Untersuchung der Äquivalenz der naturwissenschaftlichen Messung in der Sekundarstufe I

Dissertation zur Erlangung des Doktorgrades

der Philosophischen Fakultät

der Christian-Albrechts-Universität

zu Kiel

vorgelegt von

Diplom-Psychologin Helene Wagner

Kiel, im Februar 2018

Erstgutachter: Prof. Dr. Olaf Köller

Zweitgutachterin: Prof. Dr. Friederike Zimmermann

Tag der mündlichen Prüfung: 23.05.2018

Durch den Prodekan für Studium und Lehre, Prof. Dr. Elmar Eggert

zum Druck genehmigt am: 05.06.2018

Die Forschungsarbeiten, die in dieser Dissertation berichtet werden, wurden im Rahmen des Projekts „PISA, Bildungsstandards und das NEPS – Bildungsverläufe in Deutschland. Vergleich der Rahmenkonzepte und Validierung der NEPS-Testinstrumente in den Naturwissenschaften und in der Mathematik“, das vom Bundesministerium für Bildung und Forschung gefördert wurde (Projektnummer 323-21381-LSA009) durchgeführt.

Danksagung

Die Danksagung am Anfang einer Abschlussarbeit ist zu einer schönen Tradition geworden. Sie bietet die Möglichkeit all den Menschen zu danken, die den Prozess der Fertigstellung dieser Arbeit begleitet haben und ohne die diese Dissertation nie zustande gekommen wäre.

Mein besonderer Dank gilt Prof. Dr. Köller. Ich danke ihm für die Möglichkeit der Promotion am IPN und die stetige Unterstützung meiner Arbeit. Insbesondere möchte ich mich bei ihm für unsere regelmäßigen Publikationstreffen bedanken, die mir einerseits als Deadline für die Fertigstellung meiner Publikationen gedient und andererseits den Rahmen für inhaltliche Fragen gegeben haben. Außerdem möchte ich mich ganz herzlich bei meiner Zweitgutachterin, Prof. Dr. Friederike Zimmermann, sowie bei Prof. Dr. Thilo Kleickmann, Prof. Dr. Kerstin Kremer und Prof. Dr. Uta Klusmann bedanken, die Interesse an meiner Arbeit gezeigt und diese begutachtet haben.

Mein größter Dank gilt meiner Erstbetreuerin, Dr. Inga Hahn, für ihre Zuverlässigkeit und ihr Engagement bei der Betreuung meiner Dissertation. Insbesondere möchte ich mich bei ihr für die vielen wertvollen Kommentare und Anregungen bedanken, die zur Qualität meiner Arbeit entscheidend beigetragen haben. Mein herzlichster Dank gilt ebenfalls meiner Zweitbetreuerin Dr. Katrin Schöps für die hilfreichen Tipps und Verbesserungsvorschläge bei der Fertigstellung dieser Dissertation. Meinen beiden Betreuerinnen danke ich für ihr ehrliches Interesse an meiner Arbeit und dafür, dass sie mir immer mit Rat und Tat zur Seite gestanden haben.

Weiterhin möchte ich mich bei meinen NEPS-Kolleginnen und –Kollegen für eine tolle Zusammenarbeit bedanken, die mir sehr viel Spaß gemacht hat. Mein besonderer Dank gilt hier Dr. Jan Marten Ihme für seine Unterstützung in den methodischen Fragen und Jana Kähler für die Unterstützung im NEPS-Projekt. An dieser Stelle möchte ich mich auch bei Prof. Dr. Timo Ehmke, Dr. Ann-Katrin van den Ham sowie Dr. Annika Nissen von der Uni Lüneburg für einen bereichernden inhaltlichen und methodischen Austausch in unserem gemeinsamen Projekt bedanken. Mein herzlicher Dank gilt auch den Kolleginnen und Kollegen am IPN, die das Audit *Beruf und Familie* ins Leben gerufen haben. Die Errungenschaften des Audits haben mir in vielerlei Hinsicht die Promotion erleichtert.

Darüber hinaus möchte ich mich bei meinen Eltern, meinem Ehemann und meinen Schwiegereltern für Ihre Unterstützung bei allen den Haushalt sowie die Kinderbetreuung betreffenden Tätigkeiten bedanken. Ich danke auch meinen Jungs Robin und Erik dafür, dass sie mich zwischendurch daran erinnert haben, dass es andere schöne Dinge im Leben außer den Publikationen gibt.

Inhaltsverzeichnis

1

Rahmenschrift

1.1. Einleitung

„Es gibt nur eins, was auf Dauer teurer ist als Bildung, keine Bildung."

(John F. Kennedy)

Heutzutage wird die Teilhabe an Bildung als Bürger -mehr sogar- als Menschenrecht verstanden (Europäische Menschenrechtskonvention, Artikel 2 des Zusatzprotokolls). Allerdings benötigen die Bürgerinnen und Bürger von heute nicht einfach nur einen Zugang zur Bildung, vielmehr muss diese bestimmten Qualitätsanforderungen genügen, um die Menschen auf die Herausforderungen einer modernen und sich ständig verändernden Welt vorzubereiten.

Um die Qualität von Bildungssystemen zu überprüfen, wurden sowohl national als auch international verschiedene groß angelegte Schulleistungsstudien ins Leben gerufen. Die regelmäßige Teilnahme Deutschlands an internationalen Schulleistungsstudien wurde am 24. Oktober 1997 von der Ständigen Konferenz der Kultusminister der Länder der Bundesrepublik Deutschland (KMK) als Reaktion auf die enttäuschenden Ergebnisse bei TIMSS (Baumert et al., 1997) festgelegt. So wurde als Teil der Gesamtstrategie zum Bildungsmonitoring in Deutschland die Teilnahme an PISA *(Programme for International Student Assessment*) in der Sekundarstufe I sowie IGLU (*Internationale Grundschul-Lese-Untersuchung)* und TIMSS (*Trends in International Mathematics and Science Study*) in der Primarstufe beschlossen. Weiterhin wurde gefordert, deutschlandweit verbindliche Bildungsstandards in den Kernfächern für mehrere Schulabschlüsse einzuführen, deren Erreichen regelmäßig im Rahmen der Ländervergleichsstudien überprüft werden sollte.

Die Kompetenzmessungen werden in den oben genannten Studien durch ein querschnittliches Design realisiert. Allerdings bieten solche Querschnittstudien lediglich eine Momentaufnahme des Leistungsstandes der entsprechenden Schülerinnen und Schüler und lassen keine Aussagen über ihre Leistungsentwicklungen zu. Zur Verwirklichung dieser Zielsetzung sind Untersuchungen im Längsschnittdesign erforderlich. Aus diesem Grund wurde 2009 das Nationale Bildungspanel (*National Educational Panel Study* – NEPS) ins Leben gerufen, dessen Ziel es ist, „zentrale Bildungsprozesse und –verläufe in Deutschland über die gesamte Lebensspanne zu beschreiben und zu analysieren" (Blossfeld, Schneider & Doll, 2009, S. 249). Dabei wird zum Zweck der gemeinsamen Interpretation der Ergebnisse die Anschlussfähigkeit an nationale und internationale Large-Scale-Assessments in Deutschland angestrebt (Blossfeld, 2008).

Mit der Anschlussfähigkeit der Tests können verschiedene Ziele verfolgt werden: Zum einen wird durch die inhaltliche Anlehnung an bereits bestehende Testinstrumente eine Basis für die gemeinsame Interpretation ihrer Ergebnisse gelegt, die zur Klarheit und Transparenz der abgeleiteten Schlussfolgerungen beitragen würde. Zum anderen ermöglicht die Vergleichbarkeit der Testinstrumente aus verschiedenen Studien eine gegenseitige Erweiterung ihrer Testwertinterpretationen. Gelingt beispielsweise die Übertragung der Kompetenzskalen des Ländervergleichs (LV) in Biologie, Chemie und Physik auf die naturwissenschaftlichen Werte des NEPS, so könnte mithilfe der längsschnittlichen Analysen im NEPS untersucht werden, welche Faktoren des Kompetenzerwerbs besonders gut dafür geeignet sind, die Leistung im LV vorherzusagen.

Bisher wurde die Anschlussfähigkeit der NEPS-Testinstrumente lediglich im Bereich der Mathematik überprüft (van den Ham, Ehmke, Nissen, & Roppelt, 2016; Nissen, Ehmke, Köller, & Duchhardt, 2015). In den Naturwissenschaften fehlte dieser Vergleich bis jetzt. Diese Lücke wird durch die vorliegende Dissertation geschlossen, in der die Anschlussfähigkeit des NEPS-Naturwissenschaftstests für die neunte Klassenstufe am Ländervergleich 2012 in den Fächern Biologie, Chemie und Physik sowie an der Messung der naturwissenschaftlichen Kompetenz in PISA 2012 untersucht wird.

Im Fokus der empirischen Studien steht die Überprüfung der Konstrukt-Äquivalenz des NEPS-Naturwissenschaftstests mit den PISA- und LV-Naturwissenschaftstests. Die erste Studie befasst sich mit der konzeptionellen Vergleichbarkeit des NEPS-Tests mit den PISA- und LV-Tests. Die Untersuchung der dimensionalen und der skalenbezogenen Äquivalenz erfolgt für den NEPS- und den PISA-Test in der zweiten Studie sowie für den NEPS-Test und die LV-Tests in der dritten Studie. Im theoretischen Hintergrund der vorliegenden Arbeit wird zunächst die naturwissenschaftliche Kompetenz definiert sowie ihre Messung in den zu vergleichenden Studien dargestellt. Anschließend werden die Ziele dieser Arbeit herausgearbeitet. Im nächsten Schritt erfolgt die Darstellung des Ansatzes von Kolen und Brennen (2004) zum Vergleich der Testinstrumente aus verschiedenen Studien. Darauf aufbauend werden die Methoden und die Studien zur Verlinkung der Ergebnisse von verschiedenen Testinstrumenten beschrieben. Abgeschlossen wird der Theorieteil mit der Darstellung des kulturvergleichenden Ansatzes von van de Vijver (1998) zur Untersuchung der Konstrukt-Äquivalenz, aus dem die zentralen Fragestellungen abgeleitet werden. In den darauf folgenden Kapiteln werden die empirischen Studien der Dissertation vorgestellt. Anschließend werden die Ergebnisse dieser Studien zusammengefasst und diskutiert.

Abgeschlossen wird die vorliegende Arbeit mit der Implikation der Befunde und den Limitationen der empirischen Studien.

1.2. Theoretischer Hintergrund

1.2.1. Definition der naturwissenschaftlichen Kompetenz

Eine der Kernkompetenzen, die in vielen nationalen und internationalen Schulleistungsstudien untersucht wird, ist die naturwissenschaftliche Kompetenz. Sie ist eine Voraussetzung für die gesellschaftliche Teilhabe in einer durch Naturwissenschaften und Technik geprägten Welt (Prenzel, 2000; Prenzel et al., 2001; Rost et al., 2004) und wird als Prädiktor für eine erfolgreiche wirtschaftliche, soziale und kulturelle Teilnahme an der modernen Wissensgesellschaft verstanden (OECD, 2006; Prenzel, Schöps et al. 2007).

Die besondere Bedeutung der naturwissenschaftlichen Kompetenz für die moderne Gesellschaft wurde in der Studie von Hanushek und Wößmann (2015) hervorgehoben. Ihre Analysen zeigen, dass der in der PISA-Studie gemessene Erfolg der Schülerinnen und Schüler in den Naturwissenschaften am Ende der Sekundarstufe I den stärksten Prädiktor für das Wirtschaftswachstum eines Landes darstellt. Somit leistet die Messung der naturwissenschaftlichen Kompetenz zusätzlich zu den eigentlichen Zielen der Studien einen wichtigen Beitrag zur Abschätzung des wirtschaftlichen Erfolges eines Landes.

Das NEPS lehnt sich in der Definition der naturwissenschaftlichen Kompetenz zum einen an den Kompetenzbegriff an, wie er von Weinert (2001) beschrieben und von Klieme und Leutner (2006) erweitert wurde. Zum anderen ist die Definition der naturwissenschaftlichen Kompetenz im NEPS an das Konzept der Scientific Literacy (American Association for the Advancement of Science, 2009; Bybee, 1997) angelehnt, die gleichzeitig die Grundlage für die Kompetenzmessung in PISA bildet. Scientific Literacy entspricht der Vorstellung einer naturwissenschaftlichen Grundbildung für alle und wird als Voraussetzung für lebenslanges Lernen verstanden (OECD, 2006). Die Auswahl der Konzepte im NEPS erfolgte in Anlehnung an PISA (OECD, 2006), die Standards der AAAS (2009) und die Bildungsstandards für den Mittleren Bildungsabschluss (KMK, 2005a, KMK, 2005b, KMK, 2005c) und bietet dadurch die Basis für den Vergleich des NEPS-Testinstruments mit den Testinstrumenten von PISA und dem LV. Inwieweit dieser Vergleich und die darauf beruhende gegenseitigen Verknüpfung für die Studien NEPS, PISA und LV gewinnbringend sein kann, wird in den kommenden Abschnitten erläutert.

1.2.2. NEPS und Ländervergleich

Das Ziel des Ländervergleichs liegt in der Untersuchung der Kompetenzen von Schülerinnen und Schülern im Hinblick auf die durch die KMK verabschiedeten Bildungsstandards. Diese Untersuchung findet in den Fächern Deutsch, Mathematik, Fremdsprachen und Naturwissenschaften in einem regelmäßigen Zyklus in der vierten und/oder neunten Klassenstufe statt und ermöglicht die Abbildung des Bildungstrends aller Bundesländer in der Bundesrepublik Deutschland.

