.07	0.00	0.00(0.03)	-0.06/0.08	0.00	0.00(0.03)	-0.10/0.07	0.00
			t_2	0.00(0.03)	-0.07/0.07	0.00	0.00(0.04)	-0.10/0.07	0.00	0.00(0.03)	-0.07/0.09	0.00	0.00(0.03)	-0.07/0.06	0.00
			t_3	0.00(0.03)	-0.08/0.07	0.00	0.00(0.03)	-0.10/0.09	0.00	0.01(0.03)	-0.07/0.11	0.01	0.01(0.03)	-0.07/0.08	0.01
			t_4	0.00(0.03)	-0.09/0.10	0.00	0.00(0.03)	-0.08/0.09	0.00	0.00(0.03)	-0.08/0.06	0.00	0.00(0.04)	-0.08/0.10	0.00
		Misfit	t_1	-0.04(0.03)	-0.14/0.05	-0.04	-0.03(0.03)	-0.11/0.04	-0.03	-0.04(0.03)	-0.13/0.04	-0.04	-0.04(0.03)	-0.10/0.04	-0.04
			t_2	-0.02(0.03)	-0.10/0.06	-0.02	-0.01(0.03)	-0.08/0.06	-0.01	-0.01(0.04)	-0.12/0.08	-0.01	-0.01(0.03)	-0.08/0.06	-0.01
			t_3	-0.01(0.03)	-0.09/0.07	-0.01	0.00(0.03)	-0.09/0.11	0.00	-0.01(0.04)	-0.08/0.08	-0.01	0.00(0.03)	-0.10/0.09	0.00
			t_4	-0.01(0.04)	-0.10/0.08	-0.01	-0.01(0.03)	-0.09/0.08	-0.01	-0.01(0.04)	-0.09/0.09	-0.01	-0.01(0.04)	-0.10/0.08	-0.01

Linking Method	*N*	Rasch model	Time Point	Anchor: 3 *M*(*SD*)	Anchor: 3 *Min*/*Max*	Anchor: 3 *RB*	Anchor: 5 *M*(*SD*)	Anchor: 5 *Min*/*Max*	Anchor: 5 *RB*	Anchor: 7 *M*(*SD*)	Anchor: 7 *Min*/*Max*	Anchor: 7 *RB*	Anchor: 9 *M*(*SD*)	Anchor: 9 *Min*/*Max*	Anchor: 9 *RB*
weighted Mean/Mean	500	Fit	t_1	0.00(0.08)	-0.18/0.24	0.00	-0.02(0.09)	-0.18/0.34	-0.02	0.01(0.07)	-0.14/0.16	0.01	0.01(0.09)	-0.19/0.20	0.01
			t_2	0.00(0.09)	-0.20/0.21	0.00	-0.01(0.08)	-0.19/0.29	-0.01	0.01(0.10)	-0.20/0.31	0.01	0.00(0.07)	-0.20/0.17	0.00
			t_3	0.01(0.09)	-0.22/0.21	0.01	-0.01(0.07)	-0.17/0.17	-0.01	0.01(0.09)	-0.21/0.24	0.01	0.00(0.09)	-0.18/0.23	0.00
			t_4	0.01(0.08)	-0.18/0.17	0.01	-0.02(0.08)	-0.27/0.15	-0.02	0.01(0.08)	-0.20/0.26	0.01	0.00(0.08)	-0.18/0.18	0.00
		Misfit	t_1	-0.03(0.08)	-0.19/0.20	-0.03	-0.03(0.07)	-0.21/0.23	-0.03	-0.03(0.08)	-0.17/0.21	-0.03	-0.03(0.08)	-0.27/0.23	-0.03
			t_2	0.00(0.08)	-0.18/0.25	0.00	-0.01(0.08)	-0.20/0.18	-0.01	-0.02(0.08)	-0.16/0.27	-0.02	-0.01(0.10)	-0.21/0.27	-0.01
			t_3	-0.02(0.09)	-0.21/0.23	-0.02	0.00(0.08)	-0.18/0.18	0.00	0.00(0.08)	-0.18/0.21	0.00	0.00(0.09)	-0.24/0.21	0.00
			t_4	-0.02(0.09)	-0.20/0.19	-0.02	-0.01(0.08)	-0.21/0.20	-0.01	0.01(0.09)	-0.23/0.28	0.01	0.00(0.09)	-0.19/0.20	0.00
	3,000	Fit	t_1	0.00(0.03)	-0.07/0.07	0.00	0.00(0.03)	-0.08/0.07	0.00	0.00(0.03)	-0.06/0.08	0.00	0.00(0.03)	-0.10/0.07	0.00
			t_2	0.00(0.03)	-0.07/0.07	0.00	0.00(0.04)	-0.10/0.07	0.00	0.00(0.03)	-0.07/0.09	0.00	0.00(0.03)	-0.07/0.06	0.00
			t_3	0.00(0.03)	-0.08/0.07	0.00	0.00(0.03)	-0.10/0.09	0.00	0.01(0.03)	-0.07/0.11	0.01	0.01(0.03)	-0.07/0.08	0.01
			t_4	0.00(0.03)	-0.09/0.10	0.00	0.00(0.03)	-0.08/0.09	0.00	0.00(0.03)	-0.08/0.06	0.00	0.00(0.04)	-0.08/0.10	0.00
		Misfit	t_1	-0.04(0.03)	-0.14/0.05	-0.04	-0.03(0.03)	-0.11/0.04	-0.03	-0.04(0.03)	-0.13/0.04	-0.04	-0.04(0.03)	-0.10/0.04	-0.04
			t_2	-0.02(0.03)	-0.10/0.06	-0.02	-0.01(0.03)	-0.08/0.06	-0.01	-0.01(0.04)	-0.12/0.08	-0.01	-0.01(0.03)	-0.08/0.06	-0.01
			t_3	-0.01(0.03)	-0.09/0.07	-0.01	0.00(0.03)	-0.09/0.11	0.00	-0.01(0.04)	-0.08/0.08	-0.01	0.00(0.03)	-0.10/0.09	0.00
			t_4	-0.01(0.04)	-0.10/0.08	-0.01	-0.01(0.03)	-0.09/0.08	-0.01	-0.01(0.04)	-0.09/0.09	-0.01	-0.01(0.04)	-0.10/0.08	-0.01

