Vergleichbarkeit der naturwissenschaftlichen Kompetenz in der neunten Klasse im Nationalen Bildungspanel und im IQB-Ländervergleich 2012

Helene Wagner · Inga Hahn · Katrin Schöps · Nicole Mahler · Olaf Köller

Online publiziert: 9. September 2019
© Der/die Autor(en) 2019

Zusammenfassung Verschiedene Schulleistungsstudien in Deutschland messen die Kompetenz von Schülerinnen und Schülern in den Naturwissenschaften am Ende der Sekundarstufe I, wie z. B. der *IQB-Ländervergleich* (LV; neuerdings *IQB-Bildungstrend*) und das *Nationale Bildungspanel* (NEPS). Die vorliegende Untersuchung beschäftigt sich mit der Überprüfung der dimensional Äquivalenz und der Skalenäquivalenz der Testwerte von NEPS und dem LV mit dem Ziel, die Skalen beider Studien miteinander zu verlinken. Zu diesem Zweck wurde eine Linking-Studie durchgeführt, in der 678 Schülerinnen und Schüler in einem Single-Group-Design die Aufgaben aus beiden Studien bearbeitet haben. Die Analysen zeigten eine hohe Vergleichbarkeit der NEPS- und LV-Testinstrumente hinsichtlich der faktoriellen Struktur der Daten sowie der Verteilung von Kompetenzwerten. Das Linking erfolgte mittels Equipercents Equating und zeigte eine hohe Übereinstimmung der Testwertverteilung beider Studien sowie der Verteilung der Personen auf die Kompetenzstufen des LV.

Dr. H. Wagner (✉) · Dr. I. Hahn · Dr. K. Schöps · Prof. Dr. O. Köller
Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik,
Olshausenstr. 62, 24118 Kiel, Deutschland
E-Mail: h.wagner@ipn.uni-kiel.de

Dr. I. Hahn
E-Mail: hahn@ipn.uni-kiel.de

Dr. K. Schöps
E-Mail: schoeps@ipn.uni-kiel.de

Prof. Dr. O. Köller
E-Mail: koeller@ipn.uni-kiel.de

Dr. N. Mahler
Institut zur Qualitätsentwicklung im Bildungswesen, Unter den Linden 6, 10099 Berlin, Deutschland
E-Mail: nicole.mahler@iqb.hu-berlin.de

Schlüsselwörter NEPS · Ländervergleich · Schulleistungsstudie · Naturwissenschaftliche Kompetenz · Linking

Comparability of the Science Competency of 9th Graders in the German National Educational Panel Study and in the German National Assessment Study

Abstract Several large-scale assessments in Germany, e.g. the German National Assessment Study and the German National Educational Panel Study, measure the scientific competency of students at the end of secondary school. The goal of the present study was to assess the dimensional equivalence and the scale equivalence of their test scores with the aim of linking the scales of both studies. For this purpose, a linking study was conducted with $N=678$ students who worked on science items from both studies in a single group design. The analyses showed a high similarity of the NEPS and LV test instruments with regard to the factorial structure of the data and the distribution of the test scores. The linking was performed using equipercentile equating and showed a high consistency between the two studies in the distribution of the test scores and also in the assignment of students to the proficiency levels of the German National Assessment Study.

Keywords German National Assessment Study · Large-scale assessments · Linking · NEPS · Science competency

1 Einleitung

Studien wie die *Third International Mathematics and Science Study* (TIMSS), das *Programme for International Student Assessment* (PISA) und die *Überprüfung des Erreichens der Bildungsstandards in den Fächern Biologie, Chemie und Physik für den Mittleren Schulabschluss* untersuchen regelmäßig den Stand der naturwissenschaftlichen Kompetenz der Schülerinnen und Schüler (SuS) in Deutschland. Neben einem Vergleich der Kompetenzen von SuS der verschiedenen (Bundes-)Länder stehen dabei auch die unterschiedlichen Bildungssysteme auf dem Prüfstand.

Wie die Studie von Hanushek und Wößmann (2015) zeigte, steht der in PISA gemessene Bildungserfolg der SuS am Ende der Sekundarstufe I im Zusammenhang mit dem wirtschaftlichen Erfolg des jeweiligen Landes. Die dort dargestellten Regressionsanalysen ergaben, dass die naturwissenschaftliche Kompetenz, verglichen mit der mathematischen Kompetenz und der Lesekompetenz, das Wirtschaftswachstum der OECD-Staaten am besten vorhersagt. Auch wenn die von Hanushek und Wößmann (2015) verwendeten Analysen streng genommen die kausale Richtung des gefundenen Zusammenhangs offenlassen, interpretieren die Autoren die Ergebnisse dahin gehend, dass ein Anstieg in der naturwissenschaftlichen Kompetenz das Wirtschaftswachstum der OECD-Staaten positiv beeinflusst. Unter Berücksichtigung dieses Befundes gewinnt die Frage danach, wie sich die naturwissenschaftliche Kompetenz vom Kindes- bis ins hohe Erwachsenenalter entwickelt, zusätzlich an

Bedeutung. Unter anderem dieser Frage geht das *Nationale Bildungspanel* (National Educational Panel Study-NEPS) seit 2009 nach.

Um die Anschlussfähigkeit des NEPS an nationale und internationale Large-Scale-Assessments in Deutschland zu ermöglichen, wird eine studienübergreifende Interpretation der Ergebnisse angestrebt (Blossfeld 2008). Diese kann zum einen zur Klarheit und Transparenz der abgeleiteten Schlussfolgerungen aus verschiedenen Studien beitragen. Zum anderen ermöglicht sie eine Erweiterung der jeweiligen Testwertinterpretationen. Gelingt beispielsweise die Übertragung der Kompetenzskalen des Ländervergleichs (LV) in Biologie, Chemie und Physik auf die naturwissenschaftliche Skala des NEPS, so könnte mithilfe der längsschnittlichen Analysen im NEPS untersucht werden, welche Faktoren sich besonders gut eignen, um die Leistung im LV vorherzusagen. Gleichzeitig können die LV-Kompetenzskalen, wenn sie mit der Leistung der NEPS-Teilnehmerinnen und -Teilnehmer am Ende der Sekundarstufe I verlinkt werden, selbst als Prädiktor für die Vorhersage des Erfolgs in den Naturwissenschaften am Ende der Sekundarstufe II verwendet werden. Schließlich könnte so untersucht werden, ob naturwissenschaftliche Kompetenzen, wie sie im LV erhoben werden, später jenseits der Effekte der Abiturdurchschnittsnote und anderer kognitiver Maße den Erfolg im Studium eines naturwissenschaftlichen Faches vorhersagen.

Die Anschlussfähigkeit des NEPS-Naturwissenschaftstests für die neunte Klassenstufe wurde bereits am PISA-Naturwissenschaftstest untersucht (Wagner et al. 2014, 2018). Die Ergebnisse dieses Vergleichs zeigten eine hohe Überschneidung der Rahmenkonzeptionen der Tests sowie eine hohe Vergleichbarkeit ihrer Kompetenzwerte und ihrer Kompetenzskalen. Da die Rahmenkonzeptionen der naturwissenschaftlichen Kompetenz beider Studien allerdings auf einem gemeinsamen Konzept, nämlich dem der Scientific Literacy (Bybee 1997; OECD 2006), beruhen, war eine hohe Vergleichbarkeit der Testinstrumente zu erwarten. Die Rahmenkonzeptionen von NEPS und dem LV wurden bisher lediglich auf Ebene der einzelnen Konzepte untersucht. Hier zeigte sich trotz der Unterschiede in der fachlichen Spezifität ihrer Inhalte (naturwissenschaftliche Grundbildung vs. Kompetenz in Biologie, Chemie und Physik) eine hohe inhaltliche Überschneidung der Tests (Wagner et al. 2014).

Um die weiter oben erwähnte Erweiterung der Testwertinterpretationen zu erreichen, müssen die NEPS- und LV-Rahmenkonzeptionen miteinander verlinkt werden. Die Voraussetzung dafür ist neben dem bereits erfolgten Vergleich der Rahmenkonzeptionen die Untersuchung der Vergleichbarkeit ihrer Kompetenzwerte und ihrer Kompetenzskalen (van de Vijver 1998; Pietsch et al. 2009).