Das Ziel des NEPS liegt ebenfalls in der Messung von Kompetenzen, allerdings hat diese nationale Studie einen längsschnittlichen Charakter und umfasst die Untersuchung der Kompetenzentwicklung vom Kindes- bis ins hohe Erwachsenenalter u.a. in den Domänen Lesen, Mathematik, ICT- (Informations and Communications Technology) Literacy und Naturwissenschaften. Folglich werden im NEPS in regelmäßigen Zeitabständen dieselben Personen getestet, um dadurch ihre Kompetenzentwicklung abzubilden. Die Zielsetzungen beider Studien bedingen auch die Grenzen ihrer Schlussfolgerungen: Während die Ergebnisse des LV keine Aussagen über die Kompetenzentwicklung zulassen, können die Kompetenzwerte im NEPS nicht kriterial interpretiert werden. Die Anschlussfähigkeit der Testinstrumente der beiden Studien kann diesem Umstand entgegen wirken, indem die Testwertinterpretationen beider Studien erweitert werden (Nissen et al., 2015; van den Ham et al., 2016).

Aufbauend auf den Ergebnissen der vorliegenden Dissertation könnte beispielsweise untersucht werden, auf welcher Kompetenzstufe des LV die Kompetenz der Schülerinnen und Schüler am Ende der Sekundarstufe I liegen muss, damit sie einen erfolgreichen Abschluss am Ende der Sekundarstufe II erlangen können. Der Bezug der NEPS-Testwerte zu den nationalen Bildungsstandards würde sich damit für beide Studien als gewinnbringend erweisen.

1.2.3. NEPS und PISA

PISA dient in erster Linie dem Bildungsmonitoring und schafft die Möglichkeit, Vergleichsaussagen auf internationaler Ebene zu treffen. Bedingt durch den internationalen Charakter und den Fokus der Studie auf Grundbildung, verzichtet PISA auf eine enge Orientierung an Lehrplänen. Die Untersuchung der Kompetenzen (Lesen, Mathematik, Naturwissenschaften, Problemlösen und neuerdings finanzielle Allgemeinbildung) findet bei PISA im regelmäßigen Zyklus an fünfzehnjährigen Schülerinnen und Schülern statt. Für jede

dieser Erhebungen sieht PISA einen inhaltlichen Schwerpunkt vor, auf den in der jeweiligen Erhebungswelle der größte Teil der Testzeit entfällt. Es wird jedoch für die anderen Domänen ausreichend Testzeit vorgesehen, um aussagekräftige Vergleiche des Kompetenzstandes von Schülerinnen und Schülern zwischen den Staaten und über die Zeit vornehmen zu können (Prenzel, Artelt et al., 2007).

Durch die Untersuchung der Anschlussfähigkeit des NEPS-Tests an den PISA-Test und der anschließenden Verlinkung ihrer Skalen könnten die NEPS-Testwerte im internationalen Referenzrahmen von PISA eingeordnet werden. Doch auch die Testwertinterpretationen von PISA könnten durch die Verlinkung mit NEPS-Testwerten erweitert werden. Die Bereicherung für PISA könnte beispielsweise in der längsschnittlichen Untersuchung der Kompetenzentwicklung der Schülerinnen und Schüler bestehen, deren NEPS-Kompetenzwerte sich in den unteren bzw. höheren PISA-Kompetenzstufen wiederfinden. Auf diese Weise könnte man mehr über besondere Bedingungsfaktoren der Kompetenzentwicklung in verschiedenen Kompetenzbereichen erfahren.

Weiterhin kann die Anschlussfähigkeit der Testinstrumente dafür genutzt werden, um die neu entwickelten Tests anhand bereits bestehender Testverfahren zu validieren. Ein Beispiel hierfür ist die Studie von Hartig und Frey (2012), in welcher der neu entwickelte LV-Test in Mathematik anhand des PISA-Mathematiktests validiert wurde. Auf ähnliche Weise könnte die Untersuchung der Vergleichbarkeit der naturwissenschaftlichen Messung in NEPS und PISA aufgrund ihrer inhaltlichen Nähe neue Erkenntnisse hinsichtlich der Validität des NEPS-Tests liefern.

1.2.4. Ziele dieser Arbeit

Auf Basis der Ausführungen in den vorhergehenden Abschnitten soll im Rahmen der vorliegenden Arbeit geprüft werden, inwiefern die naturwissenschaftliche Messung in der NEPS-Studie mit den entsprechenden Messungen in den Studien PISA und Ländervergleich äquivalent ist. Im Fokus der Untersuchung steht der Stand der naturwissenschaftlichen Kompetenz am Ende der Sekundarstufe I. Diese Altersstufe ist durch den Übergang in die berufliche Ausbildung bzw. in die Oberstufe gekennzeichnet und stellt somit einen wichtigen Schnitt in der Bildungskarriere der Schülerinnen und Schüler dar.

Sollte sich der NEPS-Naturwissenschaftstest als anschlussfähig erweisen, werden im nächsten Schritt die NEPS-Testwerte mit den PISA- und LV-Skalen verknüpft. Die Ergebnisse dieser Verlinkung könnten die Vergleichbarkeit der naturwissenschaftlichen Messung am Ende der Sekundarstufe I in Deutschland auf nationaler und internationaler

Ebene sicherstellen, neue Erkenntnisse hinsichtlich der Validität des NEPS-Tests liefern und nicht zuletzt die Interpretationsmöglichkeiten der Testwerte aller drei Studien erweitern.

1.2.5. Vergleich von Testinstrumenten

Die Anschlussfähigkeit der Testinstrumente und deren Verlinkung setzen voraus, dass die zu vergleichenden Studien ähnliche Inhalte auf eine ähnliche Art und Weise messen. Dies kann in Anlehnung an den Ansatz von Kolen und Brennan (2004) anhand der folgenden vier Aspekte untersucht werden:

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

In Abhängigkeit davon, inwiefern die Testinstrumente zweier Studien hinsichtlich der genannten Kriterien äquivalent sind, können verschiedene Methoden zur Verlinkung ihrer Testwerte angewendet werden.

1.2.6. Linking-Methoden

Mislevy (1992) und Linn (1993) unterscheiden fünf Arten des Linkings: *Moderation, Projection, Concordance, Vertical Scaling* und *Equating*. In Tabelle 1 wird die Abhängigkeit der Linking-Methoden vom Ähnlichkeitsgrad der zu vergleichenden Testinstrumente veranschaulicht. Außerdem zeigt Tabelle 1 das Linking-Kontinuum (Ryan & Brockmann, 2009), das die Stärke des Linkings für die oben genannten Methoden wiedergibt. Das Linking-Kontinuum kann in folgender Hinsicht interpretiert werden: je stärker das Linking ist, desto enger sind die verlinkten Testwerte miteinander verknüpft.

Tab 1. Abhängigkeit der Linking-Methoden vom Ähnlichkeitsgrad der Testinstrumente

	Linking-Methode	Schluss-folgerungen	Ziel-populationen	Mess-bedingungen	Konstrukte
starkes Linking	Equating	gleich	gleich	gleich	gleich
	Vertical Scaling	gleich	ungleich	gleich/ähnlich	gleich/ ähnlich
	Concordance	gleich	gleich/ähnlich	(un)gleich	gleich
	Projection	(un)gleich	gleich	ungleich	(un)gleich
schwaches Linking	Moderation	(un)gleich	(un)gleich	ungleich	(un)gleich

Demnach ist die Moderation die schwächste Form des Linkings und kann verwendet werden, wenn die Tests in Bezug auf ihre Schlussfolgerungen, Konstrukte, Populations- und Messmerkmale unterschiedlich sind. Diese Art des Linkings kommt oft bei Tests mit verschiedenen Rahmenkonzeptionen aber ähnlichen Konstrukten zum Einsatz (vgl. National Center for Educational Statistics, 2013; Phillips, 2007). Die Werte der durch die Moderation-Methode verknüpften Tests haben unterschiedliche Bedeutungen und können nur im Hinblick auf ihre Mittelwerte verglichen werden (wie die z-Werte zweier Tests).

Für die Verknüpfung mit der Projektion müssen die Tests gleiche Zielpopulationen haben, können sich aber in Schlussfolgerungen, Konstrukten und Messbedingungen unterscheiden (vgl. Wu, 2010). Werden die Tests A und B mit der Projektion-Methode verlinkt, können zwar die Werte des Tests A aus den Werten des Tests B vorhergesagt werden, aber nicht umgekehrt.

Die Concordance befindet sich im mittleren Bereich des Linking-Kontinuums (Tab. 1) und kann verwendet werden, um Tests mit gleichen Schlussfolgerungen und Konstrukten sowie ähnlichen Populationen, aber unterschiedlichen Messbedingungen zu verbinden. Am besten eignet sich diese Methode, um die Schwellenwerte der Kompetenzstufen verschiedener Tests miteinander zu vergleichen (vgl. Dorans, Lyu, Pommerich & Houston, 1997).

Für die Verknüpfung anhand des Vertical Scaling müssen die Tests ähnliche Schlussfolgerungen, Konstrukte und Messbedingungen aufweisen, können sich aber in ihren Populationen unterscheiden. Zur Verknüpfung zweier Tests mit dem Vertical Scaling können IRT-Methoden angewendet werden (vgl. Holland & Dorans, 2006; Nissen, Ehmke, Köller & Duchhardt, 2015; Pietsch, Böhme, Robitzsch & Stubbe, 2009).

Die stärkste Linking-Methode ist das Equating. Diese Methode wird oft verwendet, wenn parallele Tests existieren (vgl. Cartwright, 2012; Cartwright, Lalancette, Mussio & Xing, 2003; van den Ham, Ehmke, Nissen, & Roppelt, 2016; Nissen et al., 2015). Für die

Verknüpfung mit Hilfe des Equatings müssen die Tests in Bezug auf alle Merkmale des Ansatzes von Kolen und Brennan (2004) ähnlich sein. 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. Welche Studien mit dieser Methode bis jetzt verknüpft wurden und welche Schlussfolgerungen aus diesen Verlinkungen abgeleitet worden sind, wird im nächsten Abschnitt exemplarisch dargestellt.

1.2.7. Linking-Studien

Seit PISA 2000 gibt es zahlreiche Bestrebungen, nationale und internationale Assessments zu verlinken (Cartwright, 2012; Cartwright et al., 2003; Hambleton, Sireci & Smith, 2009; National Center for Educational Statistics, 2013; Nissen et al., 2015; Pietsch et al., 2009). An dieser Stelle werden drei dieser Studien exemplarisch vorgestellt, weil sie aufgrund der jeweils verlinkten Studien (PISA, NEPS, LV) und der angewendeten Methode (Equipercentile Equating) für die vorliegende Arbeit besonders relevant sind.

In der Studie von Cartwright et al. (2003) wurden die Ergebnisse von fünfzehnjährigen Schülerinnen und Schülern aus Kanada im nationalen Test für Leseverständnis der Foundation Skills Assessment (FSA) über ein Equipercentile Equating mit den Ergebnissen von PISA 2000 verlinkt. Der Vergleich der Standards in diesen Studien auf Grundlage der verlinkten FSA- und PISA-Werte zeigte, dass der Schwellenwert der höchsten FSA-Kompetenzstufe über dem Schwellenwert der höchsten PISA-Kompetenzstufe liegt. Würden also beide Tests mit dem Ziel eingesetzt, die besten fünfzehnjährigen Leserinnen und Leser in Kanada zu identifizieren, so würde ihre Anzahl im PISA-Test höher ausfallen als auf Grundlage des FSA-Tests. Der Vergleich auf den mittleren Kompetenzstufen führte zu ähnlichen Ergebnissen: Schülerinnen und Schüler, deren Lesekompetenz nach FSA dem Regelstandard entspricht (meeting expectations), wurden in PISA teilweise auf Kompetenzstufe 5 verortet. Umgekehrt erreichen Schülerinnen und Schüler, deren Leistung im FSA-Test den Regelstandard verfehlt (not meeting expectations), im PISA-Test die Kompetenzstufe 2 und werden somit in ihrer Leistung als ausreichend bewertet.

In einer anderen Studie zur Verlinkung eines internationalen Assessments mit einem nationalen Testinstrument wurden keine Kompetenzstandards der Studien, sondern die Linking-Methoden hinsichtlich ihrer Klassifizierungskonsistenz verglichen. In dieser Studie haben Nissen et al. (2015) das IRT-Linking dem Equipercentile Equating gegenüber gestellt, indem sie die Mathematik-Skala des TIMSS-Tests für die vierte Klasse auf den Mathematiktest des NEPS für die fünfte Klasse übertragen haben. Die Verlinkung der beiden

Skalen an 733 Viertklässlerinnen und Viertklässlern zeigte eine zufriedenstellende Klassifizierungskonsistenz gegenüber den internationalen TIMSS-Kompetenzstufen. Darüber hinaus stellten die Autoren 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. Aufgrund der geringfügigen Unterschiede in der Klassifizierung zu den Kompetenzstufen empfehlen die Autoren, Schlussfolgerungen lediglich über Gruppen von Schülerinnen und Schülern und nicht über Einzelpersonen zu ziehen.

Neben der Verlinkung nationaler und internationaler Testinstrumente gewinnt die Verlinkung nationaler Tests untereinander in der empirischen Bildungsforschung immer mehr an Bedeutung. So wurden in der Studie von van den Ham et al. (2016) die Ergebnisse des NEPS-Mathematiktests für die neunte Klassenstufe auf die Kompetenzstufen des LV übertragen. Das Linking erfolgte via Equipercentile Equating und zeigte für die Gesamtpopulation vergleichbare Verteilungen der Schülerinnen und Schüler auf die Kompetenzstufen des LV. Jedoch legt die gemeinsame Skalierung der Testwerte nahe, dass beide Skalen trotz des hohen latenten Zusammenhangs von $r = 0.92$ nicht ohne weiteres austauschbar sind. Die prozentuale Übereinstimmung der Studien hinsichtlich der Zuordnung zu den Kompetenzstufen (PÜ = Prozentsatz der Schüler, die derselben Kompetenzstufe zugeordnet sind) lag auf der Individualebene bei $PÜ = 48\%$ und wies auf beträchtliche Unterschiede hin. Auch Cohens Kappa lag nur bei $k = 0.31$. Aus diesem Grund empfehlen die Autoren lediglich die Schlussfolgerungen auf der Populationsebene.