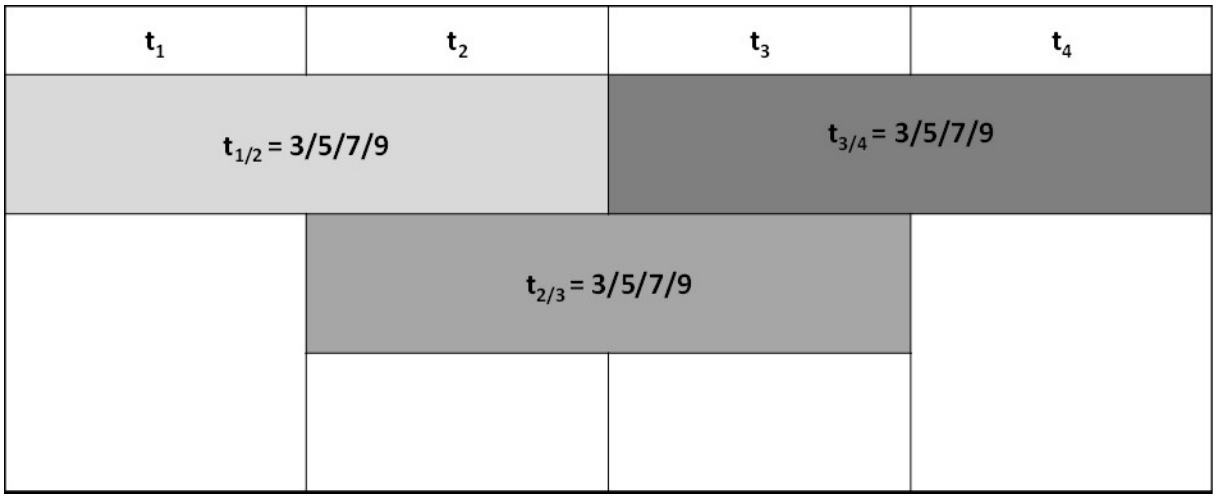
Supplementary Table 5. Descriptive Statistics for the RMSE of Sample Variance Split by Experimental Conditions and Linking Methods. Anchor: = Number of anchor items used for linking; *M*(*SD*) = mean and standard deviation of the RMSE of sample variance; *Min/Max* = *minimum*/*maximum* of the RMSE of sample variance; FPC = fixed parameter calibration; Mean/Mean = mean/mean linking; weighted Mean/Mean = weighted m/m. The RMSE was averaged over 100 replications.

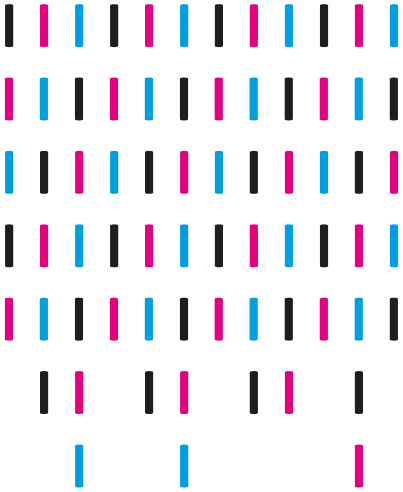
Linking Method	*N*	Rasch model	Time Point	Anchor: 3 *M*(*SD*)	Anchor: 3 *Min*/*Max*	Anchor: 5 *M*(*SD*)	Anchor: 5 *Min*/*Max*	Anchor: 7 *M*(*SD*)	Anchor: 7 *Min*/*Max*	Anchor: 9 *M*(*SD*)	Anchor: 9 *Min*/*Max*
FPC	500	Fit	t_1	0.07(0.05)	0.00/0.24	0.08(0.06)	0.00/0.34	0.06(0.04)	0.00/0.16	0.07(0.05)	0.0/0.20
			t_2	0.07(0.05)	0.00/0.23	0.07(0.05)	0.00/0.29	0.07(0.06)	0.00/0.32	0.06(0.04)	0.0/0.18
			t_3	0.07(0.05)	0.00/0.21	0.06(0.04)	0.00/0.19	0.07(0.06)	0.00/0.26	0.07(0.05)	0.0/0.22
			t_4	0.06(0.05)	0.00/0.18	0.07(0.05)	0.00/0.25	0.07(0.05)	0.00/0.28	0.07(0.05)	0.0/0.18
		Misfit	t_1	0.07(0.05)	0.00/0.20	0.06(0.05)	0.00/0.23	0.07(0.05)	0.00/0.21	0.07(0.05)	0.0/0.27
			t_2	0.07(0.05)	0.00/0.25	0.06(0.05)	0.00/0.20	0.06(0.05)	0.00/0.26	0.08(0.06)	0.0/0.26

Linking Method	*N*	Rasch model	Time Point	Anchor: 3 *M*(*SD*)	Anchor: 3 *Min*/*Max*	Anchor: 5 *M*(*SD*)	Anchor: 5 *Min*/*Max*	Anchor: 7 *M*(*SD*)	Anchor: 7 *Min*/*Max*	Anchor: 9 *M*(*SD*)	Anchor: 9 *Min*/*Max*
			t_3	0.07(0.05)	0.00/0.24	0.07(0.05)	0.00/0.19	0.07(0.05)	0.00/0.19	0.07(0.05)	0.0/0.21
			t_4	0.07(0.05)	0.00/0.21	0.06(0.05)	0.00/0.21	0.07(0.06)	0.00/0.26	0.07(0.05)	0.0/0.23
		Fit	t_1	0.03(0.02)	0.00/0.07	0.03(0.02)	0.00/0.08	0.03(0.02)	0.00/0.08	0.03(0.02)	0.0/0.10
			t_2	0.03(0.02)	0.00/0.07	0.03(0.02)	0.00/0.10	0.03(0.02)	0.00/0.09	0.03(0.02)	0.0/0.07
			t_3	0.03(0.02)	0.00/0.08	0.03(0.02)	0.00/0.09	0.03(0.02)	0.00/0.11	0.03(0.02)	0.0/0.07
	3,000		t_4	0.03(0.02)	0.00/0.10	0.03(0.02)	0.00/0.09	0.03(0.02)	0.00/0.08	0.03(0.02)	0.0/0.09
		Misfit	t_1	0.04(0.03)	0.00/0.14	0.03(0.03)	0.00/0.11	0.04(0.03)	0.00/0.13	0.04(0.03)	0.0/0.10
			t_2	0.03(0.02)	0.00/0.10	0.03(0.02)	0.00/0.09	0.03(0.02)	0.00/0.12	0.03(0.02)	0.0/0.09
			t_3	0.03(0.02)	0.00/0.09	0.03(0.02)	0.00/0.10	0.03(0.02)	0.00/0.09	0.03(0.02)	0.0/0.10
			t_4	0.03(0.02)	0.00/0.10	0.03(0.02)	0.00/0.10	0.03(0.02)	0.00/0.09	0.03(0.02)	0.0/0.10
		Fit	t_1	0.07(0.05)	0.00/0.24	0.08(0.06)	0.00/0.34	0.06(0.04)	0.00/0.16	0.07(0.05)	0.0/0.20
			t_2	0.07(0.05)	0.00/0.21	0.06(0.05)	0.00/0.29	0.08(0.06)	0.00/0.31	0.06(0.04)	0.0/0.20
			t_3	0.07(0.05)	0.00/0.22	0.06(0.04)	0.00/0.17	0.07(0.06)	0.00/0.24	0.07(0.05)	0.0/0.23
	500		t_4	0.06(0.05)	0.00/0.18	0.07(0.05)	0.00/0.27	0.07(0.05)	0.00/0.26	0.07(0.05)	0.0/0.18
		Misfit	t_1	0.07(0.05)	0.00/0.20	0.06(0.05)	0.00/0.23	0.07(0.05)	0.00/0.21	0.07(0.05)	0.0/0.27
			t_2	0.07(0.05)	0.00/0.25	0.06(0.05)	0.00/0.20	0.06(0.05)	0.00/0.27	0.08(0.06)	0.0/0.27
			t_3	0.07(0.05)	0.00/0.23	0.06(0.05)	0.00/0.18	0.07(0.05)	0.00/0.21	0.07(0.05)	0.0/0.24
Mean/Mean			t_4	0.07(0.05)	0.00/0.20	0.07(0.05)	0.00/0.21	0.07(0.06)	0.00/0.28	0.07(0.05)	0.0/0.20
		Fit	t_1	0.03(0.02)	0.00/0.07	0.03(0.02)	0.00/0.08	0.03(0.02)	0.00/0.08	0.03(0.02)	0.0/0.10
			t_2	0.03(0.02)	0.00/0.07	0.03(0.02)	0.00/0.10	0.03(0.02)	0.00/0.09	0.03(0.02)	0.0/0.07
			t_3	0.03(0.02)	0.00/0.08	0.03(0.02)	0.00/0.10	0.03(0.02)	0.00/0.11	0.03(0.02)	0.0/0.08
	3,000		t_4	0.03(0.02)	0.00/0.10	0.03(0.02)	0.00/0.09	0.03(0.02)	0.00/0.08	0.03(0.02)	0.0/0.10
		Misfit	t_1	0.04(0.03)	0.00/0.14	0.03(0.03)	0.00/0.11	0.04(0.03)	0.00/0.13	0.04(0.03)	0.0/0.10
			t_2	0.03(0.02)	0.00/0.10	0.03(0.02)	0.00/0.08	0.03(0.02)	0.00/0.12	0.03(0.02)	0.0/0.08
			t_3	0.03(0.02)	0.00/0.09	0.02(0.02)	0.00/0.11	0.03(0.02)	0.00/0.08	0.03(0.02)	0.0/0.10
			t_4	0.03(0.02)	0.00/0.10	0.03(0.02)	0.00/0.09	0.03(0.02)	0.00/0.09	0.03(0.02)	0.0/0.10
		Fit	t_1	0.07(0.05)	0.00/0.24	0.08(0.06)	0.00/0.34	0.06(0.04)	0.00/0.16	0.07(0.05)	0.0/0.20
			t_2	0.07(0.05)	0.00/0.21	0.06(0.05)	0.00/0.29	0.08(0.06)	0.00/0.31	0.06(0.04)	0.0/0.20
weighted			t_3	0.07(0.05)	0.00/0.22	0.06(0.04)	0.00/0.17	0.07(0.06)	0.00/0.24	0.07(0.05)	0.0/0.23
Mean/Mean	500		t_4	0.06(0.05)	0.00/0.18	0.07(0.05)	0.00/0.27	0.07(0.05)	0.00/0.26	0.07(0.05)	0.0/0.18
		Misfit	t_1	0.07(0.05)	0.00/0.20	0.06(0.05)	0.00/0.23	0.07(0.05)	0.00/0.21	0.07(0.05)	0.0/0.27
			t_2	0.07(0.05)	0.00/0.25	0.06(0.05)	0.00/0.20	0.06(0.05)	0.00/0.27	0.08(0.06)	0.0/0.27