In der vorliegenden Arbeit sollen nun zunächst die dimensionale Struktur sowie die Verteilung der Kompetenzwerte des NEPS-Naturwissenschaftstests 2009 mit der dimensional Struktur sowie der Verteilung der Kompetenzwerte der LV-Tests 2012 zur Überprüfung des Erreichens der Bildungsstandards in den Fächern Biologie, Chemie und Physik für den Mittleren Schulabschluss verglichen werden. Sollten die Testinstrumente der beiden Studien hinsichtlich der genannten Merkmale miteinander vergleichbar sein, werden im nächsten Schritt ihre Kompetenzwerte miteinander verlinkt.

2 Theoretischer Hintergrund

Kolen und Brennan (2004) schlagen zum Vergleich von Testinstrumenten und ihren empirischen Befunden folgende Kriterien vor:

- *Schlussfolgerungen*: Inwiefern lassen sich aus den Testwerten der zu vergleichenden Tests ähnliche Schlussfolgerungen ableiten?
- *Zielpopulationen*: Inwieweit werden die Testinstrumente bei derselben Zielpopulation eingesetzt?
- *Merkmale und Umstände der Messung*: Inwieweit ähneln sich die Tests hinsichtlich der Messbedingungen, insbesondere in Bezug auf die verwendeten Aufgabenformate, Durchführungsbedingungen und die Testlänge?
- *Operationalisierte Konstrukte*: Inwieweit erfassen die Tests dieselben inhaltlichen Teilbereiche und kognitiven Prozesse?

Nur, wenn die Tests hinsichtlich der genannten Kriterien ähnlich sind, ist es sinnvoll, ihre Testwerte miteinander zu verlinken. Weiterhin hat die Vergleichbarkeit der Testinstrumente eine Auswirkung auf die Wahl der Linking-Methode und später auf die Robustheit des Linkings.

2.1 Vergleich der Rahmenkonzeptionen der Naturwissenschaftstests von NEPS und dem LV

Um abzuschätzen, inwiefern die Testinstrumente der Studien NEPS und LV miteinander vergleichbar sind, werden ihre Rahmenkonzeptionen einander auf Basis der Kriterien von Kolen und Brennan (2004) gegenübergestellt.

2.1.1 Schlussfolgerungen

Das Ziel des LV liegt in der länderübergreifenden Überprüfung des Erreichens der 2004 beschlossenen Bildungsstandards in den Fächern Biologie, Chemie und Physik für den Mittleren Schulabschluss (KMK 2006). Der Vergleich der Kompetenzstände im LV wird in einem querschnittlichen Design realisiert, d.h. es werden jeweils im Abstand von sechs Jahren die Leistungen von SuS in den Fächern Biologie, Chemie und Physik am Ende der Sekundarstufe I (neunte Jahrgangsstufe) untersucht. Auf diese Weise ist es zwar möglich, einen Trend in den erreichten Kompetenzen abzubilden, allerdings kann mit Hilfe querschnittlicher Analysen keine Aussage über die Entwicklung der Kompetenz der SuS gemacht werden.

Diese Forschungslücke versucht seit 2009 die NEPS-Studie (Blossfeld et al. 2011) zu schließen, die längsschnittliche Analysen der Kompetenzentwicklung der Menschen in Deutschland von ihrer Geburt bis ins hohe Erwachsenenalter auf Basis eines Multi-Kohorten-Sequenz-Designs liefert. Dieses Design erlaubt auch die Feststellung des Kompetenzstandes der SuS unterschiedlicher Startkohorten in der neunten Jahrgangsstufe. Somit können aus den Testwerten des NEPS und des LV trotz unterschiedlicher Zielsetzungen ähnliche Schlussfolgerungen abgeleitet werden, nämlich, über welche naturwissenschaftlichen Fähigkeiten und Fertigkeiten die SuS am Ende der Sekundarstufe I verfügen.

2.1.2 Zielpopulationen

Die in diesem Artikel zu vergleichenden Tests der Studien NEPS (Maurice et al. 2011) und LV (Siegle et al. 2013) haben dieselbe Zielpopulation, nämlich SuS der neunten Jahrgangsstufe.

2.1.3 Merkmale und Umstände der Messung

Die Datenerhebung und -verarbeitung wurden für NEPS 2010 und LV 2012 vom IEA Data Processing Center (DPC) in Hamburg koordiniert und in Form von Papier- und Bleistifttests standardisiert durchgeführt. Ein wichtiger Unterschied zwischen den Tests des NEPS und des LV besteht in den Antwortformaten. Im NEPS-Naturwissenschaftstest für die neunte Jahrgangsstufe werden ausschließlich geschlossene Antwortformate in Gestalt einfacher *multiple-choice*- oder *multiple-true-false*-Aufgaben (Schöps und Saß 2013) verwendet. In den Tests des LV werden zusätzlich zu den geschlossenen Antwortformaten (59 % des Gesamttests) halboffene (19 %) und offene Formate (22 %) eingesetzt, die eine frei formulierte Antwort erfordern (Kauertz und Fischer 2013).

Die Naturwissenschaftstests des NEPS und des LV unterscheiden sich außerdem stark in der Anzahl ihrer Items: die Bildungsstandards-Tests enthalten 386 Items, die in mehrere Aufgabenblöcke aufgeteilt sind. Jedes Testheft enthält sechs Aufgabenblöcke, die zur Ausbalancierung von Reihenfolge- und Ermüdungseffekten in ihrer Blockposition variiert werden. Die Bearbeitungszeit für jedes Testheft liegt pro Person bei zwei Stunden (Siegle et al. 2013). Der NEPS-Test besteht dagegen aus nur 28 Items, die in der gleichen Reihenfolge in einer Testzeit von 28 min bearbeitet werden (Schöps und Saß 2013). Die Unterschiede in Testzeit und Testlänge können dazu führen, dass die Tests des NEPS und des LV ihre Konstrukte in unterschiedlicher Breite messen (vgl. hierzu Wagner et al. 2014).

Weitere Unterschiede zwischen NEPS und dem LV bestehen bei den statistischen Analysen im Umgang mit fehlenden Werten. Im NEPS werden die Kategorien *nicht erreicht*, *unplausibler Wert* und *Angabe verweigert* sowohl bei der Bestimmung von Itemschwierigkeiten als auch bei der Schätzung von Personenparametern als fehlend betrachtet (Pohl und Carstensen 2012). Im LV werden dagegen alle fehlenden Werte bei der Schätzung der Item- und Personenparameter als falsch kodiert.

Ein weiterer Unterschied betrifft die Dimensionalität. Da die Tests des LV die Kompetenzen fächerspezifisch (Biologie, Chemie oder Physik) und inhaltspezifisch (Umgang mit Fachwissen oder Erkenntnisgewinnung) erheben, werden die Personenparameter hier in einem mehrdimensionalen Rasch-Modell geschätzt. NEPS versteht die naturwissenschaftliche Kompetenz dagegen als ein dimensionales Konstrukt. Dementsprechend wird die Fähigkeit der Personen in einem eindimensionalen Schätzverfahren ermittelt.

Abb. 1 Übersicht über die Rahmenkonzeptionen des NEPS-Naturwissenschaftstests und der LV-Tests 2012 zur Überprüfung des Erreichens der Bildungsstandards in Biologie, Chemie und Physik. (In Anlehnung an Hahn et al. 2013; Kauertz et al. 2010)

2.1.4 Operationalisierte Konstrukte: Vergleich der naturwissenschaftlichen Kompetenz im NEPS mit den fächerspezifischen Kompetenzen im LV 2012

Die Rahmenkonzeption der LV-Tests 2012 in Biologie, Chemie und Physik (Abb. 1, links) basiert auf einem fachdidaktisch ausgerichteten, mehrdimensionalen Kompetenzmodell (Kauertz et al. 2010; Wellnitz et al. 2012) und umfasst die Dimensionen *Kompetenzbereiche*, *Kognitive Prozesse* und *Komplexität*. Mit den Tests des LV 2012 sollte überprüft werden, ob die dort formulierten (schulischen) Lernziele am Ende der Sekundarstufe I erreicht werden. Aus diesem Grund werden die Kompetenzen im LV unter Berücksichtigung des Faches (*Physik*, *Chemie* oder *Biologie*) und des jeweiligen Kompetenzbereiches (*Umgang mit Fachwissen* oder *Erkenntnisgewinnung*) erhoben.