Die Darstellung der ausgewählten Studien zeigt, dass sowohl die Verlinkung nationaler und internationaler Assessments als auch die Verlinkung nationalen Tests untereinander gewinnbringend und von Interesse für die Testanwender sowie die Scientific Community sein kann. Dabei liegt der Fokus der Verlinkung in den meisten Studien auf der Lese- und Mathekompetenz (vgl. Ehmke, Köller, Nissen & van den Ham, 2014). Lediglich in den USA fand zusätzlich zu diesen Domänen die Verlinkung in den Naturwissenschaften statt (National Center for Education Statistics, 2013; Phillips, 2007). Im deutschen Sprachraum wurde eine solche Verknüpfung der naturwissenschaftlichen Skalen bis jetzt noch nicht durchgeführt. An dieser Lücke setzt die vorliegende Arbeit an und untersucht die Möglichkeit der Übertragung der naturwissenschaftlichen Skalen von PISA und dem LV auf die Naturwissenschaftswerte im NEPS.

Wie im Abschnitt zu den Linking-Methoden bereits ausgeführt, setzt ein starkes Linking eine hohe Vergleichbarkeit der Testinstrumente hinsichtlich ihrer

Schlussfolgerungen, Zielpopulationen, Messmerkmale und Testkonstrukte voraus. Der Aspekt der Konstrukt-Äquivalenz kann in Anlehnung an den kulturvergleichenden Ansatz von van de Vijver (1998), der im kommenden Abschnitt dargestellt wird, weiter ausdifferenziert werden.

1.2.8. Konstrukt-Äquivalenz

Eine Gegenüberstellung von Ergebnissen zweier oder mehrerer Studien hinsichtlich eines zu messenden Konstrukts setzt voraus, dass zwischen latentem, also nicht direkt beobachtbarem Konstrukt, den verwendeten Testmodellen (z.B. ein- vs. mehrparametrische IRT-Modelle) und den eingesetzten Testitems gleiche Beziehungen bestehen (van de Vijver, 1998). Dies wird erreicht, wenn neben der konzeptionellen Äquivalenz auch die dimensionale Äquivalenz und die Skalenäquivalenz vorliegen. Inwieweit dies für die naturwissenschaftsbezogenen Kompetenzmessungen in NEPS 2010, dem LV 2012 und PISA 2012 zutrifft, kann anhand folgender Kriterien beurteilt werden:

- *Konzeptionelle Äquivalenz*: Inwieweit ist die theoretische Rahmenkonzeption des NEPS für die Domäne Naturwissenschaften vergleichbar mit der Rahmenkonzeption der naturwissenschaftlichen Grundbildung von PISA und der Rahmenkonzeption des LV in den Fächern Biologie, Chemie und Physik?

- *Dimensionale Äquivalenz*: Wie hoch ist der Zusammenhang zwischen den Naturwissenschaftsskalen des NEPS, des LV und des PISA? Inwiefern messen diese Naturwissenschaftsskalen dasselbe Konstrukt?

- *Skalenäquivalenz*: Inwiefern zeigt sich die Verteilung der NEPS-Naturwissenschaftswerte äquivalent mit den Testwertverteilungen der entsprechenden Skalen des LV und des PISA?

Die Überprüfung der Konstrukt-Äquivalenz der naturwissenschaftlichen Messung im NEPS bildet mit den entsprechenden Kompetenzmessungen in PISA und dem LV das Kernstück dieser Dissertation. Sie teilt sich in drei empirische Studien, die den im Anschluss dargestellten Fragestellungen folgen.

1.3. Fragestellungen dieser Arbeit

- **F1**: Inwieweit ist die theoretische Rahmenkonzeption des NEPS für die Domäne Naturwissenschaften vergleichbar mit der Rahmenkonzeption der naturwissenschaftlichen Grundbildung von PISA und der Rahmenkonzeption des LV in den Fächern Biologie, Chemie und Physik?

Diese Fragestellung wurde in der ersten Studie „Konzeptionelle Äquivalenz von Kompetenzmessungen in den Naturwissenschaften zwischen NEPS, IQB-Ländervergleich und PISA“ untersucht. Die Analyse der konzeptionellen Äquivalenz in dieser Studie basiert auf den Urteilen von sieben Expertinnen und Experten, die die Items des NEPS-Tests den verschiedenen Teilbereichen der Rahmenkonzeptionen von PISA und den Bildungsstandards zugeordnet haben.

Die Ergebnisse der Studie zeigen, dass der NEPS-Test zwar mit den Rahmenkonzeptionen von PISA und dem LV kompatibel ist, diese jedoch nicht vollständig abdecken kann. Dies bedeutet, dass das NEPS-Testinstrument trotz der hohen Überschneidung mit den PISA- und Bildungsstandards-Rahmenkonzeptionen die naturwissenschaftliche Kompetenz nicht in der gleichen konzeptionellen Breite erfasst, wie die PISA- und LV-Tests. Die Überschneidung der Studien in Kernbereichen der naturwissenschaftlichen Kompetenz kann jedoch als substantiell bezeichnet werden und liefert die Basis für die nachfolgenden Schritte in der Untersuchung der Konstrukt-Äquivalenz des NEPS-Tests mit den PISA- und LV-Tests.

Die Untersuchung der konzeptionellen Äquivalenz des NEPS-Tests mit den PISA- und LV-Testinstrumenten liefert Einblicke in die Gemeinsamkeiten und Unterschiede dieser Tests auf der Ebene ihrer Rahmenkonzeptionen. Sie stellt den ersten Schritt in der Untersuchung der Konstrukt-Äquivalenz des NEPS-Tests dar, die als eine der Voraussetzungen für die Übertragung der PISA- und LV-Skalen auf die NEPS-Testwerte gilt. Zur vollständigen Überprüfung der Konstrukt-Äquivalenz müssen die dimensionale Äquivalenz sowie die skalenbezogene Äquivalenz der Tests untersucht werden. Diese Untersuchung folgt in den Studien zwei und drei dieser Dissertation.

- **F2**: Messen die Naturwissenschaftsskalen des NEPS und des PISA dasselbe Konstrukt und inwiefern zeigt sich die Verteilung der NEPS-Naturwissenschaftswerte äquivalent zur Testwertverteilung in PISA?

Diese Fragestellung wurde in der zweiten Studie „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“ an einer Stichprobe von 1.528 Schülerinnen und Schülern der neunten Klassenstufe untersucht, die die Aufgaben aus NEPS und PISA bearbeitet haben.

Die Analyse der dimensionalen Äquivalenz des NEPS- und des PISA-Tests zeigt einen hohen Zusammenhang zwischen diesen Testdimensionen sowie eine hohe Tendenz der Testwerte zur Eindimensionalität. Ähnlich hoch ist die Vergleichbarkeit des NEPS- und des PISA-Tests hinsichtlich der Leistungsbewertung der getesteten Schülerinnen und Schülern. Somit belegen die Ergebnisse der zweiten Studie die Vergleichbarkeit der Konstrukte in den Testinstrumenten von NEPS und PISA sowie die Vergleichbarkeit ihrer Skalen. Weiterhin zeigen sich die Ergebnisse der Konstrukt-Äquivalenz-Überprüfung beeinflussbar vom Umgang der Studien mit fehlenden Werten. Die Analysen legen nahe, dass sich die Vergleichbarkeit der Testinstrumente erhöht, wenn fehlende Werte in beiden Studien ignoriert werden.

Die Untersuchung der dimensionalen und der skalenbezogenen Äquivalenz der Testinstrumente in dieser Studie belegt die Vergleichbarkeit der naturwissenschaftlichen Messungen beider Studien und liefert somit die Basis für die Übertragung der naturwissenschaftlichen Skala von PISA auf die NEPS-Testwerte. Darüber hinaus unterstreicht diese Studie die Rolle des Umgangs mit fehlenden Werten auf die Vergleichbarkeit der Testwerte in NEPS und PISA. Da dieser Effekt bislang noch nicht untersucht wurde, sind die Ergebnisse der zweiten Studie für die Bildungsforschung von besonderer Relevanz.

- **F3**: Messen die Naturwissenschaftsskalen des NEPS und des LV dasselbe Konstrukt und inwiefern zeigt sich die Verteilung der NEPS-Naturwissenschaftswerte äquivalent zur Testwertverteilung in den LV-Tests?

Die Überprüfung dieser Fragestellung erfolgte in der dritten Studie „Vergleichbarkeit der naturwissenschaftlichen Kompetenz in der neunten Klasse im Nationalen Bildungspanel und im IQB-Ländervergleich 2012“, in der 678 Schülerinnen und Schüler die Aufgaben aus NEPS und dem LV bearbeitet haben.

Die Untersuchung der dimensionalen Äquivalenz in der dritten Studie zeigt einen hohen latenten Zusammenhang zwischen den NEPS- und LV-Testwerten, dessen Höhe nah an den Zusammenhängen der einzelnen LV-Tests untereinander liegt. Trotz des hohen

Zusammenhangs deuten die Ergebnisse der Faktorenanalyse auf die dimensionale Trennung des NEPS-Tests von den LV-Tests hin. Beide Befunde können als eine Einschränkung in der Vergleichbarkeit der Ergebnisse aus NEPS und dem LV gewertet werden. Dagegen zeigen sich die NEPS- und LV-Tests vergleichbar hinsichtlich ihrer Leistungsbewertung der Schülerinnen und Schülern. Somit legen die Ergebnisse der dritten Studie nahe, dass die NEPS- und LV-Tests in Kernbereichen der naturwissenschaftlichen Kompetenz miteinander vergleichbar sind.

Die Untersuchung der dimensionalen Äquivalenz und der skalenbezogenen Äquivalenz in der dritten Studie spricht für die Vergleichbarkeit der NEPS- und LV-Testinstrumente in Kernbereichen der naturwissenschaftlichen Kompetenz sowie die Vergleichbarkeit ihrer Skalen. Beide Befunde schaffen somit die Basis für die Verlinkung der NEPS- und LV-Skalen, die die Interpretation der jeweiligen Studienergebnisse erweitern könnten.

In den folgenden drei Kapiteln werden die empirischen Studien, die für die Untersuchung der in diesem Abschnitt formulierten Fragestellungen durchgeführt wurden, im Detail vorgestellt.

1.4. Literatur

American Association for the Advancement of Science (2009). *Benchmarks for science literacy. Project 2061*. New York: Oxford University Press.

Baumert, J., Lehmann, R., Lehrke, M., Schmitz, B., Clausen, M., Hosenfeld, I., . . . Neubrand, J. (1997). *TIMSS – Mathematisch-naturwissenschaftlicher Unterricht im internationalen Vergleich*. Deskriptive Befunde. Opladen: Leske + Budrich.

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.

Blossfeld, H.-P., Schneider T. & Doll, J. (2009). Die Längsschnittstudie Nationales Bildungspanel: Notwendigkeit, Grundzüge und Analysepotential. *Pädagogische Rundschau*, *63*, 249–259.

Bybee, R. W. (1997). Towards an understanding of scientific literacy. In W. Gräber & C. Bolte (Hrsg.), *Scientific literacy – An international symposium.* Kiel, 37–68.

Cartwright, F. (2012). *Linking the British Columbia English examination to the OECD combined reading scale*. Prepared for the British Columbia Ministry of Education.

Cartwright, F., Lalancette, D., Mussio, J. & Xing, D. (2003). *Linking provincial student assessments with national and international assessments.* Education, skills and learning, research papers, Bd. 005.Ottawa: Statistics Canada.

Dorans, N. J., Lyu, C. F., Pommerich, M. & Houston, W.M. (1997). Concordance Between ACT Assessment and Recentered SAT I Sum Scores. *College and University*, *73* (2), 24-32.

Ehmke, T., Köller, O., Nissen, A. & van den Ham, A-K. (2014). Äquivalenz von Kompetenzmessungen in Schulleistungsstudien. *Unterrichtswissenschaft*, *42* (4), 290-300.

Hambleton, R. K., Sireci, S. G. & Smith, Z. R (2009). How Do Other Countries Measure Up to the Mathematics Achievement Levels on the National Assessment of Educational Progress? *Applied Measurement in Education*, *22*(4), 376-393.

Hanushek, E. A. & Wößmann, L. (2015). *The knowledge capital of nations: education and the economics of growth.* Cambridge, MA: MIT Press.

Hartig, J. & Frey, A. (2012). Validität des Tests zur Überprüfung des Erreichens der Bildungsstandards in Mathematik. Zusammenhänge mit den bei PISA gemessenen Kompetenzen und Varianz zwischen Schulen und Schulformen. *Diagnostica, 58,* 3-14.

Holland, P. W. & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Hrsg.), *Educational measurement*, 4. Aufl., S. 187-220. Westport, CT: Praeger.

Klieme, E. & Leutner, D. (2006). Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen. Beschreibung eines neu eingerichteten Schwerpunktprogramms der DFG. *Zeitschrift für Pädagogik, 52* (6), 876-903.

KMK (2005a) = Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland. (2005a). *Beschlüsse der Kultusministerkonferenz: Bildungsstandards im Fach Biologie für den Mittleren Schulabschluss.* Beschluss vom 16.12.2004. München: Luchterhand.

KMK (2005b) = Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland. (2005b). *Beschlüsse der Kultusministerkonferenz: Bildungsstandards im Fach Chemie für den Mittleren Schulabschluss.* Beschluss vom 16.12.2004. München: Luchterhand.

KMK (2005c) = Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland. (2005c). *Beschlüsse der Kultusministerkonferenz: Bildungsstandards im Fach Physik für den Mittleren Schulabschluss.* Beschluss vom 16.12.2004. München: Luchterhand.

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.

Mislevy, R. J. (1992). *Linking educational assessments: concepts, issues, methods, and prospects.* Princeton, NJ: ETS Policy Information Center.