Linking Method	*N*	Rasch model	Time Point	Anchor: 3		Anchor: 5		Anchor: 7		Anchor: 9	
				***M*(*SD*)**	***Min*/*Max***	***M*(*SD*)**	***Min*/*Max***	***M*(*SD*)**	***Min*/*Max***	***M*(*SD*)**	***Min*/*Max***
			t_3	0.07(0.05)	0.00/0.23	0.06(0.05)	0.00/0.18	0.07(0.05)	0.00/0.21	0.07(0.05)	0.0/0.24
			t_4	0.07(0.05)	0.00/0.20	0.07(0.05)	0.00/0.21	0.07(0.06)	0.00/0.28	0.07(0.05)	0.0/0.20
	3,000	Fit	t_1	0.03(0.02)	0.00/0.07	0.03(0.02)	0.00/0.08	0.03(0.02)	0.00/0.08	0.03(0.02)	0.0/0.10
			t_2	0.03(0.02)	0.00/0.07	0.03(0.02)	0.00/0.10	0.03(0.02)	0.00/0.09	0.03(0.02)	0.0/0.07
			t_3	0.03(0.02)	0.00/0.08	0.03(0.02)	0.00/0.10	0.03(0.02)	0.00/0.11	0.03(0.02)	0.0/0.08
			t_4	0.03(0.02)	0.00/0.10	0.03(0.02)	0.00/0.09	0.03(0.02)	0.00/0.08	0.03(0.02)	0.0/0.10
		Misfit	t_1	0.04(0.03)	0.00/0.14	0.03(0.03)	0.00/0.11	0.04(0.03)	0.00/0.13	0.04(0.03)	0.0/0.10
			t_2	0.03(0.02)	0.00/0.10	0.03(0.02)	0.00/0.08	0.03(0.02)	0.00/0.12	0.03(0.02)	0.0/0.08
			t_3	0.03(0.02)	0.00/0.09	0.02(0.02)	0.00/0.11	0.03(0.02)	0.00/0.08	0.03(0.02)	0.0/0.10
			t_4	0.03(0.02)	0.00/0.10	0.03(0.02)	0.00/0.09	0.03(0.02)	0.00/0.09	0.03(0.02)	0.0/0.10

Supplementary Figure 1. Anchor-Items Design of the Simulation Study. Four test forms were administered from t_1 to t_4, consisting of 25 items each, sharing either a number of 3, 5, 7 or 9 common items with their adjacent test form(s). Blank squares symbolize test form unique items.





NEPS SURVEY PAPERS

Luise Fischer, Theresa Rohm, Timo Gnambs, &

Claus H. Carstensen

LINKING THE DATA OF THE COMPETENCE TESTS

NEPS Survey Paper No. 1
Bamberg, April 2016

LIfBi
LEIBNIZ INSTITUTE FOR
EDUCATIONAL TRAJECTORIES

NEPS
National Educational Panel Study

Survey Papers of the German National Educational Panel Study (NEPS)
at the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg

The NEPS Survey Paper Series provides articles with a focus on methodological aspects and data handling issues related to the German National Educational Panel Study (NEPS).

The NEPS Survey Papers are edited by a review board consisting of the scientific management of LIfBi and NEPS.

They are of particular relevance for the analysis of NEPS data as they describe data editing and data collection procedures as well as instruments or tests used in the NEPS survey. Papers that appear in this series fall into the category of 'grey literature' and may also appear elsewhere.

The NEPS Survey Papers are available at https://www.neps-data.de (see section “Publications“).

Editor-in-Chief: Corinna Kleinert, LIfBi/University of Bamberg/IAB Nuremberg

Contact: German National Educational Panel Study (NEPS) – Leibniz Institute for Educational Trajectories – Wilhelmsplatz 3 – 96047 Bamberg – Germany – contact@lifbi.de

Linking the Data of the Competence Tests

Luise Fischer, Theresa Rohm, Timo Gnambs, & Claus H. Carstensen

Leibniz Institute for Educational Trajectories, Bamberg, Germany

E-mail address of lead author:

luise.fischer@lifbi.de

Bibliographic data:

Fischer, L., Rohm, T., Gnambs, T., & Carstensen, C. H. (2016). *Linking the data of the competence tests* (NEPS Survey Paper No. 1). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study. doi:10.5157/NEPS:SP01:1.0

Linking the Data of the Competence Tests

Abstract

The National Educational Panel Study (NEPS) aims at investigating the development of competencies across the whole life span and developing tests for assessing different competence domains. In order to compare competencies across different measurement occasions and examine competence development over time the different measurements must be placed onto a common scale. This goal is achieved by linking two measurements of the same construct. The present document describes the linking procedure adopted for the NEPS. Moreover, this approach is demonstrated using data from Starting Cohort 3 that links mathematical and reading competences across Grades 5 and 7. First, the procedure on how to derive linked item parameters in an anchor-items design is described. After showing that the mathematical tests administered in Grades 5 and 7 are unidimensional, it is demonstrated that all common items administered in both grades showed negligible differential item functioning. Therefore, the two mathematical tests were linked using six measurement invariant items. Subsequently, the procedure on how to derive linked item parameters in an anchor-groups design is described. It is demonstrated that the reading competence tests administered in Grades 5 and 7 were unidimensional and showed no differential item functioning. Therefore, the two reading tests were linked using responses from an independent link sample. Finally, the naming conventions in the scientific use file for repeatedly administered items and different competence scores are presented.