Die Entwicklung der LV-Aufgaben erfolgte im Rahmen des Projekts Evaluation der Standards in den Naturwissenschaften für die Sekundarstufe I (Walpuski et al. 2008). Zur Schätzung der Schülerkompetenzen wurde im LV ein sechsdimensionales Modell (ein Wert je Kompetenzbereich und Fach) verwendet, dessen empirische Validierung aussteht. Bisher wurden lediglich die Unterscheidung des Umgangs mit Fachwissen und Erkenntnisgewinnung in den Biologie-Aufgaben (Kampa und Köller 2016) sowie ein fach- und prozessbezogenes Modell im Bereich der Erkenntnisgewinnung (Wellnitz et al. 2017) empirisch überprüft.

Die Rahmenkonzeption des NEPS-Naturwissenschaftstests (Abb. 1, rechts) orientiert sich bei der Modellierung der Kompetenz am Literacy-Konzept (Bybee 1997; OECD 2006) im Sinne einer funktionalen Grundbildung, die in den Alltagskontext

eingebunden ist. Dementsprechend verzichtet NEPS bei der Definition der Inhalte des Naturwissenschaftstests auf eine enge Orientierung an Lehrplänen. Stattdessen wird die naturwissenschaftliche Kompetenz in der NEPS-Rahmenkonzeption als ein eindimensionales Konstrukt mit den beiden Facetten der *inhaltsbezogenen* und der *prozessbezogenen Komponente* definiert. Darüber hinaus erfolgt die Erfassung der naturwissenschaftlichen Kompetenz in der NEPS-Rahmenkonzeption, anders als im LV, eingebettet in ausgewählte Kontexte. Folglich unterscheiden sich die Rahmenkonzeptionen der Studien in ihren Konstruktdefinitionen. Aber wie vergleichbar sind die Tests beider Studien auf der Aufgabenebene?

Die Ähnlichkeit der operationalisierten Konstrukte kann in Anlehnung an van de Vijver (1998) hinsichtlich der konzeptionellen Äquivalenz, der dimensionalen Äquivalenz und der Skalenäquivalenz beurteilt werden (vgl. Pietsch et al. 2009). Untersuchungen zur konzeptionellen Äquivalenz beider Tests konnten zeigen, dass der NEPS-Test auf der Aufgabenebene große Gemeinsamkeiten mit der Rahmenkonzeption der Bildungsstandards aufweist (Wagner et al. 2014).

Wie die Studie von Wagner et al. (2014) gezeigt hat, liegt die größte Überschneidung der Testinstrumente in ihren Kompetenzbereichen (Abb. 1, gestrichelte Linie zwischen den Rahmenkonzeptionen der Studien). So konnten jedem Teilbereich des Umgangs mit Fachwissen bzw. der Erkenntnisgewinnung der LV-Rahmenkonzeption ein oder mehrere NEPS-Items zugeordnet werden. Diese Überschneidung liefert die erste Basis für die Verlinkung der LV- und NEPS-Kompetenzskalen.

2.2 Linking-Methoden und -Studien

Je nachdem, wie stark sich Tests hinsichtlich der von Kolen und Brennan (2004) vorgeschlagenen Merkmale ähneln, können verschiedene Linking-Methoden angewendet werden. Mislevy (1992) und Linn (1993) unterscheiden fünf Arten des Linkings: *Moderation*, *Projection*, *Concordance*, *Vertical Scaling* und *Equating*.

Das Equating ist die stärkste Linking-Methode. Die Verknüpfung der Skalen zweier Tests (A und B) mit der Equating-Methode ermöglicht die Vorhersage der Werte des Tests A aus den Werten des Tests B und umgekehrt. Für die Verknüpfung mit Hilfe des Equating müssen die Tests in Bezug auf alle Merkmale des Ansatzes von Kolen und Brennan (2004) ähnlich sein. Nissen et al. (2015) haben in ihrer Studie das Linking auf Grundlage der Item-Response-Theorie (IRT) und das Equipercentile Equating hinsichtlich der Qualität des Linkings miteinander verglichen, indem sie die Mathematik-Werte des TIMSS 2011-Tests für die vierte Klasse mit den entsprechenden Werten des NEPS-Tests 2010 für die fünfte Klasse verknüpft haben. Dabei stellten sie fest, dass die Verteilungen der mit dem Equipercentile Equating verlinkten Werte im Vergleich zum IRT-Linking ähnlicher sind. Somit bietet das Equipercentile Equating einen Vorteil gegenüber dem IRT-Linking.

Seit PISA 2000 gibt es weltweit Bestrebungen, nationale und internationale Assessments zu verlinken, um auf diese Weise ihre Testwertinterpretationen zu erweitern. So wurden in der Studie von Wagner et al. (2018) die Werte des Naturwissenschaftstests von PISA 2012 mit den Werten des NEPS-Naturwissenschaftstests 2010 für die neunte Klasse verknüpft. Die Untersuchung der Vergleichbarkeit der dimensional Struktur von NEPS und PISA ergab einen hohen latenten Zusammenhang

($r=0,85$) zwischen den Naturwissenschaftsskalen der beiden Studien. Die Verlinkung der beiden Skalen mit dem Equipercile Equating zeigte für beide Studien äquivalente Verteilungen der SuS auf die PISA-Kompetenzstufen.

In einer weiteren Studie übertrugen van den Ham et al. (2017) die Ergebnisse des NEPS-Mathematiktests für die neunte Klassenstufe auf die Kompetenzstufen des LV. Die gemeinsame Skalierung der Testwerte zeigte, dass beide Skalen trotz des hohen latenten Zusammenhangs von $r=0,92$ nicht ohne Weiteres austauschbar sind. Das Linking erfolgte via Equipercile Equating und zeigte vergleichbare Verteilungen der SuS auf die Kompetenzstufen des LV.

3 Fragestellungen

Das Ziel dieser Arbeit liegt in der Verlinkung der Kompetenzskalen des NEPS und des LV. Die Voraussetzung für ein robustes Linking ist die Vergleichbarkeit der Testwerte. In diesem Zusammenhang werden folgende Fragestellungen untersucht:

Dimensionale Äquivalenz

- (1) Inwiefern ist die dimensionale Struktur der Naturwissenschaftswerte im NEPS und dem LV miteinander vergleichbar?

Skalenäquivalenz

- (2) Inwiefern zeigen sich bei den Kompetenzskalen des NEPS und des LV ähnliche Verteilungen?

Linking

- (3) Wie hoch ist die Klassifikationskonsistenz der verlinkten Werte hinsichtlich der Zuordnung zu den Kompetenzstufen der LV-Tests?

4 Methode

4.1 Stichprobe und Untersuchungsdesign

Zur Überprüfung der Äquivalenz der naturwissenschaftlichen Messung in NEPS und dem LV wurde im Frühling 2012 eine Linking-Studie durchgeführt. Ein weiteres Ziel dieser Studie lag in der Untersuchung der Effektivität des Programms *Steigerung der Effizienz des mathematisch-naturwissenschaftlichen Unterrichts* (SINUS; Prenzel und Ostermeier 2006). Aus diesem Grund bestand die Stichprobe aus 80 SINUS-Schulen (1728 SuS). Die Datenerhebung erfolgte in fünf Bundesländern (Bayern, Hamburg, Hessen, Schleswig-Holstein und Thüringen) und umfasste fünf verschiedene Schulformen: Gymnasium ($N=32$), integrierte Gesamtschule ($N=19$), Realschule ($N=16$), Schulen mit mehreren Bildungsgängen ($N=12$) und eine Hauptschule.