National Center for Education Statistics (2013). *U.S. States in a Global Context: Results from the 2011 NAEP-TIMSS Linking Study.* Washington, DC: Institute of Education Sciences.

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 equipercentile and IRT methods. *Studies in Educational Evaluation, 47,* 58–67. DOI: 10.1016/j.stueduc.2015.07.003.

OECD (2006). *Assessing scientific, reading and mathematical literacy: A framework for PISA 2006.* Paris: OECD.

Phillips, G. W. (2007). *Expressing International Educational Achievement in Terms of US Performance Standards: Linking NAEP Achievement Levels to TIMSS*. Washington, DC: American Institutes for Research.

Pietsch, M. Böhme, K., Robitzsch, A. & T. C. Stubbe (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 und Basel: Belz Verlag.

Prenzel, M. (2000). Lernen über die Lebensspanne aus einer domänenspezifischen Perspektive: Naturwissenschaften als Beispiel. In F. Achtenhagen & W. Lempert (Hrsg.), *Lebenslanges Lernen im Beruf - seine Grundlegung im Kindes- und*

Jugendalter. Band IV. Formen und Inhalte von Lernprozessen (S. 175-192). Opladen: Leske + Budrich.

Prenzel, M., Rost, J., Senkbeil, M., Häußler, P. & Klopp, A. (2001). Naturwissenschaftliche Grundbildung: Testkonzeption und Ergebnisse. In J. Baumert, E. Klieme, M. Neubrand, M. Prenzel, U. Schiefele, W. Schneider, P. Stanat, K.-J. Tillmann & M. Weiß (Hrsg.), *PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (S. 191-248). Opladen: Leske + Budrich.

Prenzel, M., Artelt, C., Baumert, J., Blum, W., Hammann, M., Klieme, E. & Pekrun, R. (Hrsg.) (2007), *PISA 2006 – Die Ergebnisse der dritten internationalen Vergleichsstudie*. Münster: Waxmann.

Prenzel, M., Schöps, K., Rönnebeck, S., Senkbeil, M., Walter, O., Carstensen, C. H. & Hammann, M. (2007). Naturwissenschaftliche Kompetenz im internationalen Vergleich. In M. Prenzel, C. Artelt, J. Baumert, W. Blum, M. Hammann, E. Klieme & R. Pekrun (Hrsg.), *PISA 2006 – Die Ergebnisse der dritten internationalen Vergleichsstudie* (S. 63-105). Münster: Waxmann.

Rost, J., Prenzel, M., Carstensen, C.-H., Senkbeil, M. & Groß, K. (Hrsg.) (2004). *Naturwissenschaftliche Bildung in Deutschland. Methoden und Ergebnisse von PISA 2000*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Ryan, J. & Brockmann, F. (2009). *A practitioner's introduction to equating with primers on classical test theory and item response theory.* CCSSO: Washington, DC.

van den Ham, A.-K., Ehmke, T., Nissen, A. & Roppelt, A. (2016): Assessments verbinden, Interpretationen erweitern? *Zeitschrift für Erziehungswissenschaft. DOI:* 10.1007/s11618-016-0686-2.

van de Vijver, F.J.R. (1998). Towards a Theory of Bias and Equivalence. In J. Harkness (Hrsg.), *ZUMA-Nachrichten Spezial, 3*, 41-65. Mannheim: ZUMA.

Weinert, F. E. (2001). Concept of competence: A conceptual clarification. In S. Rychen, D. Salganik & L. Hersh (Hrsg.), *Defining and selecting key competencies* (S. 54–65). Seattle: Hogrefe & Huber.

Wu, M. (2010). *Comparing the Similarities and Differences of PISA 2003 and TIMSS.* OECD Education Working Papers, No. 32, Paris: OECD Publishing.

2

Studie I:

Konzeptionelle Äquivalenz von Kompetenzmessungen in den Naturwissenschaften zwischen NEPS, IQB-Ländervergleich und PISA*

---

*Erschienen in:

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.

2.1. Einleitung

Ziel der Gesamtstrategie zum Bildungsmonitoring in Deutschland ist es, Lernergebnisse systematisch und wissenschaftlich abgesichert festzustellen, mögliche Gründe für eventuell unbefriedigende Ergebnisse zu analysieren und daraus geeignete Reformmaßnahmen durch die Bildungspolitik und –verwaltung abzuleiten (KMK, 2006). Um die Kompetenzen der Schülerinnen und Schüler in Deutschland international verorten zu können, nimmt Deutschland regelmäßig an internationalen Schulleistungsstudien wie z.B. TIMSS (*Trends in International Mathematics and Science Study*) und PISA (*Programme for International Student Assessment)* teil. Für einen nationalen Vergleich der Qualitätsentwicklung in den Schulen der einzelnen Bundesländer wurden abschlussbezogene länderübergreifende Bildungsstandards formuliert, die sich an einem gemeinsam vereinbarten Maßstab ausrichten und regelmäßig überprüft werden.

Die von PISA, TIMSS und den Tests zur Überprüfung der Bildungsstandards gesetzten Ziele können effizient und zuverlässig in einem querschnittlichen Design realisiert werden. Solche sogenannten Querschnittsstudien ermöglichen eine Momentaufnahme des Leistungsstandes, lassen aber keine Aussagen über Leistungsentwicklungen zu. Zur Verwirklichung dieser Zielsetzung sind Untersuchungen im Längsschnittdesign erforderlich. Aus diesem Grund wurde 2009 das Nationale Bildungspanel (*National Educational Panel Study* – NEPS) ins Leben gerufen, dessen Ziel es ist, „zentrale Bildungsprozesse und –verläufe in Deutschland über die gesamte Lebensspanne zu beschreiben und zu analysieren“ (Blossfeld et al., 2009, S. 249). Ein weiteres Ziel besteht in der Untersuchung von Kompetenzentwicklungen. Hier werden unter anderem der wechselseitige Einfluss von Kompetenzentwicklungen und Entscheidungen an kritischen Übergängen der Bildungskarriere sowie Einflüsse des sozioökonomischen und kulturellen Hintergrundes und der Lerngelegenheiten auf die Kompetenzentwicklung in den Blick genommen (Blossfeld et al., 2009).

Eine wichtige Anforderung an das NEPS ist zum einen die Anschlussfähigkeit an nationale und internationale Large-Scale-Assessments in Deutschland (Blossfeld, 2008) und zum anderen die Möglichkeit einer gemeinsamen Interpretation der Ergebnisse dieser Studien. Um die Umsetzung dieser Forderung zu prüfen, wurde 2012 die im Folgenden beschriebene Validierungsstudie durchgeführt. Sie hatte zum Ziel, den NEPS-Naturwissenschaftstest für die neunte Klassenstufe mit den Testinstrumenten aus PISA 2012 und den länderübergreifenden Bildungsstandards für den Mittleren Schulabschluss in den Fächern Biologie, Physik und

Chemie zu verknüpfen. Auf diese Weise soll der NEPS-Test validiert, seine Anschlussfähigkeit geprüft und in einem internationalen Referenzmaßstab beziehungsweise in den nationalen Standards verortet werden.

2.2. Theoretischer Hintergrund

2.2.1. Zur Vergleichbarkeit von Studienergebnissen

Nachdem im ersten Beitrag des vorliegenden Themenheftes die wichtigsten Schritte für das methodische Vorgehen bei der Untersuchung der Vergleichbarkeit von Studien erläutert wurden, sollen sie an dieser Stelle nur kurz aufgegriffen werden.

Um Testinstrumente und empirische Befunde aus Erhebungen vergleichen zu können, müssen nach Kolen und Brennan (2004) vier Aspekte berücksichtigt werden (vgl. Pietsch et al., 2009):

- *Schlussfolgerungen*: Welche Erkenntnisse können aus den empirischen Erhebungen gezogen werden? Teilen die Tests die gleiche Messintention?
- *Merkmale und Umstände der Messung*: Inwieweit unterscheiden sich die verwendeten Aufgabenformate, die Durchführungsbedingungen oder die Testdauer?
- *Zielpopulationen*: In welchen Zielpopulationen werden die Testinstrumente eingesetzt?
- *Operationalisierte Konstrukte*: Welche Konstrukte werden durch die Testinstrumente erhoben? Inwieweit erfassen diese dieselben inhaltlichen Teilbereiche und kognitiven Prozesse?

Ob und inwieweit die ersten drei Aspekte für die Naturwissenschaftstests von NEPS, PISA und Bildungsstandards zutreffen, wird im nächsten Kapitel diskutiert. Der letzte Aspekt des Vergleichs wird in Anlehnung an van de Vijver (1998) weiter ausdifferenziert.

Eine Gegenüberstellung von Ergebnissen zweier oder mehrerer Studien hinsichtlich eines zu messenden Konstrukts setzt voraus, dass zwischen latentem, also nicht direkt beobachtbarem Konstrukt, den verwendeten Testmodellen (z.B. ein- vs. mehrparametrische IRT-Modelle) und den eingesetzten Testitems gleiche Beziehungen bestehen. Dies wird erreicht, wenn neben der konzeptionellen Äquivalenz auch die dimensionale Äquivalenz und die Skalenäquivalenz vorliegen. Inwieweit dies für die naturwissenschaftsbezogenen

Kompetenzmessungen in NEPS, PISA 2012 und in den länderübergreifenden Bildungsstandards zutrifft, kann anhand folgender Kriterien beurteilt werden:

- *Konzeptionelle Äquivalenz*: Inwieweit ist die theoretische Rahmenkonzeption des NEPS für die Domäne Naturwissenschaften vergleichbar mit der Rahmenkonzeption der naturwissenschaftlichen Grundbildung von PISA und der Überprüfung des Erreichens der Bildungsstandards in den Fächern Biologie, Chemie und Physik?
- *Dimensionale Äquivalenz*: Inwieweit ist die faktorielle Struktur des latenten Konstrukts „naturwissenschaftliche Kompetenz“ des NEPS vergleichbar mit der faktoriellen Struktur der naturwissenschaftlichen Grundbildung des PISA-Tests und der Tests zur Überprüfung des Erreichens der Bildungsstandards in den Fächern Biologie, Chemie und Physik?
- *Skalenäquivalenz*: Inwieweit führen die Kompetenzstufenmodelle in den drei Studien zu äquivalenten Personenklassifikationen?

In der vorliegenden Studie liegt der Schwerpunkt auf der Untersuchung der konzeptionellen Äquivalenz der Naturwissenschaftstests von NEPS, PISA und Bildungsstandards. Im Folgenden werden in diesem Zusammenhang zunächst die ersten drei Aspekte des Modells von Kolen und Brennan (2004) diskutiert.

2.2.2. Inhaltlicher Vergleich zwischen NEPS, PISA und den Bildungsstandards

Messintention

PISA dient in erster Linie dem Bildungsmonitoring und schafft die Möglichkeit, Vergleichsaussagen auf internationaler Ebene zu treffen. Die Studie orientiert sich bei der Untersuchung der Kompetenzen von Schülerinnen und Schülern an der Vorstellung einer Grundbildung (*Literacy*), die erforderlich ist, um sich als mündiger Bürger am politischen, wirtschaftlichen und gesellschaftlichen Leben zu beteiligen. Eine solche Grundbildung muss „die Voraussetzungen für ein weiterführendes Lernen“ schaffen (Prenzel, Drechsel, Carstensen, & Ramm, 2004, S. 17). Bedingt durch den internationalen Charakter und den Fokus der Studie auf Grundbildung, verzichtet PISA auf eine enge Orientierung an Lehrplänen.

Auch die Bildungsstandards haben eine Monitoringfunktion, allerdings findet das Monitoring auf nationaler Ebene statt. Zu diesem Zweck wurden die fachspezifischen Ziele des Unterrichts in Form von Kompetenzen und Regelstandards formuliert (vgl. Kauertz,

Fischer, Mayer, Sumfleth & Walpuski, 2010; KMK, 2005a, b, c). Im Unterschied zu PISA werden die zu testenden Kompetenzen in den Bildungsstandards fächerspezifisch erhoben.

Die NEPS-Rahmenkonzeption der naturwissenschaftlichen Kompetenz orientiert sich ebenso wie PISA am *Literacy*-Konzept (Weinert et al., 2011). Analog zu PISA steht die naturwissenschaftliche Grundbildung im Fokus der Untersuchung und nicht das fachspezifische Wissen.

Anders als bei PISA und den Bildungsstandards können im NEPS Bildungsprozesse in einem längsschnittlichen Design über die gesamte Lebensspanne der teilnehmenden Personen untersucht werden (Blossfeld et al., 2009). Der Schwerpunkt der Studie liegt also nicht im Beschreiben der Kompetenzstände der einzelnen (Bundes-) Länder, sondern im Verstehen des Wechselspiels zwischen Bildungsbiographien und den damit einhergehenden Bildungsentscheidungen einerseits und der Kompetenzentwicklung der Personen andererseits.

Zusammenfassend ist festzustellen, dass alle drei Studien (PISA, Bildungsstandards und NEPS) einen wichtigen Beitrag im Bereich des Bildungsmonitorings leisten, sich aber in ihrem Schwerpunkt und der Spezifizierung der naturwissenschaftlichen Kompetenz unterscheiden. Letztere bildet den Ausgangspunkt für die Rahmenkonzeption einer Studie und kann folglich Auswirkungen auf die Vergleichbarkeit der Testinstrumente haben. Daher würde man beispielsweise eine stärkere inhaltliche Nähe zwischen den Tests von NEPS und PISA vermuten.

Merkmale und Umstände der Messung

Der Prozess der Datenerhebung und –verarbeitung wird für die Studien NEPS, PISA und Bildungsstandards seit Jahren von demselben Erhebungsinstitut koordiniert und streng standardisiert umgesetzt. Aufgrund dieser starken Standardisierung kann davon ausgegangen werden, dass die Durchführungsbedingungen der Messungen für alle drei Studien sehr ähnlich sind.