Keywords

link design, link method, item response theory, longitudinal

1. Rationale for Linking

Within the National Educational Panel Study (NEPS; Blossfeld et al., 2011) different competences (e.g., reading, mathematics) are measured across the life span. The respective competence tests are constructed in such a way as to allow for an accurate assessment of these competences within each age group. Therefore, respondents from different age groups typically do not receive identical tests. For example, a test that is optimally challenging for students is likely to be too difficult for adolescents; similarly, a test targeted at adolescents is probably too easy for students. Rather, the NEPS administers different tests to participants that are specifically targeted at their age and competence level. As a consequence, the competence scores from different measurement occasions cannot be directly compared; differences in observed scores would reflect differences in competences as well as differences in item difficulties. In order to examine developmental trajectories and compare competences across different measurement occasions, the different measurements must be placed onto a common scale. This goal is achieved by linking two measurements of the same construct. Thus, linking is a necessary prerequisite to compare the competence data in the NEPS across the life course and investigate educational trajectories (Pohl, Haberkorn, & Carstensen, 2015).

The present study describes the linking procedure adopted for the NEPS. Moreover, this approach is demonstrated using data from Starting Cohort 3 that links mathematical and reading competences across Grades 5 and 7.

2. NEPS Linking Procedure

Specific test designs and statistical procedures are needed to place different measurements on a common scale. In the NEPS two different test designs, the *anchor-items design* and the *anchor-group design*, with their corresponding link methods are used.

2.1 Link designs

In order to link competence scores from two tests common information on the two tests is needed; that is, the same respondents must provide answers to at least a subset of items from both tests. This common information can subsequently be used to place the two measurements on a common scale. In the NEPS competence tests for, among others, the domains reading, mathematics, scientific literacy, information and communication technologies (ICT), and English are administered. Because memory effects are to be expected in some domains if the same item is administered repeatedly to the same respondents (see Pohl et al., 2015), two different link designs are adopted to build an overlap of information between different measurement points.

2.1.1 Anchor-items design

For items measuring mathematical competence in secondary education in NEPS no memory effects are to be expected, since the items are similar to the tasks typically used in schools and the time interval between assessments is rather long with two years. Therefore, it is feasible to include a subset of items from a previously administered test in the test administered at a subsequent measurement occasion. If various assumptions are met (see section 2.3), these “common items” that are included in both tests can be used to link the two tests and create a common scale.

2.1.2 Anchor-group design

For the competence domains of reading, science literacy, and ICT an anchor-group design is needed because memory effects might distort responses if the same items are repeatedly administered to the same respondents. Therefore, common information on two tests is created using an independent link sample that is not part of the original sample whose responses are to be linked. This link sample is drawn from the same population as the respective starting cohort. In the NEPS Linking studies, the age of the link sample either matches the participants' age taking the latter test or falls somewhere between the age groups of the two measurement occasions. Both tests are administered the link sample within a single measurement occasion. Therefore, no developmental changes can occur that might influence the relationship between the two tests.

2.2 Link method

Because Scientific Use Files (SUFs) are published sequentially, the data of a former measurement point is already scaled and released, when data of a latter measurement point is available for scaling. In order to leave the earlier released reference scale unchanged, the data of subsequent measurement points are linked to that initial scale. Leaving the reference scale unchanged imposes some restrictions on potential link methods: If the data from both measurement points were scaled within a single analysis (i.e., "concurrent calibration"; Kolen & Brennan, 2014), the reference scale would change. Moreover, in the NEPS competence tests are scaled using models of the Rasch family (e.g., the Partial Credit model; PCM; Masters, 1982). Therefore, link methods that change the variance of item difficulties and person abilities (e.g., "mean/sigma linking"; Marco, 1977) would risk biasing the measurement of competence development. Therefore, these approaches were not taken into account. In a series of preliminary studies different procedures were evaluated in terms of their suitability for linking different competence domains, such as reading, science, or mathematics, across multiple measurement occasions (cf. Fischer, Rohm, & Carstensen, 2015a, 2015b). Based on these analyses the method of "mean/mean linking" (Loyd & Hoover, 1980) was adopted for the NEPS.

2.2.1 Linking in the anchor-items design

For competence domains using an anchor-items design two independently scaled tests (i.e., the mathematics tests at the first and the second measurement occasion) are linked using a linear transformation of the item parameters of the second test. To link and, therefore, map the latter scale to its reference scale it is assumed that the items' difficulties are the same in both assessments and that all changes in response frequencies can be attributed to change of the person competences over time. To set the item difficulties equal between assessments, a correction term c is derived using the common items that were included in both tests. After independent calibrations of both assessments, the correction term c is derived based on those respondents that participated at both measurement occasions (also referred to as the longitudinal subsample). The correction term c is computed as the difference of the mean of the item difficulty parameters at the first measurement occasion, $M(\sigma_{1j})$, and the mean of the item difficulty parameters at the second measurement occasion, $M(\sigma_{2j})$:

$$c = M(\sigma_{1j}) - M(\sigma_{2j}) \quad (1)$$

The correction term c is then added to each item difficulty parameter of the test to be linked. This approach links two Rasch model calibrated assessments in an anchor-items design using the "mean/mean" method (Loyd & Hoover, 1980). Because "mean/mean" linking depends upon the differences in difficulties between the common items, the choice of the common items will influence the link result to some degree; that is, another set of common items could result in a slightly different correction term c. This uncertainty in the link due to the sampling of common items is reflected by the link error. The link error is based on the differences between the linked item parameters for the k common items at the first and the second measurement occasion as $\Delta\sigma_j = \sigma_{1j} - (\sigma_{2j} + c)$. Following PISA 2006 (OECD, 2014) the link error is then given as the standard deviation of these differences, $SD(\Delta\sigma_j)$, standardized at the number of common items:

$$link\ error = \frac{SD(\Delta\sigma_j)}{\sqrt{k}} \tag{2}$$

2.2.2 Linking in the anchor-group design

For domains using an anchor-group design two independently scaled tests are also linked using a linear transformation of the item parameters in the second test. However, because no common items are included in the two tests, the correction term c is derived using the responses from an independent link sample. Let A and B represent the items from the two tests that were administered at the first respectively the second measurement occasion. In the main sample (i.e., the respective starting cohort) A and B are scaled independently at the two measurement occasions. Again, only the longitudinal subsample that participated at both measurement occasions is included in these analyses. In contrast, in the link sample that took both tests all items, A and B, are scaled concurrently; thus, all items are included in a single scaling model. Subsequently, means of the item difficulty parameters for A and B are computed in the main sample and the link sample, that is, (a) the mean difficulty parameters for A in the main sample, $M(\sigma_{MS,A})$, (b) the mean item difficulty parameters for B in the main sample, $M(\sigma_{MS,B})$, (c) the mean item difficulty parameters for A in the link sample, $M(\sigma_{LS,A})$, and (d) the mean item difficulty parameters for B in the link sample, $M(\sigma_{LS,B})$. Then, the correction term c is computed as

$$c = M(\sigma_{MS,A}) - M(\sigma_{MS,B}) + M(\sigma_{LS,B}) - M(\sigma_{LS,A}). \tag{3}$$