Der Prozess der Datenerhebung und -verarbeitung wurde wie in den Haupterhebungen der Studien vom DPC durchgeführt. Insgesamt haben 678 SuS (50 % weiblich) die Aufgaben aus NEPS und dem LV bearbeitet. Der Altersdurchschnitt betrug 15,6 Jahre. Die Gesamtbearbeitungszeit der Tests lag bei 1,5 Zeitstunden. Es wurden insgesamt 148 Aufgaben der LV-Tests, die eine zufällige Auswahl der Aufgaben der Hauptstudie darstellten, bearbeitet. Die LV-Aufgaben wurden auf 12 Blöcke verteilt und im Multi-Matrix-Design dargeboten, so dass jede Testperson nur einen Teil der Aufgaben (3 Blöcke à 20 min) bearbeitete. Die Aufgabenblöcke entstammten der Haupterhebung des LV in 2012 und wurden für den Einsatz in der Linking-Studie zufällig ausgewählt. Zur Schätzung der naturwissenschaftlichen Kompetenz im NEPS haben SuS den vollständigen NEPS-Test aus der Haupterhebung 2010 bearbeitet. Die NEPS-Items wurden wie in der Hauptstudie in einer festen Reihenfolge mit einer Testzeit von 28 min dargeboten. Zur Reduktion von Reihenfolge- und Ermüdungseffekten wurden die NEPS- und LV-Testblöcke in ihrer Position rotiert.

4.2 Dimensionale Äquivalenz

Die erste Fragestellung zielte auf die Untersuchung der Vergleichbarkeit der faktoriellen Struktur der Kompetenzwerte im NEPS und dem LV. Wie der Vergleich der operationalisierten Konstrukte im Abschn. 2.1.4 gezeigt hat, unterscheiden sich die Rahmenkonzeptionen von NEPS und dem LV in ihrer Komplexität. Diese Komplexität spiegelt sich in der Anzahl der Kompetenzdimensionen, die in den jeweiligen Studien zur Schätzung der Fähigkeit von SuS herangezogen werden, wider. Während die Kompetenz von SuS im LV anhand eines sechsdimensionalen Modells unter Berücksichtigung des Kompetenzbereiches (Umgang mit Fachwissen oder Erkenntnisgewinnung) und des jeweiligen naturwissenschaftlichen Faches (Biologie, Chemie oder Physik) ermittelt wird (Mayer et al. 2013), wird die Personenfähigkeit im NEPS als ein eindimensionales Konstrukt mit den beiden Facetten der inhaltsbezogenen und prozessbezogenen Komponente geschätzt (Schöps und Saß 2013). Diese Vorgehensweisen bei der Schätzung der Personenfähigkeiten im NEPS und dem LV bildeten den Ausgangspunkt für die Untersuchung der dimensionalen Äquivalenz ihrer Kompetenzdaten. Dementsprechend beinhaltet das erste Modell in unseren Analysen (Abb. 2, Modell 1) die oben beschriebenen sechs fach- und kompetenzbezogenen LV-Dimensionen sowie eine kompetenzübergreifende NEPS-Dimension. Diesem Modell wurde ein dreidimensionales Modell (Modell 2), das die naturwissenschaftlichen Kompetenzen beider Studien als fachübergreifende Dimensionen beschreibt, gegenübergestellt. Dieses Modell überprüft die Annahme, dass sich die drei Fächer im LV nicht mehr als eigene Facetten abbilden lassen, sondern zu zwei Naturwissenschaftsfaktoren (Fachwissen und Erkenntnisgewinnung) zusammengefasst werden können, und stellt somit eine Vereinfachung des ersten Modells dar. Die fehlende analytische Trennbarkeit der fachlichen Facetten wurde auch schon anhand der PISA-Daten gezeigt (Haschke et al. 2017) und mag u. a. daran liegen, dass die naturwissenschaftlichen Denk- und Arbeitsweisen, die sich in der Erkenntnisgewinnung abbilden, alle drei Fächer betreffen. Ebenfalls bestehen in den Fachwissensitems (zumindest teilweise) inhaltliche Überschneidungen der naturwissenschaftlichen Kompetenzen. Dies gilt beispielsweise für das Basiskonzept

Energie, das in allen drei Fächern behandelt wird (Opitz et al. 2017). Für hohe inhaltliche Überschneidungen sprechen auch die sehr hohen Korrelationen zwischen den drei Fächern, die Wellnitz et al. (2017) auf der Basis von Normierungsdaten aus dem IQB berichten ($r=0,89$ bis $0,95$).

Schließlich wurde mithilfe des dritten Modells untersucht, inwiefern die Eindimensionalität des NEPS-Tests, die in der NEPS-Haupterhebung festgestellt wurde, auch für die in dieser Studie erhobenen Daten gilt. In diesem Modell wurde die naturwissenschaftliche Kompetenz im NEPS-Test, ähnlich wie in den LV-Tests im Modell 2, anhand der inhaltsbezogenen und prozessbezogenen Dimensionen spezifiziert.

Zur Untersuchung der dimensionalen Äquivalenz von NEPS und dem LV wurden die erhobenen Daten in Anlehnung an die Manuale der Studien (NEPS: Schöps und Saß 2013; LV: Hecht et al. 2013) kodiert und im Rahmen der IRT als ein

Abb. 2 Statistische Modelle zur Untersuchung dimensionaler Äquivalenz der NEPS- und LV-Daten (*FB* Umgang mit Fachwissen Biologie, *FC* Umgang mit Fachwissen Chemie, *FP* Umgang mit Fachwissen Physik, *EB* Erkenntnisgewinnung Biologie, *EC* Erkenntnisgewinnung Chemie, *EP* Erkenntnisgewinnung Physik, *IK* inhaltsbezogene Komponente, *PK* prozessbezogene Komponente)

einparametrisches logistisches Modell mit der Software ConQuest (Wu et al. 2007) analysiert. Als Schätzalgorithmus wurde für alle Modelle die Montecarlo-Integration mit 5000 Nodes verwendet.

Da komplexere Modelle die Daten tendenziell besser beschreiben können, aber aufgrund der Anzahl von Parametern, die sie dafür benötigen, weniger sparsam sind, werden die Ergebnisse konfirmatorischer Faktorenanalysen mit Hilfe informationstheoretischer Indizes bewertet und verglichen. Diese Indizes: Akaikes Information Criterion (AIC), Bayesian Information Criterion (BIC) und Sample-Size Adjusted BIC (SABIC) sind Maße für die Passung des statistischen Modells auf die erhobenen Daten. Das AIC (Akaike 1973) berücksichtigt neben der logarithmierten Wahrscheinlichkeit für das zu untersuchende Modell ($\text{Devianz} = -2 * \text{Log-Likelihood}$) auch die Anzahl der geschätzten Parameter in diesem Modell. Das BIC (Schwarz 1978) berücksichtigt die Anzahl der Parameter in den Modellen stärker als das AIC, indem es diese mit der Stichprobengröße gewichtet. Weniger stark als das BIC aber stärker als das AIC wird die Anzahl der Parameter durch das an die Stichprobengröße angepasste SABIC (Sclove 1987) berücksichtigt. Beim Vergleich der Modelle anhand der informationstheoretischen Indizes erklärt das Modell mit dem kleinsten Wert des betreffenden Maßes die Daten am besten.

4.3 Skalenäquivalenz und Linking

Zur Überprüfung der Skalenäquivalenz von NEPS und dem LV wurden die Verteilungen der Kompetenzwerte der beiden Hauptstudien anhand der Daten aus der Linking-Studie miteinander verglichen. Zur Sicherstellung der Vergleichbarkeit der Linking-Studie mit den Hauptstudien wurden die NEPS-Kompetenzwerte eindimensional und die LV-Kompetenzdaten sechsdimensional skaliert. Dazu wurden die Rohwerte des NEPS-Tests und der LV-Tests basierend auf dem 1PL-Rasch-Modell unter der Fixierung der Itemparameter aus dem NEPS 2010 und dem LV 2012 analysiert. Fünf *Plausible Values* (PVs) wurden pro Schülerin/Schüler und Test gezogen und linear auf den Mittelwert von $M=500$ und eine Standardabweichung von $SD=100$ transformiert.