Auch die Testformate der Studien ähneln einander sehr stark. In allen drei Tests hat die Mehrheit der Items ein geschlossenes Antwortformat (Kauertz & Fischer, 2013; Schiepe-Tiska, Schöps, Rönnebeck, Köller & Prenzel, 2013; Schöps & Saß, 2013). Der einzige Unterschied zu den Testformaten des NEPS besteht darin, dass es bei PISA und den Bildungsstandards zudem Aufgaben gibt, die eine frei formulierte Antwort erfordern. Alle Tests wurden 2012 als Papier- und Bleistift-Tests durchgeführt.

Weitere Unterschiede bestehen im Testdesign. PISA (Prenzel, Carstensen, Frey, Drechsel & Rönnebeck, 2007) und die Bildungsstandards (Siegle, Schroeders & Roppelt,

2013) arbeiten mit einem Multi-Matrix-Design, bei dem die Testpersonen nur eine Auswahl der Items bearbeiten. Im NEPS hingegen werden alle Aufgaben von jeder Person in der gleichen Reihenfolge bearbeitet.

Folglich gibt es Unterschiede zwischen den Studien in den Testformaten und im Testdesign, die eine Auswirkung auf die Vergleichbarkeit der Studienergebnisse haben könnten. Diese Frage bedarf einer empirischen Überprüfung und wird im zweiten Schritt der Äquivalenzuntersuchung angegangen.

Zielpopulation

Das Ziel von PISA besteht darin, die Kompetenz von fünfzehnjährigen Schülerinnen und Schülern zu untersuchen. Im deutschen Schulsystem besuchen die meisten Fünfzehnjährigen die 9. oder 10. Klasse (Prenzel et al., 2004). Zielpopulation der Bildungsstandards für den Mittleren Schulabschluss (Siegle et al., 2013) und des NEPS K9-Tests (von Maurice, Sixt & Blossfeld, 2011) sind Schülerinnen und Schüler der 9.Klasse.

Rahmenkonzeptionen der Studien und deren Überschneidung

Bei der Untersuchung der konzeptionellen Äquivalenz geht es um die Frage, ob die theoretische Rahmenkonzeption des NEPS für die Domäne Naturwissenschaften mit der PISA-Rahmenkonzeption der naturwissenschaftlichen Grundbildung und der Rahmenkonzeption zur Überprüfung des Erreichens der Bildungsstandards in den Fächern Biologie, Chemie und Physik vergleichbar ist. Im Folgenden werden die Rahmenkonzeptionen der Studien vorgestellt. Anschließend werden mögliche Überschneidungen zwischen den Inhalten der Rahmenkonzeptionen aufgezeigt und die Fragestellungen dieser Untersuchungen abgeleitet.

Die PISA-Rahmenkonzeption (siehe Abb.1) unterscheidet drei Teilkompetenzen: *naturwissenschaftliche Fragestellungen erkennen*, *naturwissenschaftliche Phänomene erklären* und *naturwissenschaftliche Evidenz nutzen*. Grundlage für diese Teilkompetenzen bilden das (objektbezogene) *naturwissenschaftliche Wissen* und (Meta-) *Wissen über die Naturwissenschaften*. Unter dem *naturwissenschaftlichen Wissen* werden vier Wissenssysteme subsummiert: *Physikalische Systeme, Lebende Systeme, Erd- und Weltraumsysteme* und *Technologische Systeme*. Der Bereich *Wissen über die Naturwissenschaften* untergliedert sich in zwei Aspekte: *naturwissenschaftliches Forschen* und *naturwissenschaftliche Erklärungen*. Eine weitere Komponente, auf der die Teilkompetenzen der naturwissenschaftlichen Grundbildung beruhen, sind die *motivationalen*

Orientierungen und *Einstellungen* einer Person. Die Untersuchung der naturwissenschaftlichen Kompetenz erfolgt bei PISA situations- bzw. kontextgebunden. Dazu werden in der Rahmenkonzeption folgende fünf Kontexte differenziert: Gesundheit, natürliche Ressourcen, Umwelt, Risiken/Gefahren und zuletzt Grenzen von Naturwissenschaften und Technik.

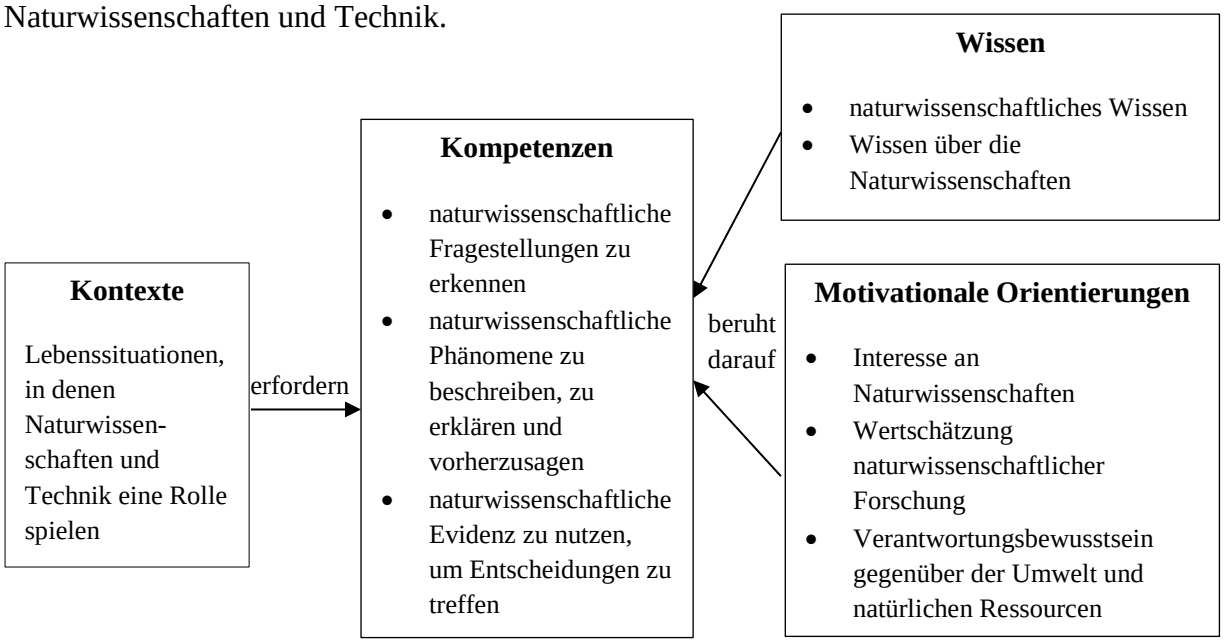

Abbildung 1: Die Rahmenkonzeption der naturwissenschaftlichen Grundbildung in PISA 2012 (Prenzel, Schöps et al., 2007)

Die Rahmenkonzeption der Bildungsstandards basiert auf dem Kompetenzstrukturmodell der naturwissenschaftlichen Fächer (s. Abbildung 2).

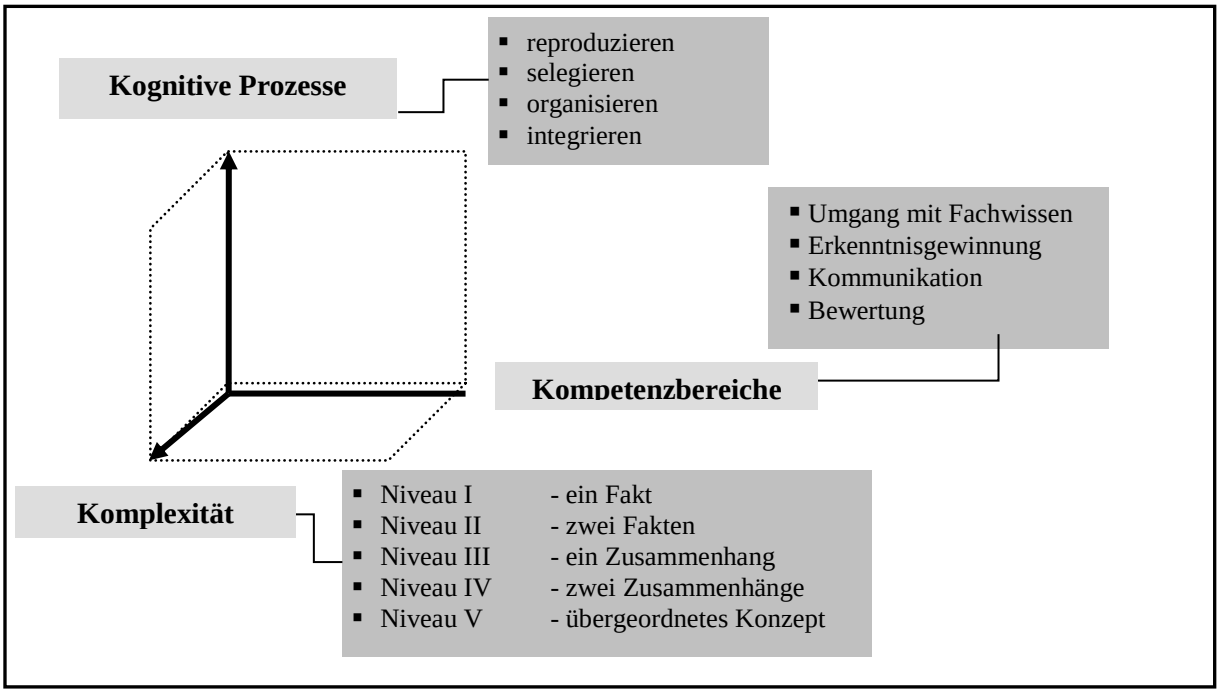

Abbildung 2: Dreidimensionales Kompetenzstrukturmodell der naturwissenschaftlichen Fächer in den länderübergreifenden Bildungsstandards (verändert nach Kauertz et al., 2010)

In diesem Modell wird die naturwissenschaftliche Kompetenz in vier Kompetenzbereiche unterteilt: *Umgang mit Fachwissen* (fachspezifisch für Biologie, Chemie und Physik), *Erkenntnisgewinnung*, *Kommunikation* und *Bewertung*. Weiterhin werden für jeden Kompetenzbereich fünf Komplexitätsstufen unterschieden: *ein Fakt*, *zwei Fakten*, *ein Zusammenhang*, *zwei Zusammenhänge* und *übergeordnetes Konzept*. Eine weitere Dimension zur Untersuchung der naturwissenschaftlichen Kompetenz stellen die kognitiven Prozesse dar, die in *Reproduzieren*, *Selegieren*, *Organisieren* und *Integrieren* unterteilt werden.

Die Rahmenkonzeption des NEPS (siehe Abb. 3) wurde in Anlehnung an die Rahmenkonzeptionen von PISA und den Bildungsstandards entwickelt. Die naturwissenschaftliche Kompetenz wird ähnlich wie in PISA in die objektbezogene Komponente (*naturwissenschaftliches Wissen*) und prozessbezogene Komponente (*Wissen über die Naturwissenschaften*) unterteilt. Das *naturwissenschaftliche Wissen* umfasst in der Rahmenkonzeption des NEPS die Konzepte *Stoffe*, *Systeme*, *Entwicklung* und *Wechselwirkungen*. Innerhalb des *Wissens über die Naturwissenschaften* werden die Konzepte *Messen und Messfehler* und *naturwissenschaftliches Denken* unterschieden. Die Erfassung der naturwissenschaftlichen Kompetenz erfolgt eingebettet in die ausgewählten Kontexte: *Gesundheit*, *Umwelt* und *Technologie*.

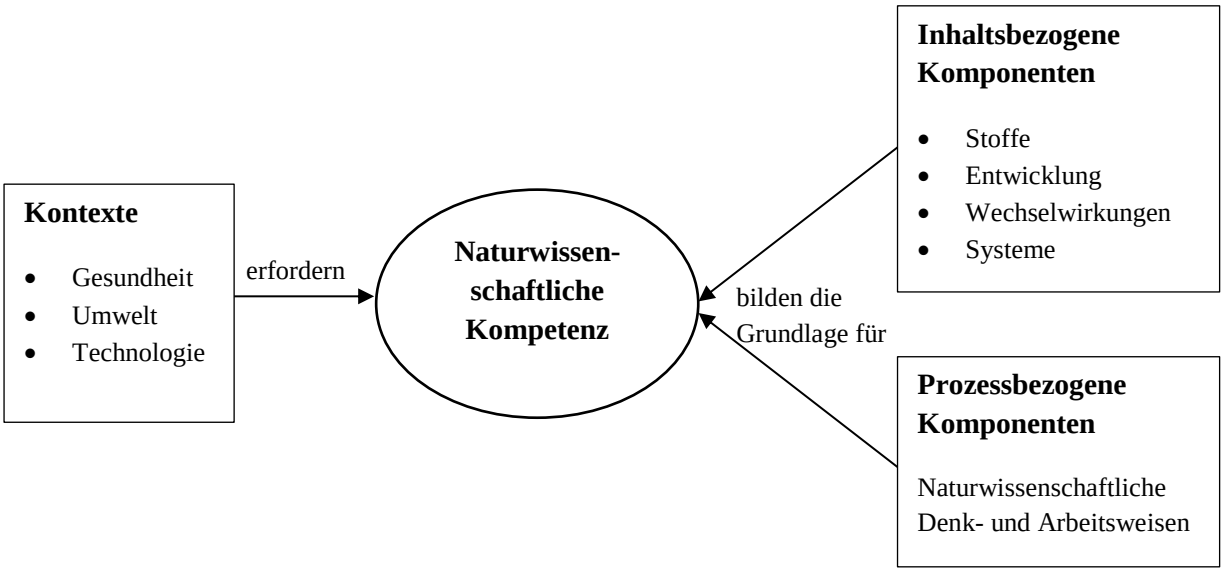

Abbildung 3: Rahmenkonzeption der naturwissenschaftlichen Grundbildung des NEPS (Hahn et al., 2013)

Die Rahmenkonzeption einer Studie dient nicht nur der Umschreibung des zu testenden Konstrukts, sie gibt auch die Anforderungen an die Konstruktion eines Tests vor. Zu den zentralen Kriterien für die Aufgabenkonstruktion zählt unter anderem die Anforderung, dass die Aufgaben „die in der Testkonzeption unterschiedenen inhaltlichen Aspekte umsetzen und repräsentieren“ sollen (Drechsel, Prenzel & Seidel, 2009, S. 366). Eine Möglichkeit dies zu

überprüfen, besteht in der Zuordnung der Items zu den einzelnen Teilbereichen einer Rahmenkonzeption durch ein Expertenpanel.