The first two terms in (3) insert the reference scale to the test that is to be linked, whereas the last two terms add the information from the link study. The latter reflects the relation between both tests that is not affected by time. The correction term c is then added to each item difficulty parameter of the test intended to be linked. Again, the uncertainty in the link due to the sampling of items is quantified by the link error. In an anchor-group design, the link error is calculated as the pooled link error for the k_A items in A and the k_B items in B. Thus, two sets of differences are derived: (a) the differences between the item parameters of A in the main sample and the link sample, $\Delta\sigma_j = \sigma_{MS,j} - \sigma_{LS,j}$ and (b) the differences between the linked item parameters of B in the main sample and the link sample, $\Delta\sigma_i = (\sigma_{MS,i} + c) - \sigma_{LS,i}$. The link error is then computed as:

$$link\ error\ = \sqrt{\left(\frac{SD(\Delta\sigma_j)}{\sqrt{k_A}}\right)^2 + \left(\frac{SD(\Delta\sigma_i)}{\sqrt{k_B}}\right)^2}. \quad (4)$$

2.3 Assumptions for linking two scales

2.3.1 Unidimensionality

To measure competence development within a domain the meaning of the underlying construct must not change over time. As a consequence, two tests that are to be linked need to be unidimensional. Unidimensionality is examined in two different ways:

In case of an anchor-items design the common items are used as an anchor test. Therefore, at each measurement occasion a one- and a two-dimensional model may be compared. For the two-dimensional model, the common items load on the first dimension and the unique items (i.e., the items included in only one test) load on the second dimension. If a model fit evaluation (i.e. Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC)) supports the one-dimensional model, unidimensionality can be assumed. Moreover, the residuals of the one-dimensional model should exhibit approximately zero-order correlations as indicated by Yen's (1984) Q_3. In case of an anchor-group design unidimensionality is evaluated with regard to the two tests administered to the link sample. Again, information criteria as well as residuals can be inspected.

2.3.2 Measurement invariance

Common items that are supposed to link two tests must exhibit measurement invariance. If a common item is not measurement invariant, it cannot be used to link two tests. An item is considered measurement invariant, when its relative position among the other items on the logit scale does not change between tests.

Using an anchor-items design, measurement invariance for a common item *j* is examined by comparing the item difficulty parameter resulting from a separate scaling of the first measurement occasion, σ_{1j}, and the parameter from the second measurement occasion, σ_{2j}. Because traditional significance tests yield an excessive power with large samples even for negligible effects, we adopted the DIF classification system of the Educational Testing Service (ETS; Holland & Wainer, 1993) that relies on an effect size in the delta metric and a significance test. The delta metric has a mean of 13 and a standard deviation of 4. Thus, one point on the delta scale corresponds to a quarter of a standard deviation. This is equivalent to Cohen's $d = 0.25$ and to $\eta^2 = 0.0154$. Following the ETS classification system items that showed a significant DIF effect greater than 1 point on the delta scale were classified as having moderate DIF. Instead of a classical null hypothesis test of no DIF, we adopted Murphy and Myors' (1999) minimum effect null hypothesis and tested for the presence of negligible DIF. Thus, a value of 1.54 percent of variance explained was used as criterion for negligible DIF. Following Lord (1980) we computed a Wald statistic for each item as

$$t_j = \frac{\sigma_{1j} - \sigma_{2j}}{\sqrt{SE(\sigma_{1j})^2 + SE(\sigma_{2j})^2}}. \quad (5)$$

The resulting t value was squared and, thus, transformed into an F distribution (with df_1 = number of measurement points and df_2 = number of participants). Adopting a value of 1.54 percent of variance explained as a minimum effect criterion, a non-central F test was used to test the assumption of non-negligible DIF (see Fischer, Gnambs, Rohm, & Carstensen, 2016).

In anchor-group designs all items can be considered common items. Therefore, measurement invariance is tested the same way as described above by comparing the item difficulty parameters from the main sample (i.e., the respective starting cohort) and the link sample. Thus, examining measurement invariance in an anchor-group design is identical to examining differential item functioning (DIF) between the two groups.

Because the link information is derived on a subsample of respondents that participated at both measurement occasions (i.e., the longitudinal subsample), it is necessary that all items exhibit measurement invariance across the whole sample, that is, between respondents that participated at only one measurement occasion and the longitudinal subsample that provided data at both measurement occasions. This is tested by estimating the item parameters separately in each subsample and examining measurement invariance as described above.

2.3.2.1 Naming Conventions

In the SUFs repeatedly administered items retain their initial variable names. To identify common items across measurement occasions, a suffix is appended to the variable name that indicates the current sample. For example: Item “mag5q301_c” was first administered in Grade 5. Because the same item was also administered in Grade 7, the variable name for the second administration is extended by the suffix ”sc3g7” (= Starting Cohort 3, Grade 7) to “mag5q301_sc3g7_c”.

3. Linking of Starting Cohort 3 – Grade 5 and Grade 7

In the following section we demonstrate in two examples how to apply the linking procedures described above to the two link designs. First, we apply the method of “mean/mean” linking to two tests on mathematical competence measured in an anchor-items design. Then, the linking procedure in an anchor-items design is illustrated using two tests on reading competence. In Starting Cohort 3, mathematical and reading competences were measured in Grade 5 and two years later in Grade 7. First, the data were scaled independently for each grade following the NEPS scaling procedure described in Pohl and Carstensen (2012). Subsequently, the competence scores were linked across the two grades whereby the scale of Grade 5 served as the underlying reference scale for Grade 7. Mathematical competences were linked using an anchor-items design, whereas reading competences were linked using an anchor-groups design.

3.1 Linking the test of mathematical competence

The scaling results of the mathematics test in Grade 5 are described in Duchhardt and Gerdes (2012) and the respective results for Grade 7 are outlined in Schnittjer, Gerken, and Fischer (2015).

3.1.1 Sample

A sample of 5,193 participants received the mathematics test in Grade 5 and 6,191 participants finished the test in Grade 7. Overall, 3,833 respondents participated at both measurement occasions. Participants with less than three valid responses were excluded from the linking procedure.

3.1.2 Test instruments

The mathematics test in Grades 5 and 7 included 24 and 23 items, respectively. Six items were included in both tests (see Table 1).