Nissen et al. (2015) haben gezeigt, dass die Verteilungen der mit dem Equipercen-tile Equating verlinkten Werte im Vergleich zum IRT-Linking ähnlicher sind. Das Equipercen-tile Equating basiert auf der Idee, dass die Testwerte aus zwei Tests mit dem gleichen Perzentilrang als gleichwertig deklariert werden (Kolen und Brennan 2004). Wenn zum Beispiel 9,74 % der SuS im NEPS-Test 427 Punkte oder weniger erreichten und 9,74 % der SuS im LV-Test 461 Punkte oder weniger erreichten, dann werden die Punktwerte 427 und 461 als äquivalent erklärt. Da das Ziel des Linkings darin lag, die NEPS-Werte in die LV-Kompetenzstufen, die auf einem sechsdimensionalen Modell beruhen, einzuordnen, wurde die Schätzung der Personenfähigkeit im LV an das sechsdimensionale Modell angelehnt. Das Linking erfolgte getrennt für jeden der fünf PVs mit der Computer-Software LEGS (Brennan 2003). Aufgrund der Sensitivität des Equipercen-tile Equating für Unregelmäßigkeiten in der Verteilung von Testergebnissen (Livingston 2004), wurden die NEPS-Äquivalente in der LV-Metrik, angelehnt an Brennan (2003), mit einem Wert von 0,3 nachgeglättet, um auf diese Weise eine gleichmäßige Verteilung der Kompetenzwerte zu erreichen.

Im letzten Schritt wurden die untersuchten SuS auf die Kompetenzstufen des LV auf Basis ihrer Werte im LV und der NEPS-Äquivalente in der LV-Metrik eingeordnet. Die Grenzen für die Kompetenzstufen im LV (Biologie: Mayer et al. 2013; Chemie: Walpuski et al. 2013; Physik: Kauertz et al. 2013) variieren in Abhängigkeit vom Kompetenzbereich (Umgang mit Fachwissen oder Erkenntnisgewinnung) und dem Fach (Biologie, Chemie oder Physik). Die Einordnung in die Kompetenzstufen des LV wurde für jeden der fünf PVs getrennt vorgenommen. Anschließend wurden die Ergebnisse der Untersuchung der Klassifikationskorrektheit für jeden Inhaltsbereich des LV getrennt gemittelt. Zur Beurteilung der Äquivalenz von Verteilungen im NEPS und dem LV wurde mittels Chi-Quadrat-Tests die Annahme überprüft, ob die Verteilung der SuS auf die Kompetenzstufen des LV auf Grundlage der Tests aus beiden Studien gleich ist.

5 Ergebnisse

5.1 Dimensionale Äquivalenz

Das Ziel der Untersuchung der dimensionalen Äquivalenz lag in der Überprüfung der Vergleichbarkeit der dimensional Struktur der NEPS- und LV-Kompetenzwerte. Dazu wurden drei statistische Modelle geschätzt und hinsichtlich der Passung auf die erhobenen Daten miteinander verglichen (Abb. 2).

Die informationstheoretischen Indizes AIC und SABIC in Tab. 1 fallen für das Modell 3 am kleinsten aus und zeigen somit im Vergleich mit den anderen Modellen die beste Passung. Dieses Modell unterscheidet sowohl im NEPS als auch im LV inhaltsbezogene und prozessbezogene Komponenten der naturwissenschaftlichen Kompetenz. Im Unterschied zu AIC und SABIC bescheinigt das BIC dem zweiten Modell die bessere Modellpassung. Entsprechend kann geschlussfolgert werden, dass die dimensionale Struktur der NEPS- und LV-Daten am besten durch die inhalts- und prozessbezogenen Komponenten beschrieben werden kann. Dabei zeigen die Ergebnisse, dass Modell 1, das die fachspezifischen Faktoren im LV modelliert, die schlechteste Modellgüte aufweist. Somit ergeben die im Rahmen der

Tab. 1 Ergebnisse der Untersuchung der dimensional Struktur der NEPS- und LV-Daten

	<i>N</i> _{Parameter}	Devianz	AIC	BIC	SABIC
<i>Modell 1</i>					
(6 LV-Dimensionen & 1 NEPS-Dimension)	225	51.711	52.161	53.178	52.465
<i>Modell 2</i>					
(2 LV-Dimensionen & 1 NEPS-Dimension)	203	51.737	52.143	<i>53.061</i>	52.417
<i>Modell 3</i>					
(2 LV-Dimensionen & 2 NEPS-Dimensionen)	207	51.722	<i>52.136</i>	53.072	<i>52.415</i>

Die kleinsten Werte sind kursiv hervorgehoben

AIC Akaike information criterion, BIC Bayesian information criterion, SABIC Sample-Size Adjusted BIC

vorliegenden Studie erhobenen Daten keine fundierte empirische Unterstützung für die Annahme analytisch trennbarer fachspezifischer Faktoren in den LV-Aufgaben.

Die im Modell 3 ermittelten messfehlerkorrigierten Korrelationen der Kompetenzbereiche liegen bei 0,87 im LV (zwischen Umgang mit Fachwissen und Erkenntnisgewinnung) und bei 0,94 im NEPS (zwischen den inhaltsbezogenen und prozessbezogenen Komponenten). Die Korrelationen der LV-Kompetenzbereiche mit den NEPS-Kompetenzbereichen liegen im Bereich von 0,75 bis 0,82. Dementsprechend teilt der NEPS-Test 56–67 % seiner Varianz mit den LV-Tests.

5.2 Skalenäquivalenz und Linking zwischen NEPS und dem Bereich Umgang mit Fachwissen des LV

Die Untersuchung der dimensionalen Äquivalenz zeigte, dass sich die NEPS- und LV-Kompetenzdaten am besten durch ein Modell abbilden lassen, das in beiden Studien lediglich die inhalts- und die prozessbezogene Komponenten der naturwissenschaftlichen Kompetenz unterscheidet. Dies legt einen Vergleich der Verteilungen der Personenfähigkeiten im NEPS und im LV anhand dieses Modells nahe. Da allerdings das Ziel der Äquivalenzprüfung darin besteht, anhand der vorliegenden Linking-Studie die Vergleichbarkeit der Verteilungen der Kompetenzwerte in den Hauptstudien zu untersuchen, wurde die Schätzung der Personenfähigkeit im LV an das sechsdimensionale Modell angelehnt. Dieses Vorgehen folgt der Logik der Berichterstattung durch das IQB (vgl. Hecht et al. 2013). Des Weiteren wurde dieses Modell dem Linking zwischen NEPS und dem LV zugrunde gelegt. Auf diese Weise konnten die verlinkten Kompetenzwerte in die LV-Kompetenzstufen, die ebenfalls auf dem sechsdimensionalen Modell beruhen, eingeordnet werden.

Zur Untersuchung der Skalenäquivalenz und damit der zweiten Fragestellung werden die Verteilungen der Personenschätzungen im NEPS und im LV miteinander verglichen. In Tab. 2 wurden die Parameter zur Beschreibung der Verteilung von Personenfähigkeiten im NEPS und den Personenfähigkeiten im Umgang mit Fachwissen im LV zusammengefasst. Die Ergebnisse zeigen einen signifikanten Unter-

Tab. 2 Verteilung der Personenfähigkeit im NEPS, den Tests zum Umgang mit Fachwissen im LV und im NEPS in der LV-Metrik

	<i>M (SE)</i>	<i>SD</i>	Schiefte	<i>SD</i>	Exzess	<i>SD</i>
NEPS	540 (3,13)	81	−0,01	0,09	−0,37	0,19
Umgang mit Fachwissen Biologie (FB)	589 (3,65)	95	−0,11	0,09	−0,38	0,19
NEPS in der FB-Metrik	589 (3,82)	99	−0,10	0,09	−0,39	0,19
Umgang mit Fachwissen Chemie (FC)	601 (3,13)	81	−0,07	0,09	−0,07	0,19
NEPS in der FC-Metrik	601 (3,21)	83	−0,09	0,09	−0,24	0,19
Umgang mit Fachwissen Physik (FP)	610 (3,08)	80	−0,17	0,09	−0,21	0,19
NEPS in der FP-Metrik	608 (3,20)	83	−0,12	0,09	−0,24	0,19

M Mittelwert, *SE* Standardfehler, *SD* Standardabweichung

Tab. 3 Prozentuale Verteilung der SuS auf die LV-Kompetenzstufen im Umgang mit Fachwissen des LV und im NEPS in der LV-Metrik

		LV-Kompetenzstufen (in %)					Gesamt
		I	II	III	IV	V	
Biologie	LV	3	12	33	36	16	100
	NEPS _{LV}	3	12	33	36	16	100
Chemie	LV	3	11	37	30	19	100
	NEPS _{LV}	3	11	37	30	19	100
Physik	LV	1	7	28	36	28	100
	NEPS _{LV}	1	7	28	36	28	100

schied ($p < 0,05$) in den NEPS- und LV-Mittelwerten bei jeweils gleich ausgeprägter Schiefe und Exzess.