Bei der Überprüfung der konzeptionellen Äquivalenz geht es um die Frage, ob die Tests der zu vergleichenden Studien die gleichen Inhalte abfragen. Für jede der betrachteten Studien wurde die Passung der Items zu der jeweiligen Rahmenkonzeption in einem sehr aufwendigen Evaluationsprozess bereits überprüft (Prenzel, Carstensen et al., 2007; Pant et al., 2013; Hahn et al., 2013). Der Fokus dieser Studie liegt auf der Untersuchung der Übereinstimmungen zwischen den Inhalten der NEPS-Items und den Inhalten der Rahmenkonzeptionen von PISA und Bildungsstandards.

Die Abbildung 4 zeigt die inhaltliche Überschneidung der Rahmenkonzeption des NEPS mit der Rahmenkonzeption von PISA im Bereich des *naturwissenschaftlichen Wissens* und mit der Rahmenkonzeption der Bildungsstandards im Kompetenzbereich *Fachwissen*.

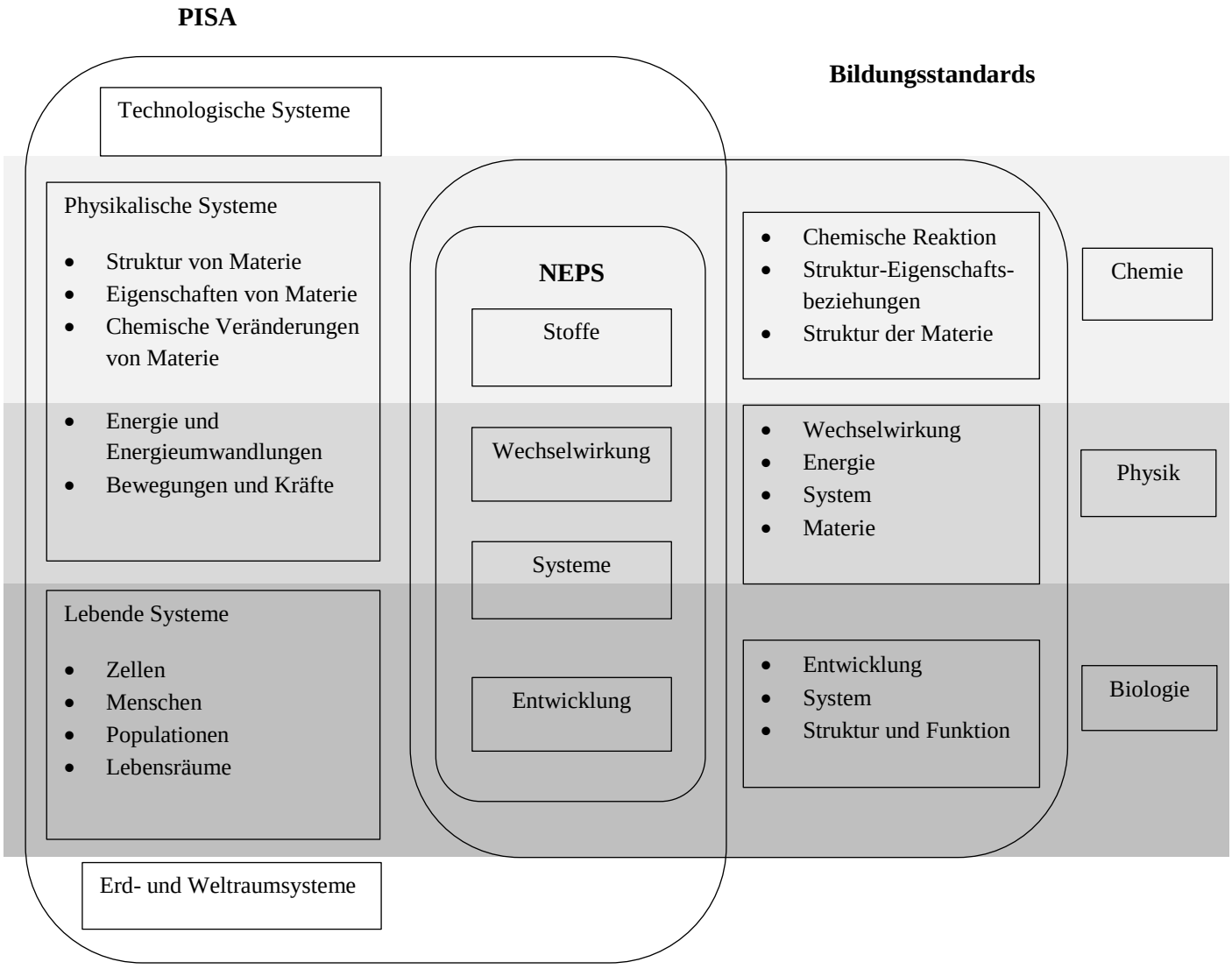

Abbildung 4: Überschneidung der Rahmenkonzeption des NEPS mit den Rahmenkonzeptionen von PISA und Bildungsstandards (verändert nach Hahn et al., 2013)

Demnach kann das NEPS-Konzept *Stoffe* in den *Umgang mit Fachwissen Chemie* (Bildungsstandards) und *Physikalische Systeme* (PISA) eingeordnet werden. Ein Äquivalent für das Konzept *Wechselwirkungen* im NEPS findet sich in der Rahmenkonzeption der Bildungsstandards im Kompetenzbereich *Umgang mit Fachwissen Physik* und in der Rahmenkonzeption von PISA im Wissenssystem *Physikalische Systeme*. In der Rahmenkonzeption des NEPS umfasst das Konzept *Systeme* biologische und technologische Systeme. Aus diesem Grund steht dieses Konzept an der Schnittstelle zwischen dem *Umgang mit Fachwissen Biologie* und -*Physik* in den Bildungsstandards und den *Physikalischen*- und *Lebenden Systemen* bei PISA. Schließlich kann auch eine Überschneidung des NEPS-Konzepts *Entwicklung* mit dem Kompetenzbereich *Umgang mit Fachwissen Biologie* (Bildungsstandards) einerseits und dem Wissenssystem *Lebende Systeme* (PISA) andererseits postuliert werden. Die Frage, ob die Überschneidung zwischen den Rahmenkonzeptionen der Studien in dieser Form auf der Testebene tatsächlich besteht, bedarf einer empirischen Überprüfung.

Die Beschreibung der Rahmenkonzeptionen verdeutlicht, dass die Studien eine unterschiedliche Anzahl von Komponenten zur Erfassung der naturwissenschaftlichen Kompetenz unterscheiden. Während die NEPS-Rahmenkonzeption „nur“ die Wissenskomponenten betrachtet, werden in der Rahmenkonzeption von PISA (zusätzlich zu den Wissensbereichen) Teilkompetenzen der naturwissenschaftlichen Grundbildung differenziert. Die Bildungsstandards-Rahmenkonzeption hat dagegen die höchste Anzahl von Komponenten: die naturwissenschaftliche Kompetenz wird hier anhand eines dreidimensionalen Kompetenzmodells beschrieben. Daher stellt sich die Frage, ob der Unterschied in der Anzahl der Komponenten auch einen Unterschied in der Breite der Rahmenkonzeptionen widerspiegelt.

2.2.3. Ableitung der Fragestellungen

Die konzeptionelle Äquivalenz des NEPS-Tests und der Tests von PISA und den Bildungsstandards wird im Rahmen der vorliegenden Arbeit anhand von folgenden Fragen untersucht:

1) Inwieweit besteht eine Übereinstimmung zwischen den Inhalten der NEPS-Items und den Inhalten der Rahmenkonzeptionen von PISA und den Bildungsstandards? (Passung der NEPS-Items zu den Inhalten der Rahmenkonzeptionen von PISA und den Bildungsstandards)

2) Inwieweit besteht eine Überschneidung zwischen den Konzepten der NEPS-Rahmenkonzeption und den Kompetenzbereichen der Bildungsstandards-Rahmenkonzeption bzw. den Wissensbereichen der Rahmenkonzeption von PISA? (inhaltliche Überschneidung der Rahmenkonzeptionen)

3) Inwieweit können Inhalte der Rahmenkonzeptionen von PISA und den Bildungsstandards durch die NEPS-Items abgedeckt werden? (Ähnlichkeit der konzeptionellen Breite der Rahmenkonzeptionen)

Da zum Zeitpunkt der Studie noch keine ausreichende Operationalisierung der Kompetenzbereiche *Kommunikation* und *Bewertung* in der Rahmenkonzeption der Bildungsstandards vorlag, werden diese Bereiche bei der Überprüfung der konzeptionellen Äquivalenz der Tests nicht berücksichtigt.

2.3. Methoden

2.3.1. Stichprobe und Design

Zur Überprüfung der konzeptionellen Äquivalenz konnten sieben Expertinnen und Experten gewonnen werden. Die Auswahl der Expertinnen und Experten wurde unter folgenden Gesichtspunkten vorgenommen:

- Da das Ziel der Studie darin bestand, NEPS-Items anhand ihrer naturwissenschaftlichen Inhalte einzuschätzen, wurde die Mehrheit der Expertinnen und Experten aus dem naturwissenschaftlich-fachdidaktischen Bereich ausgewählt. Daher kommen fünf von sieben Expertinnen und Experten aus dem Bereich der Fachdidaktik, eine Person aus dem Bereich der Erziehungswissenschaften und Psychologie und eine weitere Person aus dem Lehramt.
- Da die NEPS-Items auf der Rahmenkonzeption von PISA und Bildungsstandards eingeschätzt werden sollten, wurde bei der Auswahl der Expertinnen und Experten zusätzlich darauf geachtet, dass sie umfangreiche Kenntnisse der Rahmenkonzeption und des Naturwissenschaftstests einer der Vergleichsstudien haben. Vier Expertinnen und Experten waren durch vorherige Studien besonders vertraut mit PISA und drei von ihnen mit den Bildungsstandards.

Jede Expertin bzw. jeder Experte erhielt ein Paket mit einem vollständigen NEPS-Test (28 Items), der Beschreibung der Teilkompetenzen der naturwissenschaftlichen Grundbildung und der Aspekte des Wissens über die Naturwissenschaften bei PISA sowie der Beschreibung der

kognitiven Prozesse und der *Komplexität* bei den Bildungsstandards. Wegen des großen Umfangs der Beschreibung des *naturwissenschaftlichen Wissen*s in der PISA-Rahmenkonzeption und der Kompetenzbereiche *Umgang mit Fachwissen* und *Erkenntnisgewinnung* in der Rahmenkonzeption der Bildungsstandards wurde auf die entsprechende Literatur verwiesen. Expertenurteile wurden anhand von Review-Sheets erhoben, mit deren Hilfe die NEPS-Items den Inhalten der Rahmenkonzeptionen von PISA und der Bildungsstandards zugeordnet wurden. Ein Ausschnitt des Review-Sheets für die Dimension *Kompetenzbereiche* aus der Rahmenkonzeption der länderübergreifenden Bildungsstandards ist in Abbildung 5 dargestellt.

		Item 1…n
Das Item kann keinem der Kompetenzbereiche zugeordnet werden		
Umgang mit Fachwissen Chemie	Struktur der Materie	
	Struktur-Eigenschaftsbeziehungen	
	Chemische Reaktion	
	Das Item kann dem Kompetenzbereich zugeordnet werden, aber die Zuordnung zu einer der Kategorien des Kompetenzbereiches ist nicht möglich	
Umgang mit Fachwissen Physik	Materie	
	System	
	Energie	
	Wechselwirkung	
	Das Item kann dem Kompetenzbereich zugeordnet werden, aber die Zuordnung zu einer der Kategorien des Kompetenzbereiches ist nicht möglich	
Umgang mit Fachwissen Biologie	Struktur und Funktion	
	System	
	Entwicklung	
	Das Item kann dem Kompetenzbereich zugeordnet werden, aber die Zuordnung zu einer der Kategorien des Kompetenzbereiches ist nicht möglich	
Erkenntnis-gewinnung	Naturwissenschaftliche Untersuchungen	
	Naturwissenschaftliche Modell- und Theoriebildung	
	Wissenschaftstheoretische Reflexion	
	Das Item kann dem Kompetenzbereich zugeordnet werden, aber die Zuordnung zu einer der Kategorien des Kompetenzbereiches ist nicht möglich	

Abbildung 5: Ausschnitt des Review-Sheets für die Dimension „Kompetenzbereiche“ der Rahmenkonzeption der länderübergreifenden Bildungsstandards

Jedes Item durfte demnach lediglich einem Teilbereich pro Dimension zugewiesen werden. Bei fehlender Passung konnten die Items auch als „nicht zuzuordnen“ eingestuft werden. Weiterhin lag in der Studie ein vollständiges Design vor, d.h. jede Expertin und jeder Experte

schätzten alle NEPS-Items sowohl hinsichtlich der Rahmenkonzeption von PISA als auch hinsichtlich der Rahmenkonzeption der Bildungsstandards ein.

2.3.2. Generalisierbarkeitstheorie

Für eine Zusammenfassung von Expertenurteilen muss eine ausreichende Interrater-Reliabilität der erhobenen Daten gegeben sein. Nur so kann eine Aussage über die Wahrscheinlichkeit des Zustandekommens eines Ergebnisses gemacht werden. In diesem Fall über die Wahrscheinlichkeit, dass selbst bei einer anderen Stichprobe von Ratern (die nach den gleichen Kriterien ausgewählt werden), das gleiche Ergebnis zustande käme. Diese Voraussetzung wurde in der vorliegenden Arbeit mit der Generalisierbarkeitstheorie (Cronbach et al., 1972) überprüft.