3.1.3 Results

3.1.3.1 Unidimensionality

For each grade, we estimated a one-dimensional model that specified a single latent factor for all items and also a two-dimensional model that specified separate latent factors for the common items and the unique items (i.e., the items that were included at only one measurement occasion). In both grades the information criteria favored the two-dimensional model, AIC = 156983.15 and BIC = 157164.88 for Grade 5, and AIC = 132409.71 and BIC = 132599.81 for Grade 7, over the one-dimensional model, AIC = 157104.52 and BIC = 157272.79 for Grade 5, and AIC = 132466.18 and BIC = 132643.17 for Grade 7. Therefore, we also examined the residual correlations for the one-dimensional models. The corrected Q_3 statistics indicated largely unidimensional scales in Grade 5, $M(Q_3) = 0$, $SD(Q_3) = 0.03$, and Grade 7, $M(Q_3) = 0$, $SD(\mathrm{Q}_3) = 0.02$. This indicates that unidimensional scales can be assumed for the mathematics tests in Grades 5 and 7.

3.1.3.2 Measurement invariance

First, the mathematics tests in Grades 5 and 7 were scaled separately in the longitudinal subsample. Subsequently, the difficulty parameters of the common items were centered in each grade (i.e., their means were set to zero). The differences in item difficulties between Grades 5 and 7, and the tests for measurement invariance based on the Wald statistic (see Equation 3) are summarized in Table 1. For the six common items measurement invariance was supported (i.e., the minimum effects hypothesis test was not significant).

Table 1: DIF Analyses for the common items in the tests for mathematical competence in Grades 5 and 7

Grade 5	Grade 7	Δσ	$SE_{\Delta\sigma}$	*F*	*p (α = .05)*
mag5q301_c	mag5q301_sc3g7_c	0.03	0.05	0.24	*> .999*
mag5d051_c	mag5d051_sc3g7_c	-0.31	0.09	12.49	*> .999*
mag5d052_c	mag5d052_sc3g7_c	0.42	0.06	51.36	*0.72*
mag5r251_c	mag5r251_sc3g7_c	-0.12	0.05	5.30	*> .999*
mag5v321_c	mag5v321_sc3g7_c	-0.10	0.06	3.12	*> .999*
mag5r191_c	mag5r191_sc3g7_c	0.09	0.06	2.34	*> .999*

Note. Δσ = Difference in item difficulty parameters between Grades 5 and 7 (positive values indicate easier items in Grade 7); SEΔσ = Pooled standard error; F = Test statistic for the minimum effects hypothesis test based on (5); Fcrit = Critical value for the minimum effects hypothesis test for an α of .05; the degrees of freedom (df1, df2) are based on the number

of measurement points (df1 = k-1) and the number of test takers taking both tests (df2 = n-1). The critical F0154(1, 3,832) = 88.3. A non-significant test indicates measurement invariance.

Furthermore, measurement invariance for Grade 7 was supported for the two groups of one-time participants (n = 2,358) and the longitudinal subsample (n = 3,833), with none of the F-statistics exceeding the critical value of F0154(1, 6,189) = 132.1.

3.1.3.3 Linking the two tests

The mathematics tests administered in the two grades were linked using the "mean/mean" method (see section 2.2.1). In the longitudinal subsample, the mean item difficulty parameters for the six common items were -0.384 in Grade 5 and -1.110 in Grade 7 (see Table 2).

Table 2: Original and linked item difficulty parameters for the mathematics test in Grade 7.

			Item difficulties σ_j	
Item	**Common item**	**Position**	**Original**	**Linked**
mag9q071_c	No	1	-0.364	0.362
mag7v071_c	No	2	0.499	1.225
mag7r081_c	No	3	0.207	0.932
mag7q051_c	No	4	0.287	1.013
mag5q301_sc3g7_c	Yes	5	-0.267	0.459
mag9d151_c	No	6	-1.356	-0.630
mag5d051_sc3g7_c	Yes	7	-3.130	-2.404
mag5d052_sc3g7_c	No	8	-1.828	-1.103
mag9v011_c	No	9	-0.523	0.203
mag9v012_c	No	10	0.243	0.969
mag7q041_c	No	11	-0.648	0.078
mag7d042_c	No	12	-1.828	-1.102
mag7r091_c	No	13	-0.120	0.606
mag9q181_c	No	14	-1.819	-1.093
mag7d011_c	No	15	-1.346	-0.621
mag7v012_c	No	16	-0.150	0.576
mag7v031_c	No	17	-0.395	0.331
mag5r251_sc3g7_c	Yes	18	-0.502	0.224
mag7d061_c	No	19	0.660	1.386
mag5v321_sc3g7_c	Yes	20	0.284	1.010
mag9v091_c	No	21	1.189	1.914

mag5r191_sc3g7_c	Yes	22	-1.217	-0.491
mag7r02s_c	No	23	-1.258	-0.532

Note. Original item difficulty parameters were derived by an independent scaling of the item responses (see Schnittjer et al., 2015). Linked item difficulty parameters were derived by adding c to the original item parameters.

Using Equation (1) this resulted in a correction term of c = -0.384 - (-1.110) = 0.726. The correction term c was added to each item difficulty parameter derived in Grade 7 and, thus, resulted in the linked item parameters (see Table 2). The corresponding link error according to Equation (2) was 0.09 (for more detailed information see Fischer et al., 2016.).

Person abilities were subsequently estimated using the linked item difficulty parameters in Grade 7. In the SUF, manifest scale scores are provided in the form of two different weighted maximum likelihood estimates (WLE; see Pohl & Carstensen, 2012), “mag7_sc1” and “mag7_sc1u”, including their respective standard error, “mag7_sc2” and “mag7_sc2u”. Both WLE scores are linked to the underlying reference scale of Grade 5. The uncorrected score “mag7_sc1u” (uncorrected for the position of the math test within the booklet) can be used, if the research focus lies on longitudinal issues, such as competence development, since the positon of the domains stays the same over subsequent assessment and therefore, resulting differences in WLE scores can be interpreted as development trajectories across measurement points. Conversely, the corrected score “mag7_sc1” is corrected for the position of the math test within the booklet and can thus not be used for longitudinal purposes but for cross-sectional research questions.

3.2 Linking the test of reading competence

The scaling results of the reading competence tests in Grade 5 are presented in Pohl, Haberkorn, Hardt, and Wiegand (2012), whereas the respective results for Grade 7 can be found in Krannich et al. (in prep.). Moreover, assumptions for the linking of reading competences are discussed in Pohl et al. (2015).

3.2.1 Sample

Overall, 5,193 participants were administered the reading test in Grade 5, 6,186 participants took the test in Grade 7, and 3,829 respondents participated at both measurement occasions. Participants with less than three valid responses were excluded from the link procedure.

3.2.2 Test instruments

The reading test in Grade 5 included 32 items whereas the test in Grade 7 consisted of 40 items. Because retest effects are expected for the reading items, no common items could be administered in the two tests. Instead, an overlap of information was accomplished by using an independent link sample including 608 participants attending Grade 7. While participants in the main study took the two reading competence tests with a time-lag of two years between Grade 5 and Grade 7, participants of the link study took both tests at one measurement point in Grade 7. In the link sample the two tests were presented in random order: 309 participants received the test for Grade 5 first and subsequently the Grade 7 test; 299 participants took the Grade 7 test before working on the Grade 5 test. Moreover, because in Grade 7 two different test versions (i.e., an easy and a difficult test) were administered to participants (see Krannich et al., in prep.), the participants in the link sample

were randomly assigned either test version. In the link sample the test was scaled concurrently, whereas in the main sample the tests in Grades 5 and 7 were scaled independently.