Der Unterschied in den Mittelwerten der Studien weist auf eine unterschiedliche Bewertung der Leistung von SuS in Abhängigkeit vom bearbeiteten Test hin. Werden den Personen beide Tests vorgelegt, erreichen sie im Durchschnitt im LV-Test einen höheren Skalenwert als im NEPS-Test, wenn die Items auf die Schwierigkeiten der jeweiligen Hauptstudie fixiert sind. Durch die Verlinkung wurden die beiden Verteilungen einander angeglichen. Die Mittelwerte zu den LV-Tests weisen zudem auf eine positive Selektion der Stichprobe hin. Die Mittelwerte, die um 600 schwanken, entsprechen den im Ländervergleich 2012 erreichten Werten an den Gymnasien.

Im nächsten Schritt erfolgte die Einordnung der SuS auf die Kompetenzstufen des LV auf der Grundlage der LV-Werte und der NEPS-Testwerte in der LV-Metrik (Tab. 3). Wie die Ergebnisse zeigen, sind die verlinkten NEPS- und LV-Skalen untereinander äquivalent. Zur Untersuchung der Klassifikationskonsistenz der beiden Studien und somit der dritten Fragestellung in dieser Arbeit wurde der Chi-Quadrat-Test auf Gleichverteilung durchgeführt. Der Chi-Quadrat-Wert liegt im Bereich zwischen $\chi^2 = 0,28$ im Umgang mit Fachwissen Biologie und $\chi^2 = 1,58$ im Umgang mit Fachwissen Physik und zeigt somit keine statistisch signifikante Abweichung der beiden Studien in der prozentualen Verteilung auf die Kompetenzstufen ($df = 4$, $n. s.$).

5.3 Skalenäquivalenz und Linking zwischen NEPS und dem Bereich der Erkenntnisgewinnung des LV

Bei der Untersuchung der Skalenäquivalenz zwischen NEPS und dem Bereich Erkenntnisgewinnung im LV zeigen sich, wie auch bereits bei der Untersuchung der Äquivalenz der NEPS- und LV-Skalen im Bereich des Umgangs mit Fachwissen, signifikante Mittelwertsunterschiede ($p < 0,05$) der beiden Skalen bei ansonsten gleich verteilter Schiefe und gleich verteiltem Exzess (Tab. 4).

Erneut zeigt sich, dass unsere Stichprobe positiv selektiert ist. Die erreichten Kompetenzwerte auf der LV-Skala entsprechen wiederum denen, die SuS an Gymnasien im LV 2012 erreicht haben. Durch die Verlinkung wurde die NEPS-Verteilung an die Verteilung des LV angeglichen, wodurch die Unterschiede in den Mittelwerten (44–65 Punkte) verschwunden sind.

Tab. 4 Verteilung der Personenfähigkeit im NEPS, den Tests zur Erkenntnisgewinnung im LV und im NEPS in der LV-Metrik

	<i>M (SE)</i>	<i>SD</i>	Schiefe	<i>SD</i>	Exzess	<i>SD</i>
NEPS	540 (3,13)	81	−0,01	0,09	−0,39	0,19
Erkenntnisgewinnung Biologie (EB)	584 (3,22)	84	−0,19	0,09	−0,15	0,19
NEPS in der EB-Metrik	584 (3,39)	88	−0,17	0,09	−0,19	0,19
Erkenntnisgewinnung Chemie (EC)	605 (4,11)	107	−0,14	0,09	−0,23	0,19
NEPS in der EC-Metrik	605 (4,28)	111	−0,17	0,09	−0,24	0,19
Erkenntnisgewinnung Physik (EP)	589 (3,84)	100	−0,13	0,09	−0,17	0,19
NEPS in der EP-Metrik	589 (3,98)	104	−0,17	0,09	−0,26	0,19

M Mittelwert, *SE* Standardfehler, *SD* Standardabweichung

Tab. 5 Prozentuale Verteilung der SuS auf die LV-Kompetenzstufen in Erkenntnisgewinnung des LV und im NEPS in der LV-Metrik

		LV-Kompetenzstufen (in %)					Gesamt
		I	II	III	IV	V	
Biologie	LV	4	16	43	31	6	100
	NEPS _{LV}	4	16	42	32	6	100
Chemie	LV	6	11	17	26	40	100
	NEPS _{LV}	6	11	17	26	40	100
Physik	LV	4	8	26	30	32	100
	NEPS _{LV}	4	8	26	29	33	100

Die Einordnung der verlinkten NEPS- und LV-Werte auf die Kompetenzstufen der Erkenntnisgewinnung in Tab. 5 zeigte, dass beide Skalen äquivalent verteilt sind. Der Chi-Quadrat-Wert liegt im Bereich zwischen $\chi^2=0,23$ in der Erkenntnisgewinnung Biologie und $\chi^2=0,45$ in der Erkenntnisgewinnung Physik ($df=4$, *n. s.*) und widerspricht somit der Annahme unterschiedlicher Verteilungen der NEPS- und LV-Kompetenzwerte der SuS auf die Kompetenzstufen des LV.

6 Zusammenfassung und Diskussion

Das Ziel dieser Studie war es, den Naturwissenschaftstest des NEPS für die neunte Klassenstufe mit den Kompetenzskalen des LV in den Naturwissenschaften für den Mittleren Schulabschluss zu verknüpfen. Eine hohe dimensionale Äquivalenz der Testwerte sowie eine hohe Skalenäquivalenz bilden die Basis für ein robustes Linking. Nachfolgend werden die Ergebnisse der Untersuchung beider Äquivalenzarten für die NEPS- und LV-Daten zusammengefasst und diskutiert.

6.1 Dimensionale Äquivalenz

Die erste Fragestellung zielte darauf ab, die Vergleichbarkeit der dimensional Struktur von NEPS- und LV-Kompetenzwerten zu überprüfen und wurde mithilfe von konfirmatorischen Faktorenanalysen untersucht. Die Ergebnisse dieser Analysen bescheinigten größtenteils dem Modell die beste Modellpassung, das in beiden Studien inhalts- sowie prozessbezogene Komponenten unterscheidet. Dieser Befund deckt sich für den LV mit den Ergebnissen der Studie von Kampa und Köller (2016) und spricht für die Vergleichbarkeit der dimensional Struktur der Kompetenzwerte im NEPS und dem LV.

Ebenfalls sprechen die messfehlerkorrigierten Korrelationen zwischen den Kompetenzskalen der Studien für einen hohen Zusammenhang zwischen den NEPS- und LV-Kompetenzwerten. Dieser Zusammenhang fiel kleiner aus als die messfehlerkorrigierte Korrelation der Mathematik-Werte im NEPS und dem LV für die neunte Klasse (van den Ham et al. 2017) und war ebenfalls kleiner als die messfehlerkorrigierte Korrelation der NEPS- und PISA-Naturwissenschaftswerte (Wagner et al. 2018). Seine Höhe lag allerdings sehr nah an den in der vorliegenden Studie ermittelten Zusammenhängen der einzelnen LV-Tests (Biologie, Chemie und Physik im Modell 1). Ausgehend von der Annahme, dass die Naturwissenschaftstests des LV verwandte Konstrukte messen, kann der in dieser Studie vorgefundene latente Zusammenhang als Evidenz für die hohe Ähnlichkeit der Konstrukte im NEPS und dem LV interpretiert werden. Das bedeutet, dass die Testinstrumente beider Studien in Bezug auf die sich überschneidenden Kompetenzbereiche miteinander vergleichbar sind.