Die Generalisierbarkeitstheorie unterscheidet zwischen der Facette der Differenzierung und der Facette der Generalisierung. Während sich erstere auf das eigentliche Untersuchungsobjekt bezieht, bildet die Facette der Generalisierung die Fehlerquellen der Messung ab. Das Ziel von Analysen, die auf der Generalisierbarkeitstheorie basieren, besteht darin, die Varianz der Differenzierung ($universe\ score\ \sigma^2(p)$) im Verhältnis zur Varianz der Generalisierung möglichst groß werden zu lassen (Hughes & Garrett, 1988; vgl. Eisend, 2007). Beide Facetten werden entsprechend des Untersuchungszwecks festgelegt und können daher von Studie zu Studie variieren.

Die Generalisierbarkeitstheorie bietet im Vergleich zu anderen Ansätzen der Reliabilitätsanalyse entscheidende Vorteile. Der wichtigste Vorteil besteht in der Analyse der Fehlervarianz. Die gesamte Varianz kann in der vorliegenden Studie in die Varianz der Items, die Varianz der Rater (systematischer Fehler) und die Varianz der Interaktion zwischen den Items und den Ratern (unsystematischer Fehler) zerlegt werden. Macht die Ratervarianz mit mehr als 10% einen großen Anteil der Gesamtvarianz aus (Li & Lautenschlager, 1997), spricht das für eine systematische Verzerrung der Urteile durch die Rater. Im Fall der vorliegenden Studie würde dies bedeuten, dass die Rater unterschiedliche Kriterien für die Zuordnung der Items genutzt haben.

Ein zweites wichtiges Kriterium für die Beurteilung der Interrater-Reliabilität stellt der Generalisierbarkeitskoeffizient dar. In Abhängigkeit davon, ob der absolute oder der relativen Fehler verwendet wird, ist es im Rahmen des G-Ansatzes möglich, zwei Arten des Generalisierbarkeitskoeffizienten zu berechnen (Eisend, 2007). Während der absolute Fehler ($\sigma^2 abs$) sowohl die systematischen als auch die unsystematischen Fehlervarianzen umfasst,

werden beim relativen Fehler ($\sigma^2 rel$) nur die unsystematischen Anteile der Fehlervarianz berücksichtigt.

Die Wahl des Fehlers und somit des Generalisierbarkeitskoeffizienten hängt von der Fragestellung ab. Wenn die Reliabilitätsanalyse für die Berechnung der internen Konsistenz einer Skala verwendet wird, bei der alle Personen immer wieder mit dem gleichen Testinstrument untersucht werden, spielt die Varianz der Items keine bedeutsame Rolle. In diesem Fall bietet sich der relative Generalisierbarkeitskoeffizient Eρ² als Maß der internen Konsistenz an:

$$\mathrm{E}\rho^2 = \frac{\sigma^2(p)}{(\sigma^2(p) + \sigma^2 rel)}$$

Interessiert man sich hingegen für die Interrater-Reliabilität der Daten, spielt der Haupteffekt der Rater eine entscheidende Rolle. In diesem Fall wird der absolute Fehler für die Berechnung des absoluten Generalisierbarkeitskoeffizienten Φ herangezogen (Bloch & Normann, 2011):

$$\Phi = \frac{\sigma^2\,(p)}{(\sigma^2(p) + \sigma^2 abs)}$$

Ähnlich wie die interne Konsistenz der Items kann auch die Konsistenz der Raterurteile untersucht werden. Anders als bei der Untersuchung der Interrater-Reliabilität wird die Facette der Differenzierung durch die Rater und die Facette der Generalisierung durch die Items abgebildet. In diesem Zuge ist es möglich, der Frage nachzugehen, inwiefern die Berücksichtigung der Ratereigenschaften (wie z.B. des beruflichen Hintergrundes der Rater) die Konsistenz der Raterurteile erhöht.

In der vorliegenden Studie wurde die Generalisierbarkeitstheorie zur Analyse der Interrater-Reliabilität angewendet. In diesem Fall stellten die Items die Facette der Differenzierung und die Rater die Facette der Generalisierung dar. Zur Beurteilung der Interrater-Reliabilität wurde die Varianzverteilung untersucht und der Generalisierbarkeitskoeffizient Φ berechnet.

Des Weiteren wurde für bestimmte Dimensionen der Bildungsstandards-Rahmenkonzeption die Konsistenz der Rater überprüft. Dazu wurde der relative Generalisierbarkeitskoeffizient Eρ² einmal für die Gesamtheit der Rater und in Abhängigkeit von ihrem beruflichen und fachlichen Hintergrund berechnet.

Die oben beschriebenen Anwendungsmöglichkeiten, und damit einhergehend die herkömmlichen Programme für die Anwendung des G-Ansatzes, wurden für Daten entwickelt, die mindestens Intervallskalenniveau aufweisen. Dies trifft aber nur für die zweite und dritte Dimension der Bildungsstandards-Rahmenkonzeption zu. Die komplette Rahmenkonzeption von PISA und die erste Dimension der Bildungsstandards können hingegen nur auf einer Nominalskala abgebildet werden. Aus diesem Grund wurden die Reliabilitätsanalysen für nominale Daten mithilfe des Verfahrens von Li und Lautenschlager (1999) durchgeführt. Die Auswertung der intervallskalierten Daten erfolgte mit dem Programm G-String IV (Bloch & Norman, 2011). Nachdem die Varianzanteile berechnet wurden, wurde deren Adjustierung in Anlehnung an Brennan (2001a) durchgeführt.

Ein zufriedenstellender Generalisierbarkeitskoeffizient und eine niedrige Varianz zwischen den Ratern dienen als Voraussetzung für die Generalisierbarkeit der Urteile. Der Generalisierbarkeitskoeffizient Φ darf als Cohens Kappa (k) (Cohen, 1960) interpretiert werden (Brennan, 2003). Landis und Koch (1977) schlagen für Cohens k folgende Interpretation vor: Werte zwischen 0.41 und 0.6 werden als moderate Übereinstimmung, und Werte zwischen 0.61 und 0.8 werden als substantielle Übereinstimmung bezeichnet. Im Rahmen der vorliegenden Arbeit wird ein Generalisierbarkeitskoeffizient ≥ 0.6 und ein Varianzanteil der Rater $< 10\%$ (Li & Lautenschlager, 1997) als hinreichende Interrater-Reliabilität interpretiert.

Der Generalisierbarkeitskoeffizient $E\rho^2$ kann unter Berücksichtigung des Designs der Daten (Rater x Items) als Intra-Klassen-Korrelation (Shrout & Fleiss, 1979) interpretiert werden (Brennan, 2003). Hier wird die Veränderung des Generalisierbarkeitskoeffizienten $E\rho^2$ in Abhängigkeit des fachlichen bzw. beruflichen Hintergrundes analysiert.

Erst nachdem die Interrater-Reliabilität der Daten als ausreichend bewertet wurde, konnte eine Betrachtung der Item-Zuordnung zu den Rahmenkonzeptionen der Studien PISA und Überprüfung der länderübergreifenden Bildungsstandards erfolgen. Dabei wurde ein Item nur dann als eindeutig zugeordnet betrachtet, wenn die Mehrheit der Expertinnen und Experten (mind. 4) zu einem übereinstimmenden Urteil kamen.

2.4. Ergebnisse

2.4.1. Generalisierbarkeit der Expertenurteile

Für die Zusammenfassung der Expertenurteile musste die Interrater-Reliabilität der Daten überprüft werden. Dies wurde in der vorliegenden Arbeit anhand der Varianzverteilung und des Generalisierbarkeitskoeffizienten bewertet. Abbildungen 6 und 7 zeigen die Verteilung der Varianz mit den dazugehörigen Generalisierbarkeitskoeffizienten Φ für die Einordnung der NEPS-Items in die Rahmenkonzeptionen der Bildungsstandards und von PISA.

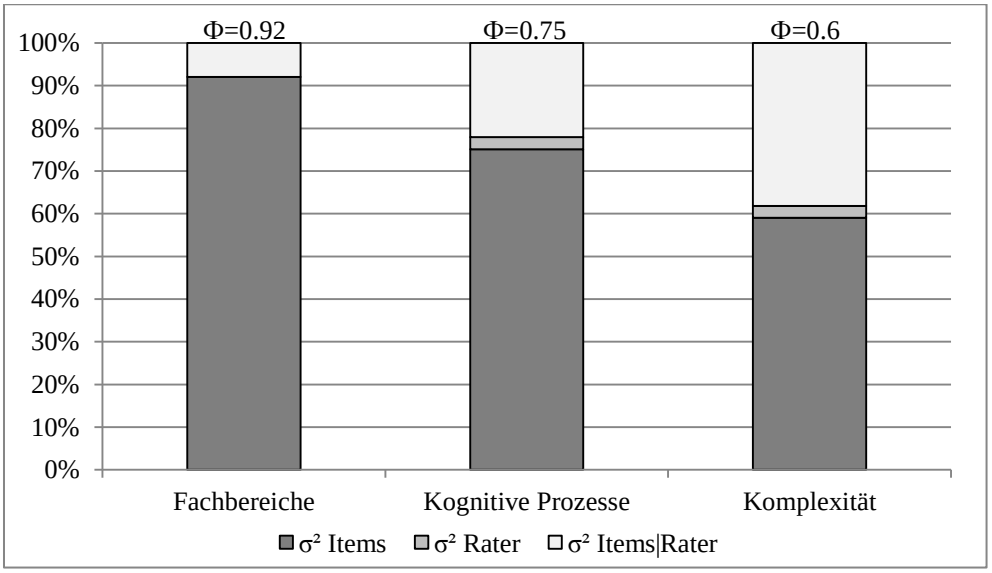

Abbildung 6: Varianzverteilungen und Generalisierbarkeitskoeffizienten für die Einordnung der NEPS-Items in die Rahmenkonzeption der Bildungsstandards

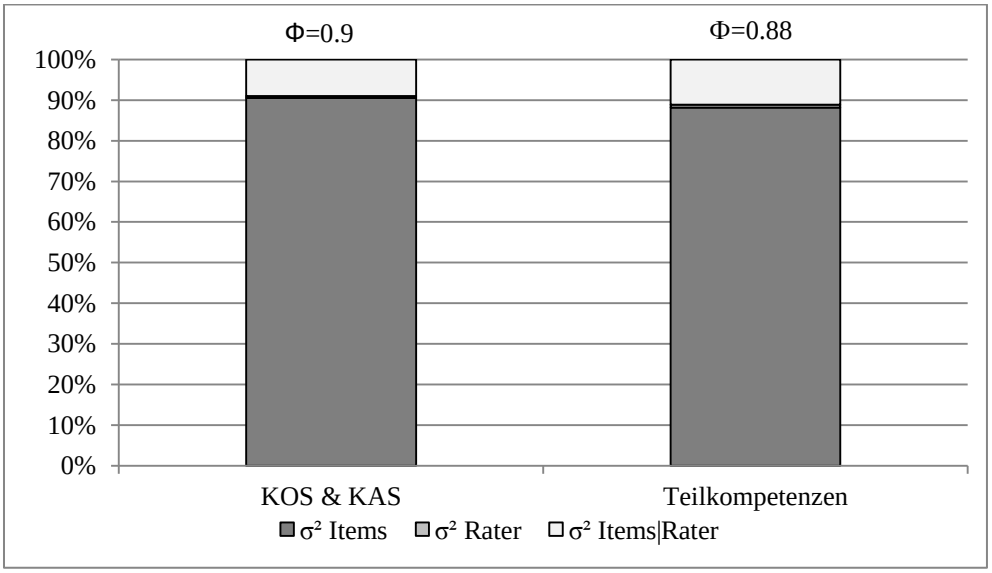

Abbildung 7: Varianzverteilung und Generalisierbarkeitskoeffizienten für die Einordnung der NEPS-Items in die PISA-Rahmenkonzeption. Anmerkungen: KOS: Knowledge of Science; KAS: Knowledge about Science.

Der Generalisierbarkeitskoeffizient Φ liegt in einem Bereich zwischen 0.6 und 0.92. Die Analyse der Varianzverteilung zeigt, dass der höchste Anteil der Varianz (59-91%) auf Itemebene zu finden ist und dass die Varianz zwischen den einzelnen Ratern gering (0-3%) ausfällt. Ferner wurde für die zweite und dritte Dimension der Bildungsstandards die Konsistenz der Raterurteile überprüft. Dazu wurde die Veränderung des relativen Generalisierbarkeitskoeffizienten Eρ² in Abhängigkeit vom beruflichen bzw. fachlichen Hintergrund der Rater betrachtet. Die Ergebnisse dieser Untersuchung sind in den Abbildungen 8 und 9 dargestellt.