3.2.3 Results

3.2.3.1 Unidimensionality

In the link sample we estimated a one-dimensional model that specified a single latent factor for all items and also a two-dimensional model that specified separate latent factors for the two tests. The information criteria slightly favored the two-dimensional model, AIC = 32125.42 and BIC = 32606.13, over the one-dimensional model, AIC = 32158.13 and BIC = 32630.02. Therefore, we also examined the residual correlations of the one-dimensional model. The corrected Q_3 statistics indicated largely unidimensional scales, $M(Q_3) = 0$, $SD(Q_3) = 0.06$. This indicates that unidimensional scales can be assumed for the reading tests in Grades 5 and 7.

3.2.3.2 Measurement invariance

Measurement invariance was examined using the same procedure as used for the mathematics test. We tested whether the item parameters derived in the link sample showed a non-negligible shift in item difficulties as compared to the longitudinal subsample from the main sample. The respective results are summarized in Tables 4 and 5 in the appendix. The analyses supported measurement invariance for all items.

Furthermore, measurement invariance for Grade 7 was supported for the two groups of one-time participants (n = 2,357) and the longitudinal subsample (n = 3829), with none of the *F*-statistics exceeding the critical value of $F_{0154}(1, 6{,}184) = 131.9$.

3.2.3.3 Linking the two tests

The reading competence tests administered in the two grades were linked using the "mean/mean" method for the anchor-group design (see section 2.2.2). The mean item difficulty parameters in the longitudinal subsample were -1.416 in Grade 5 and -1.293 in Grade 7; in the link sample, the respective mean parameters were -2.321 and -1.531 for Grades 5 and 7. Following equation (3) the correction term was calculated as c = -1.416 – (-1.293) + (-1.531) – (-2.321) = 0.667. The correction term c was added to each difficulty parameter derived in Grade 7 and, thus, resulted in the linked item parameters (see Table 3). The resulting link error according to Equation (4) was 0.07 (for more detailed information see Fischer et al., 2016.).

Table 3: Item difficulty parameters of the linked reading competence test in Grade 7

		Item difficulties σ_i	
Item	**Position**	**Original**	**Linked**
reg70110_c	1	-0,375	0,292
reg70120_c	2	-2,524	-1,856
reg7013s_c	3	-2,594	-1,927
reg70140_c	4	-3,456	-2,789
reg7015s_c	5	-2,940	-2,273
reg7016s_c	6	-1,099	-0,432
reg70210_c	7	-2,792	-2,125
reg70220_c	8	-1,941	-1,274
reg7023s_c	9	-1,932	-1,265
reg7024s_c	10	-0,754	-0,087
reg70250_c	11	-1,003	-0,336
reg7026s_c	12	-1,419	-0,752
reg70310_c	13	-2,629	-1,961
reg70320_c	14	-1,627	-0,960
reg7033s_c	15	-1,215	-0,548
reg70340_c	16	-1,533	-0,866
reg70350_c	17	-2,040	-1,373
reg70360_c	18	-1,252	-0,585
reg70410_c	19	-2,350	-1,683
reg70420_c	20	-1,892	-1,225
reg70430_c	21	-2,403	-1,736
reg70440_c	22	-1,917	-1,249
reg7045s_c	23	-0,469	0,198
reg70460_c	24	0,801	1,468
reg7051s_c	25	-1,963	-1,295
reg70520_c	26	-1,292	-0,625
reg7053s_c	27	-1,164	-0,496
reg7055s_c	28	0,124	0,791
reg70560_c	29	0,522	1,190
reg70610_c	30	-2,847	-2,180

		Item difficulties σ_i	
Item	**Position**	**Original**	**Linked**
reg70620_c	31	-0,613	0,055
reg7063s_c	32	-2,706	-2,039
reg70640_c	33	0,464	1,131
reg70650_c	34	0,229	0,897
reg7066s_c	35	-1,208	-0,541
reg7071s_c	36	-1,482	-0,815
reg70720_c	37	0,918	1,585
reg70730_c	38	0,631	1,299
reg70740_c	39	-0,911	-0,244
reg7075s_c	40	0,318	0,985

Note. Original item difficulty parameters were derived by an independent scaling of the item responses (see Krannich et al., in prep.). Linked item difficulty parameters were derived by adding c to the original item parameters.

Person abilities were subsequently estimated using the linked item difficulty parameters in Grade 7. In the SUF, manifest scale scores are provided in the form of two different WLE estimates, ”reg7_sc1” and “reg7_sc1u”, including their respective standard errors “reg7_sc2” and “reg7_sc2u”. Both WLE scores are linked to the underlying reference scale of Grade 5. The uncorrected score “reg7_sc1u” (uncorrected for the position of the reading test within the booklet) can be used, if the focus of the research lies on longitudinal issues, such as competence development since differences in WLE scores can be interpreted as development trajectories across measurement points. Again, the corrected score “reg7_sc1” was corrected for the position of the reading test within the booklet and can be used, if the research interest lies on cross-sectional issues.

4. Summary

The NEPS repeatedly measures different competences (e.g., reading, mathematics) across the life span. To study competence development and compare competences scores from different measurement occasions, the different scores must be placed onto a common scale. Otherwise changes in competences would be confounded with differences in test difficulties. Therefore, in the NEPS repeatedly measured competences are linked and placed onto a common scale. Depending on the specific competence domain the NEPS uses two different link strategies. Competences such as mathematical competence that are unlikely to be prone to memory effects are linked using an anchor-items design. In contrast, for competences that might be susceptible to memory effects (e.g., reading competence) an anchor-group design is used. After outlining the basics of these link procedures, the present study demonstrated how to link mathematical and reading competences across Grades 5 and 7. It was shown that the two mathematical tests administered in the two grades were essentially unidimensional. Moreover, six items that were included in both tests were measurement invariant and, thus, could be used to link the two tests. Therefore, the item parameters of the mathematical test administered in Grade 7 were linked to the item parameters of the

respective test administered in Grade 5. As a consequence, person abilities estimated using these linked item difficulty parameters were on the same scale as the person abilities derived in Grade 5. These ability estimates can be used for longitudinal comparisons; ability differences using these scores can be interpreted as development trajectories across the two measurement occasions. Subsequently, we also demonstrated that the reading competence tests administered in Grades 5 and 7 could be linked using an anchor-groups design. It was shown that both tests were essentially unidimensional and showed no differential item functioning. Therefore, the item parameters of the reading test administered in Grade 7 were linked to the item parameters of the respective test administered in Grade 5 using the responses of an independent link sample. As a consequence, person abilities estimated using these linked item difficulty parameters were on the same scale as the person abilities derived in Grade 5. These scores can be used to compare ability estimates across the two grades. In conclusion, the paper summarized the two link strategies adopted in the NEPS and showed how to derive ability estimates that can be compared across different measurement occasions.

References

Blossfeld, H.-P., Roßbach, H.-G., & von Maurice, J. (2011). Education as a lifelong process: The German National Educational Panel Study (NEPS). Zeitschrift für Erziehungswissenschaft, 14, 1-4.

Duchhardt, C., & Gerdes, A. (2012). NEPS Technical Report for Mathematics - Scaling Results of Starting Cohort 3 in Fifth Grade (NEPS Working Paper No. 19). Bamberg: Otto-Friedrich-Universität, Nationales Bildungspanel.