Gleichzeitig konnte bei der Untersuchung der Dimensionalität von NEPS- und LV-Kompetenzwerten keine fachliche Spezifität der Inhalte im LV festgestellt werden. Dieses Ergebnis steht scheinbar im Widerspruch zu den Ergebnissen der Studie von Wellnitz et al. (2017) für den Bereich der Erkenntnisgewinnung. Dort zeigte sich die Überlegenheit des domänenspezifischen Modells (Biologie, Chemie und Physik) im Vergleich zum domänenübergreifenden Modell (naturwissenschaftliche Untersuchungen). Gleichzeitig aber lagen die Korrelationen zwischen den fachspezifischen Faktoren bei Wellnitz et al. (2017) ($r=0,89$ bis $r=0,95$) noch höher als in der vorliegenden Untersuchung. Dass trotz dieser nahe an 1 liegenden Korrelationen ein Vorteil des dreidimensionalen Modells abgesichert werden konnte, mag an der deutlich größeren Stichprobe (über 9000 SuS) und der damit verbundenen größeren Teststärke zum Aufdecken von Modelldifferenzen gelegen haben. Auch basierte die Modellierung der Kompetenzdaten in der Studie von Wellnitz et al. (2017), anders als im LV und in unserer Studie, auf dem zweiparametrischen logistischen Modell, das zusätzlich zu den Parametern des Rasch-Modells die Trennschärfe der Items berücksichtigt. Möglicherweise hat dies auch die Befunde moderiert. Natürlich kann auch nicht ausgeschlossen werden, dass die vergleichsweise geringere Stichprobe in der Linking-Studie dazu geführt haben könnte, dass ein fachspezifisches Modell nicht gefunden werden konnte, da die statistische Power zur Aufdeckung von Mehrdimensionalität reduziert war. Zusammenfassend geben unsere Befunde, wie auch die von Haschke et al. (2017), erhebliche Evidenz dafür, dass die drei naturwissenschaftlichen Fächer Biologie, Chemie und Physik große Anteile gemeinsamer

Varianz binden, die durch die inhaltsbezogene und prozessbezogene Komponenten naturwissenschaftlicher Kompetenz erfasst werden können.

Die Untersuchung der konzeptionellen Äquivalenz zeigte bereits eine große inhaltliche Überschneidung zwischen dem NEPS-Test und den Tests des LV in den Bereichen des Umgangs mit Fachwissen und der Erkenntnisgewinnung. Auch die Analyse der latenten Zusammenhänge und der dimensional Struktur ihrer Kompetenzdaten legen nahe, dass die zu vergleichenden Testinstrumente äquivalente Kernkompetenzen messen und daher als gegenseitig anschlussfähig betrachtet werden können.

6.2 Skalenäquivalenz und Linking

Bei der Untersuchung der Skalenäquivalenz und damit der zweiten Fragestellung wurde ein Mittelwertunterschied zwischen den Originalskalen des NEPS und des LV festgestellt: Die Testpersonen erreichten im LV-Test im Durchschnitt einen höheren Fähigkeitswert als im NEPS-Test. Die im nächsten Schritt durchgeführte Verlinkung der NEPS-Skala mit den Skalen des LV glich den NEPS-Mittelwert an die entsprechenden Mittelwerte des LV an. Die Untersuchung der dritten Fragestellung bezüglich der Robustheit des Linkings ergab, dass die Skalen des NEPS und des LV in der Population der SuS der neunten Klasse äquivalente Verteilungen auf die Kompetenzstufen des LV aufweisen. Somit kann also die in der vorliegenden Studie ermittelte Linking-Funktion als ausreichend robust angesehen werden.

Die Verbindung der Kompetenzskalen des NEPS und des LV durch die vorliegende Arbeit kann als eine Erweiterung der Testwertinterpretationen für beide Studien angesehen werden. Da die NEPS-Studie bisher keinen Gebrauch von Kompetenzstufenmodellen macht, war es bis jetzt nicht möglich, die im Rahmen dieser Studie erhobenen Testwerte kriterial zu interpretieren. Der LV bietet zwar die Zuordnung der Testwerte zu den Bildungsstandards, allerdings fehlt die Erklärung dazu, wie die beobachtete Leistung zustande gekommen ist. Die Verbindung der Kompetenzskalen aus beiden Studien schafft die notwendigen Bedingungen für die Einordnung der NEPS-Ergebnisse in einen nationalen Referenzrahmen und schließt somit die Lücken der beiden Studien. Da die untersuchte Stichprobe allerdings im Vergleich zur LV-Stichprobe positiv selektiert ist, sollte die in dieser Studie ermittelte Linking-Funktion vor allem für SuS im mittleren bis hohen Kompetenzbereich robust sein, für den unteren Leistungsbereich sollten in zukünftigen Studien weniger positiv selektierte Stichproben gezogen werden. Zusammenfassend kann hier also geschlussfolgert werden, dass das Linking in der vorliegenden Studie eine Verbindung zwischen dem NEPS-Test und den LV-Tests herstellen kann, die eine Erweiterung der Testwertinterpretationen beider Studien ermöglicht und somit für die Bildungsforschung von Relevanz ist.

6.3 Limitationen

Limitierende Faktoren der vorliegenden Studie liegen zum einen in der Größe und zum anderen in der Selektivität der untersuchten Stichprobe. Kolen und Brennan (2004) nennen zwar die Stichprobengröße von $N=250$ als ausreichend, um eine

Verlinkung mittels des Equipercenile Equating durchzuführen. Jedoch wäre bei einer größeren Stichprobe vermutlich sowohl die Robustheit des Linkings als auch die Aussagekraft der Ergebnisse zur Untersuchung der dimensional Struktur höher. Eine weitere Limitation ergab sich daraus, dass die untersuchte Stichprobe positiv selektiert war und auf den Skalen des LV deutlich höhere Kompetenzwerte erreichte als die repräsentative Stichprobe von Neuntklässlern, die im LV 2012 getestet wurde. Die hohe Übereinstimmung unserer Befunde mit denen von Haschke et al. (2017), in denen eine deutlich heterogenere Stichprobe verwendet wurde, weist trotz verzerrter Stichprobe auf die Validität unserer Befunde hin. Auf die Grenzen bei der Klassifikation nach Kompetenzstufen sind wir oben eingegangen. Zuletzt sollte angemerkt werden, dass die eingesetzten LV-Aufgaben nur eine Auswahl der Aufgaben der Hauptstudie darstellen und die zugrundeliegenden Konstrukte nicht in ihrer vollen Breite erfasst wurden. Aus diesem Grund wäre es wünschenswert, die Ergebnisse der vorliegenden Studie mit größeren Itemzahlen in größeren Personenstichproben zu replizieren.

Förderung Das Projekt wurde gefördert durch das Zentrum für internationale Vergleichsstudien (ZIB) und das Bundesministerium für Bildung und Forschung (BMBF).

Open Access Dieser Artikel wird unter der Creative Commons Namensnennung 4.0 International Lizenz (<http://creativecommons.org/licenses/by/4.0/deed.de>) veröffentlicht, welche die Nutzung, Vervielfältigung, Bearbeitung, Verbreitung und Wiedergabe in jeglichem Medium und Format erlaubt, sofern Sie den/die ursprünglichen Autor(en) und die Quelle ordnungsgemäß nennen, einen Link zur Creative Commons Lizenz beifügen und angeben, ob Änderungen vorgenommen wurden.

Literatur

- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Hrsg.), *Second international symposium on information theory* (S. 267–281). Budapest: Akademiai Kiado.
- Blossfeld, H.-P., von Maurice, J., & Schneider, T. (2011). *Grundidee, Konzeption und Design des Nationalen Bildungspanels für Deutschland* (NEPS Working Paper No. 1). Bamberg: Otto-Friedrich-Universität.
- Blossfeld, H.-P. (2008). *Education as a lifelong process. A proposal for a national educational panel study (NEPS) in Germany. Part B: Theories, operationalizations and piloting strategies for the proposed measurements* (Unveröffentlichter BMBF-Antrag). Bamberg: Universität Bamberg.
- Brennan, R. L. (2003). *LEGS: A computer program for linking with the randomly equivalent groups or single-group design* (Version 2.0). Iowa City: University of Iowa – Center of Advanced Studies in Measurement and Assessment.
- Bybee, R. W. (1997). Towards an understanding of scientific literacy. In W. Gräber & C. Bolte (Hrsg.), *Scientific literacy—An international symposium* (S. 37–68). Kiel: Institut für die Pädagogik der Naturwissenschaften.
- Hahn, I., Schöps, K., Rönnebeck, S., Martensen, M., Hansen, S., Saß, S., & Prenzel, M. (2013). Assessing scientific literacy over the lifespan—A description of the NEPS science framework and the test development. *Journal for Educational Research Online*, 5(2), 110–138.
- van den Ham, A.-K., Ehmke, T., Nissen, A., & Roppelt, A. (2017). Assessments verbinden, Interpretationen erweitern? *Zeitschrift für Erziehungswissenschaft*, 20(1), 89–111.
- Hanushek, E. A., & Wößmann, L. (2015). *The knowledge capital of nations: education and the economics of growth*. Cambridge: MIT Press.
- Haschke, L., Hahn, I., Bernholt, S., & Ihme, J.-M. (2017). Zur Dimensionalität des PISA-Naturwissenschaftstests am Beispiel naturwissenschaftlicher Grundbildung und Fachkompetenzen. *Unterrichtswissenschaft*, 45(3), 254–268.
- Hecht, M., Roppelt, A., & Siegle, T. (2013). Testdesign und Auswertung des LV. In H. A. Pant, P. Stanat, U. Schröders, A. Roppelt, T. Siegle & C. Pöhlmann (Hrsg.), *LV 2012. Mathematische und Naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I* (S. 391–402). Münster: Waxmann.

- Kampa, N., & Köller, O. (2016). German national proficiency scales in biology: internal structure, relations to general cognitive abilities and verbal skills. *Science Education*, 100(5), 903–922.
- Kauert, A., & Fischer, H. E. (2013). Die Operationalisierung naturwissenschaftlicher Kompetenzen im LV 2012. In H. A. Pant, P. Stanat, U. Schröders, A. Roppelt, T. Siegle & C. Pöhlmann (Hrsg.), *LV 2012. Mathematische und Naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I* (S. 50–52). Münster: Waxmann.
- Kauert, A., Fischer, H. E., & Jansen, M. (2013). Kompetenzstufenmodelle für das Fach Physik. In H. A. Pant, P. Stanat, U. Schröders, A. Roppelt, T. Siegle & C. Pöhlmann (Hrsg.), *LV 2012. Mathematische und Naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I* (S. 92–100). Münster: Waxmann.
- Kauert, A., Fischer, H. E., Mayer, J., Sumfleth, E., & Walpuski, M. (2010). Standardbezogene Kompetenzmodellierung in den naturwissenschaftlichen Fächern der Sekundarstufe I. *Zeitschrift für Didaktik der Naturwissenschaften*, 16, 135–153.
- KMK (2006). *Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring*. München: Wolters Kluwer.
- Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: methods and practices*. New York: Springer.
- Linn, R. L. (1993). Linking results of distinct assessments. *Applied Measurement in Education*, 4, 185–207.
- Livingston, S. A. (2004). *Equating test scores (without IRT)*. Princeton: ETS Educational Testing Service.
- von Maurice, J., Sixt, M., & Blossfeld, H.-P. (2011). *The German national educational panel study: surveying a cohort of 9th graders in Germany* (NEPS Working Paper No. 3). Bamberg: Otto-Friedrich-Universität.
- Mayer, J., Wellnitz, N., Klebba, N., & Kampa, N. (2013). Kompetenzstufenmodelle für die naturwissenschaftlichen Fächer. In H. A. Pant, P. Stanat, U. Schröders, A. Roppelt, T. Siegle & C. Pöhlmann (Hrsg.), *LV 2012. Mathematische und Naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I* (S. 74–100). Münster: Waxmann.
- Mislevy, R. J. (1992). *Linking educational assessments: concepts, issues, methods, and prospects*. Princeton: ETS Policy Information Center.
- Nissen, A., Ehmke, T., Köller, O., & Duchhardt, C. (2015). Comparing apples with oranges? An approach to link TIMSS and the National Educational Panel Study in Germany via equipercenile and IRT methods. *Studies in Educational Evaluation*, 47, 58–67. <https://doi.org/10.1016/j.stueduc.2015.07.003>.
- OECD (2006). *Assessing scientific, reading and mathematical literacy: a framework for PISA 2006*. Paris: OECD.
- Opitz, S., Harms, U., Neumann, K., & Bernholt, S. (2017). Students' energy understanding across biology, chemistry, and physics contexts. *Research in Science Education (RISE)*. <https://doi.org/10.1007/s11165-017-9632-4>.
- Pietsch, M., Böhme, K., Robitzsch, A., & Stubbe, T. C. (2009). Das Stufenmodell zur Lesekompetenz der länderübergreifenden Bildungsstandards im Vergleich zu IGLU 2006. In D. Ganzer, O. Köller, A. Bremerich-Vos, M. v.d. Heuvel-Panhuizen, K. Reiss & G. Walther (Hrsg.), *Bildungsstandards Deutsch und Mathematik. Leistungsmessung in der Grundschule* (S. 393–416). Weinheim: Beltz.
- 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.
- Prenzel, M., & Ostermeier, C. (2006). Improving mathematics and science instruction: a program for the professional development of teachers. In F. K. Oser, F. Achtenhagen & U. Reynolds (Hrsg.), *Competence oriented teacher training. Old research demands and new pathways* (S. 79–96). Rotterdam: Sense Publisher.
- Schöps, K., & Saß, S. (2013). *NEPS technical report for science. Scaling results of starting cohort 4 in ninth grade* (NEPS Working Paper No. 23). Bamberg: Otto-Friedrich-Universität.
- Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2), 461–464.
- Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika*, 52, 333–343.
- Siegle, T., Schroeders, U., & Roppelt, A. (2013). Anlage und Durchführung des LV. In H. A. Pant, P. Stanat, U. Schröders, A. Roppelt, T. Siegle & C. Pöhlmann (Hrsg.), *LV2012. Mathematische und Naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I* (S. 101–121). Münster: Waxmann.
- van de Vijver, F. J. R. (1998). Towards a theory of bias and equivalence. In J. Harkness (Hrsg.), *ZUMA-Nachrichten Spezial* (Bd. 3, S. 41–65). Mannheim: ZUMA.
- Wagner, H., Hahn, I., Schöps, K., Ihme, J. M., & Köller, O. (2018). Are the tests scores of the Programme for International Student Assessment (PISA) and the National Educational Panel Study (NEPS)

- science tests comparable? An assessment of test equivalence in German Schools. *Studies in Educational Evaluation*, 59, 278–287. <https://doi.org/10.1016/j.stueduc.2018.09.002>.
- Wagner, H., Schöps, K., Hahn, I., Pietsch, M., & Köller, O. (2014). Konzeptionelle Äquivalenz von Kompetenzmessungen in den Naturwissenschaften zwischen NEPS, LV und PISA. *Unterrichtswissenschaft*, 42(4), 301–320.
- Walpuski, M., Kampa, N., Kauertz, A., & Wellnitz, N. (2008). Evaluation der Bildungsstandards in den Naturwissenschaften. *Der mathematische und naturwissenschaftliche Unterricht*, 61(6), 323–326.
- Walpuski, M., Sumfleth, E., & Pant, H. A. (2013). Kompetenzstufenmodelle für das Fach Chemie. In H. A. Pant, P. Stanat, U. Schröders, A. Roppelt, T. Siegle & C. Pöhlmann (Hrsg.), *LV 2012. Mathematische und Naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I* (S. 83–91). Münster: Waxmann.
- Wellnitz, N., Fischer, H. E., Kauertz, A., Mayer, J., Neumann, I., & Pant, H. A. (2012). Evaluation der Bildungsstandards – eine fächerübergreifende Testkonzeption für den Kompetenzbereich Erkenntnisgewinnung. *Zeitschrift für Didaktik der Naturwissenschaften*, 18, 261–291.
- Wellnitz, N., Hecht, M., Heitmann, P., Kauertz, A., Mayer, J., & Sumfleth, E. (2017). Modellierung des Kompetenzteilbereichs naturwissenschaftliche Untersuchungen. *Zeitschrift für Erziehungswissenschaft*, 20(3), 556–584.
- Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). *ACER ConQuest Version 2: Generalised item response modelling software*. Camberwell: Australian Council for Educational Research.