Fischer, L., Rohm, T., & Carstensen, C. H. (2015a, July). Comparing link methods based on their link errors. Poster presented at the International Meeting of the Psychometric Society, Beijing, China.

Fischer, L., Rohm, T., & Carstensen, C. H. (2015b, September). Ein Vergleich verschiedener Linkmethoden und ihrer Linkfehler anhand Rasch skalierter Kompetenzdaten. Paper presented at the 12th meeting of the section Methoden & Evaluation of the DGPs, Jena.

Fischer, L., Gnambs, T., Rohm, T., & Carstensen, C. H. (2016) Evaluating link methods on Rasch scaled longitudinal data in large scale assessments. Unpublished manuscript, Bamberg: Leibniz Institute for Educational Trajectories.

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

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

Krannich, M., Jost, O., Rohm, T., Koller, I., Carstensen, C. H., & Fischer, L. (in prep.). NEPS Technical Report for Reading – Scaling results of Starting Cohort 3 in seventh grade (NEPS Working Paper). Bamberg: Leibniz Institute for Educational Trajectories.

Lord F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum.

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. Journal of Educational Measurement, 17, 179-193.

Marco, G. L. (1977). Item characteristic curve solutions to the three intractable testing problems. Journal of Educational Measurement, 16, 139-160.

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

Murphy, K. R., & Myors, B. (1999). Testing the hypothesis that treatments have negligible effects: Minimum-effect tests in the general linear model. Journal of Applied Psychology, 84, 234-248.

Organisation for Economic Cooperation and Development (OECD). (2014). PISA 2012 Technical Report. Paris, France: OECD Publishing. Retrieved from https://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf

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.

Pohl, S., Haberkorn, C., & Carstensen, C. H. (2015). Measuring competencies across the lifespan - Challenges of linking test scores. In M. Stemmler, A. von Eye, & W. Wiedermann (Eds), Dependent data in social science research (pp. 281-308). Berlin, Germany: Springer.

Pohl, S., Haberkorn, K., Hardt, K., & Wiegand, E. (2012). NEPS Technical Report for Reading – Scaling results of Starting Cohort 3 in fifth grade (NEPS Working Paper No. 15). Bamberg: Otto-Friedrich-Universität, Nationales Bildungspanel.

Schnittjer, I., Gerken, A., & Fischer., L. (2015). NEPS Technical Report for Mathematics – Scaling results of Starting Cohort 3 in seventh grade (NEPS Working Paper No. XX).

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of thethree-parameter logistic model. Applied Psychological Measurement, 8, 125-145.

Appendix

Table 4: DIF Analyses for Reading Competence between the longitudinal subsample in Grade 5 and the Link Sample (LS).

Item	Δσ	$SE_{\Delta\sigma}$	*F*
reg50110_c	0.003	0.302	0.00
reg5012s_c	-0.372	0.274	1.84
reg50130_c	0.006	0.189	0.00
reg50140_c	-0.047	0.156	0.09
reg50150_c	-0.037	0.126	0.09
reg5016s_c	-0.112	0.143	0.61
reg50170_c	0.426	0.114	13.89
reg50210_c	-0.108	0.237	0.21
reg50220_c	-0.385	0.115	11.15
reg50230_c	0.219	0.239	0.84
reg50240_c	0.008	0.153	0.00
reg50250_c	-0.312	0.126	6.14
reg5026s_c	0.145	0.115	1.61
reg50310_c	0.166	0.211	0.62
reg50320_c	0.249	0.275	0.82
reg50330_c	-0.148	0.215	0.47
reg50340_c	0.315	0.175	3.25
reg50350_c	-0.251	0.124	4.08
reg50360_c	0.673	0.265	6.44
reg50370_c	-0.238	0.142	2.81
reg50410_c	0.267	0.135	3.90
reg5042s_c	-0.292	0.196	2.21
reg50430_c	0.218	0.119	3.39
reg50440_c	0.063	0.120	0.27
reg50460_c	0.102	0.130	0.61
reg50510_c	0.165	0.239	0.48
reg5052s_c	-0.106	0.184	0.33
reg50530_c	-0.461	0.132	12.27
reg50540_c	0.086	0.182	0.22
reg5055s_c	0.343	0.211	2.64

Item	Δσ	$SE_{\Delta\sigma}$	*F*
reg50560_c	-0.256	0.142	3.23
reg50570_c	-0.347	0.153	5.14

Note. Δσ = Difference in item difficulty parameters between the longitudinal subsample in Grade 5 and the link sample (positive values indicate easier items in the link sample); $SE_{\Delta\sigma}$ = Pooled standard error; *F* = Test statistic for the minimum effects hypothesis test based on(5). the critical value for the minimum effects hypothesis test using an α of .05 is F_{0154} (1, 4,435) = 99.7. A non-significant test indicates measurement invariance.

Table 5: DIF Analyses for Reading Competence between the longitudinal subsample in Grade 7 and the Link Sample (LS).

Item	Δσ	$SE_{\Delta\sigma}$	*F*
reg70110_c	-0.742	0.152	23.87
reg70120_c	-0.690	0.199	12.02
reg7013s_c	0.228	0.292	0.61
reg70140_c	0.382	0.382	1.00
reg7015s_c	-0.331	0.246	1.81
reg7016s_c	0.120	0.188	0.41
reg70210_c	-0.123	0.168	0.54
reg70220_c	0.031	0.144	0.05
reg7023s_c	0.128	0.170	0.57
reg7024s_c	0.348	0.144	5.80
reg70250_c	-0.106	0.122	0.76
reg7026s_c	-0.025	0.167	0.02
reg70310_c	0.416	0.199	4.38
reg70320_c	0.053	0.137	0.15
reg7033s_c	-0.088	0.138	0.41
reg70340_c	-0.006	0.135	0.00
reg70350_c	-0.078	0.150	0.27
reg70360_c	-0.186	0.128	2.09
reg70410_c	0.091	0.169	0.29
reg70420_c	-0.063	0.149	0.18
reg70430_c	0.114	0.178	0.41
reg70440_c	0.021	0.157	0.02
reg7045s_c	0.321	0.134	5.79
reg70460_c	-0.060	0.131	0.21
reg7051s_c	0.832	0.313	7.07
reg70520_c	-0.046	0.199	0.05
reg7053s_c	0.019	0.220	0.01
reg7055s_c	0.350	0.173	4.08
reg70560_c	-0.701	0.186	14.25
reg70610_c	-0.762	0.228	11.15
reg70620_c	0.294	0.159	3.42
reg7063s_c	-0.440	0.240	3.36

Item	Δσ	$SE_{\Delta\sigma}$	*F*
reg70640_c	0.145	0.153	0.89
reg70650_c	-0.042	0.153	0.08
reg7066s_c	-0.265	0.153	3.02
reg7071s_c	0.540	0.256	4.43
reg70720_c	0.228	0.196	1.36
reg70730_c	-0.019	0.198	0.01
reg70740_c	-0.351	0.202	3.01
reg7075s_c	0.463	0.197	5.50

Note. Δσ = Difference in item difficulty parameters between the longitudinal subsample in Grade 7 and the link sample (positive values indicate easier items in the link sample); SEΔσ = Pooled standard error; F = Test statistic for the minimum effects hypothesis test based on(5). The critical value for the minimum effects hypothesis test using an α of .05 is F0154 (1, 4,435) = 99.7. A non-significant test indicates measurement invariance.