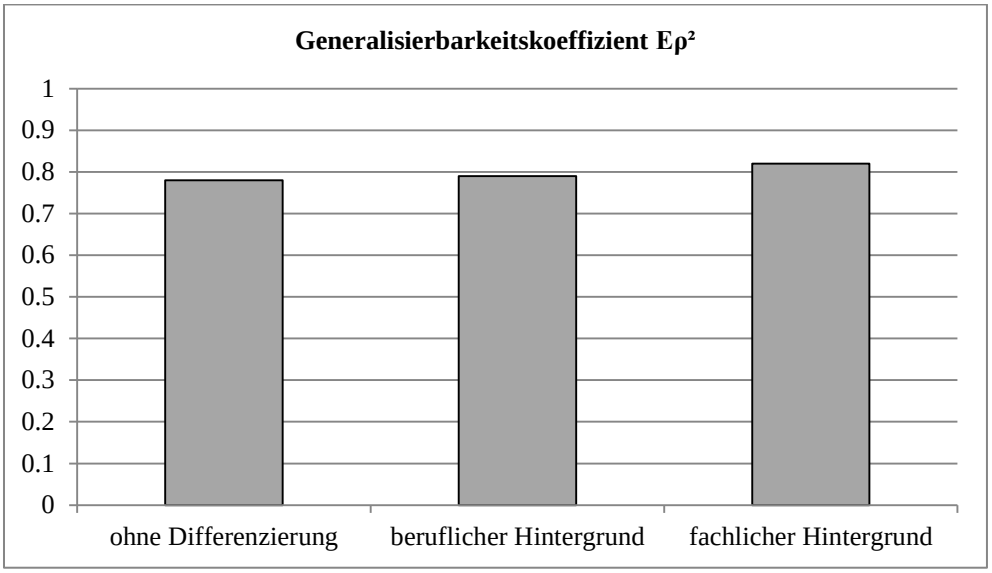

Abbildung 8: Konsistenz der Raterurteile in Abhängigkeit vom beruflichen und fachlichen Hintergrund der Rater für die Zuordnung der NEPS-Items zur Dimension *kognitive Prozesse*

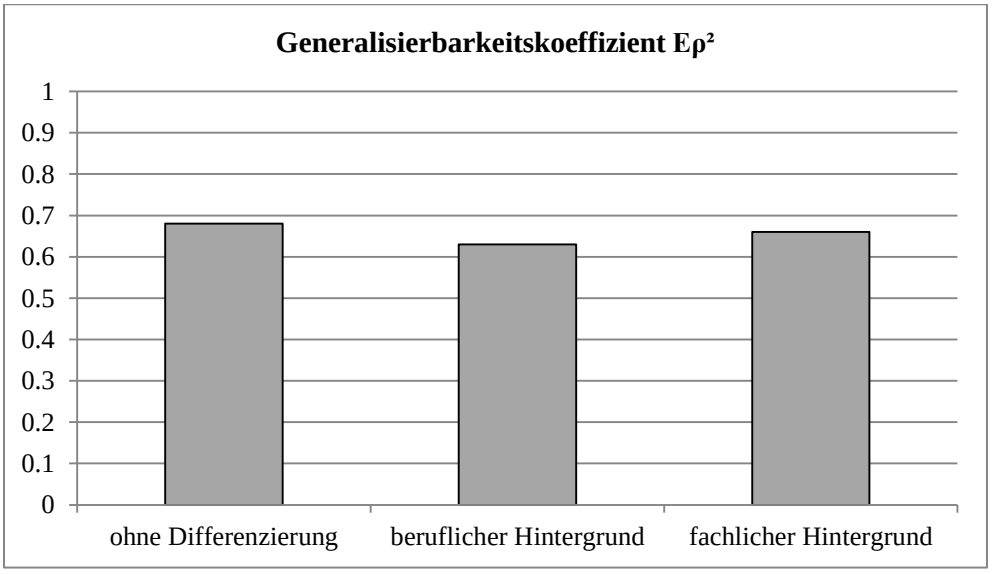

Abbildung 9: Konsistenz der Raterurteile in Abhängigkeit vom beruflichen und fachlichen Hintergrund der Rater für die Zuordnung der NEPS-Items zur Dimension *Komplexität*

Die Konsistenz der Rater für die Dimension *kognitive Prozesse* fällt insgesamt etwas höher aus als für die Dimension *Komplexität*, und die Berücksichtigung des beruflichen bzw. fachlichen Hintergrundes trägt nicht zu einer substantiellen Erhöhung der Rater-Konsistenz bei. Allerdings geht die Berücksichtigung des fachlichen Hintergrundes mit einem etwas höheren Generalisierbarkeitskoeffizienten einher als die des beruflichen Hintergrundes (Erfahrung, die die Expertinnen und Experten im Rahmen von PISA bzw. Bildungsstandards gesammelt haben).

2.4.2. Einordnung der NEPS-Testaufgaben in die Rahmenkonzeption der Bildungsstandards

Ein Großteil der NEPS-Items konnte in die Kompetenzbereiche der Rahmenkonzeption der Bildungsstandards eingeordnet werden. Vier und mehr Rater haben 26 von 28 NEPS Items (93%) einem bestimmten Teilbereich des *Umgangs mit Fachwissen* oder der *Erkenntnisgewinnung* zugeordnet. Die Items verteilen sich gleichmäßig auf die Teilbereiche mit einem etwas stärkeren Fokus auf die *Erkenntnisgewinnung* (35% der eingeordneten Items).

Bei der Zuordnung der NEPS-Items zu den *kognitiven Prozessen* zeigt sich, dass der NEPS-Test einen Schwerpunkt darauf legt, naturwissenschaftliches Wissen zu organisieren und zu integrieren (zusammen 90% der eingeordneten Items). Der Prozess *Selegieren* ist mit nur zwei NEPS-Items besetzt. Insgesamt liegt die Zuordnungsrate der NEPS-Items zur Dimension *kognitive Prozesse* bei 75%.

Bei der Dimension *Komplexität* der Bildungsstandards konnten insgesamt 64% der NEPS-Items einer bestimmten Komplexitätsstufe zugeordnet werden. Hier liegt ein sehr starker Fokus auf der Stufe *ein Zusammenhang* (72% der eingeordneten Items). Die Stufen *ein Fakt, zwei Zusammenhänge* und *übergeordnetes Konzept* weisen eine Besetzungsrate von einem bis zwei Items auf.

2.4.3. Einordnung der NEPS-Testaufgaben in die Rahmenkonzeption von PISA

Die Untersuchung der konzeptionellen Äquivalenz des NEPS-Tests bezüglich der Passung der Items zu den Wissensbereichen von PISA lässt eine insgesamt hohe Zuordnungsrate der NEPS-Items (79%) erkennen. Die Items verteilen sich relativ gleichmäßig über die Wissenssysteme *lebende Systeme*, *physikalische Systeme* und den Wissensbereich *Wissen über die Naturwissenschaften*.

Bei der Einordnung der NEPS-Items in die Teilkompetenzen der naturwissenschaftlichen Grundbildung wurden 96% der NEPS-Items einer bestimmten Teilkompetenz zugeordnet, wobei der Schwerpunkt der Zuordnung (67% der eingeordneten Items) auf der Kompetenz *naturwissenschaftliche Phänomene erklären* lag.

2.4.4. Inhaltliche Überschneidung der Rahmenkonzeptionen

Betrachtet man die Zuordnung der NEPS-Items zu den Inhalten der Rahmenkonzeptionen der Bildungsstandards und von PISA, zeigen sich Parallelen der drei Studien. Abbildung 10 verdeutlicht diese Befunde.

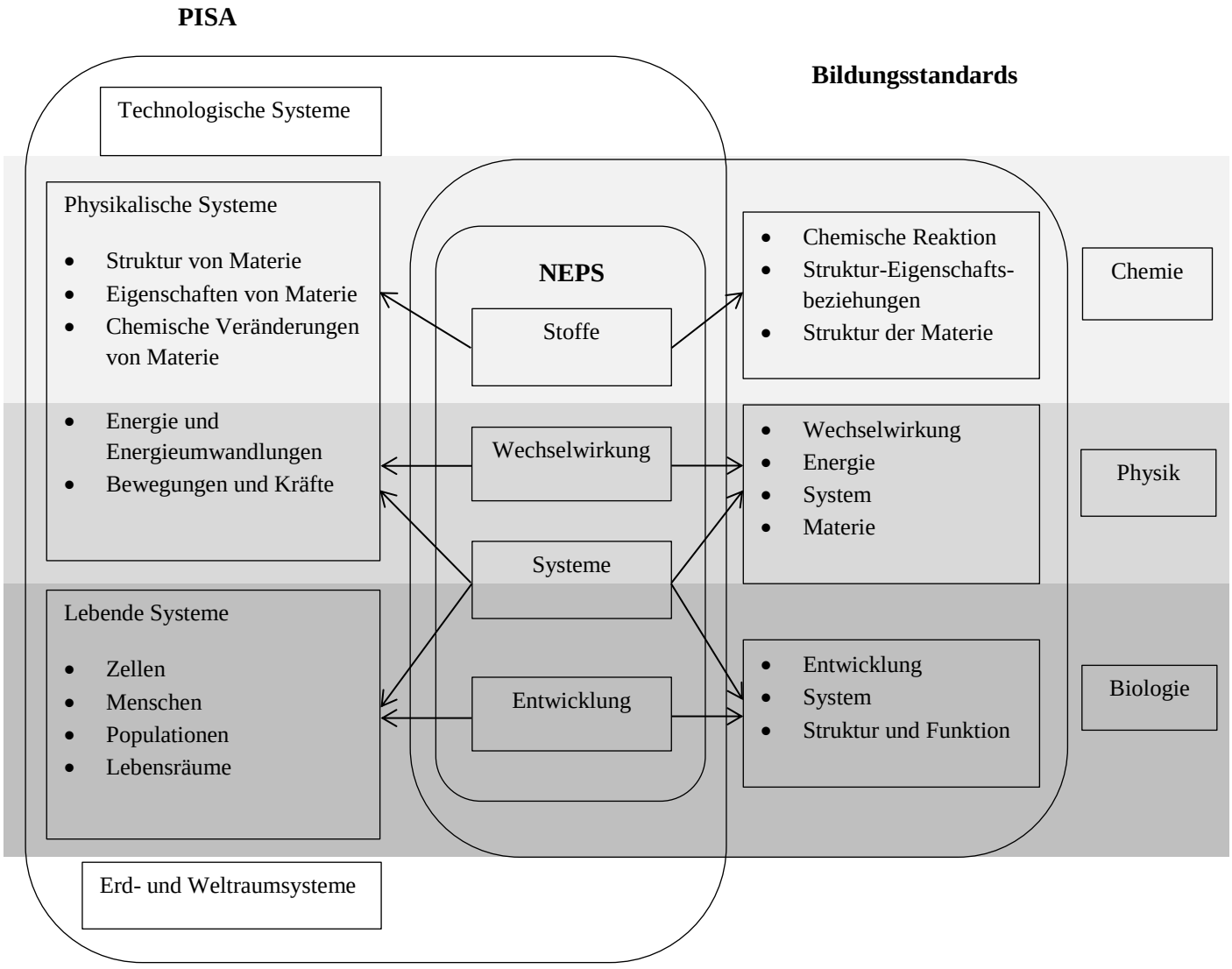

Abbildung 10: inhaltliche Überschneidung der Rahmenkonzeption des NEPS, den Bildungsstandards und von PISA.

Dem Diagramm kann entnommen werden, dass eine eindeutige Zuordnung zu den Inhalten der Rahmenkonzeptionen von PISA und den Bildungsstandards für die Mehrheit der NEPS-

Komponenten (vier von fünf) möglich war. Dieses Bild ist über die Studien hinweg konsistent: Die NEPS-Konzepte wurden in den Rahmenkonzeptionen der Bildungsstandards und von PISA meist ähnlich bezeichneten Kompetenz- bzw. Wissensbereichen zugeordnet. Beispielsweise wurden die Items des NEPS-Konzepts *Wechselwirkungen* dem *Fachwissen Physik* in der Rahmenkonzeption der Bildungsstandards und dem Wissenssystem *Physikalische Systeme* in der PISA-Rahmenkonzeption zugeordnet. Gleichzeitig kann aus dem Diagramm abgeleitet werden, dass eine eindeutige Zuordnung der Items des NEPS-Konzepts *Systeme* zu einem Kompetenzbereich der Bildungsstandards und einem Wissensbereich von PISA nicht möglich war. Die Items dieses Konzeptes verteilen sich in der Rahmenkonzeption der Bildungsstandards auf die Kompetenzbereiche *Fachwissen Physik* und *-Biologie* und *Erkenntnisgewinnung*. In der PISA-Rahmenkonzeption wurden die gleichen Items den *physikalischen-* und *lebenden Systemen* und dem *Wissen über die Naturwissenschaften* zugeordnet.

2.4.5. Ähnlichkeit der konzeptionellen Breite der Rahmenkonzeptionen

Das Experten-Review zeigt, dass nicht alle Inhalte der Rahmenkonzeptionen der Bildungsstandards und von PISA durch die NEPS-Items abgedeckt werden. Nur die Kompetenzbereiche der Bildungsstandards und die Teilkompetenzen der naturwissenschaftlichen Grundbildung der PISA-Rahmenkonzeption wurden durch die NEPS-Items inhaltlich vollständig abgebildet. Auf der Dimension *kognitive Prozesse* der Bildungsstandards haben vier von sieben Ratern dem Prozess *Reproduzieren* keine Items zugeordnet. Auch wurde die Stufe z*wei Fakten* auf der Dimension *Komplexität* von den NEPS-Items nicht besetzt. Vier von sieben Ratern konnten den Wissenssystemen *Erd- und Weltraumsysteme* und *Technologische Systeme* der PISA-Rahmenkonzeption keine Items zuordnen.

2.5. Diskussion

Das Ziel der vorliegenden Arbeit war es, die konzeptionelle Äquivalenz des NEPS-Naturwissenschaftstests, des PISA-Tests zur Erfassung naturwissenschaftlicher Grundbildung und der Tests zur Überprüfung des Erreichens der Bildungsstandards in den Fächern Biologie, Chemie und Physik zu untersuchen. Dazu wurden drei Aspekte der konzeptionellen Äquivalenz betrachtet: Die Passung der NEPS-Items zu den Inhalten der PISA- und Bildungsstandards-Rahmenkonzeptionen, die inhaltliche Überschneidung der

Rahmenkonzeptionen und die Ähnlichkeit der konzeptionellen Breite der Rahmenkonzeptionen.

Bevor die Ergebnisse eines Ratings zusammengefasst und genutzt werden können, muss zunächst die Interrater- Reliabilität geprüft werden. Zu diesem Zweck wurde in dieser Studie die Generalisierbarkeitstheorie herangezogen. Da der Generalisierbarkeitskoeffizient Φ eine zufriedenstellende Höhe erreicht und die systematische Fehlervarianz einen geringen Anteil an der Gesamtvarianz einnimmt, kann die Interrater-Reliabilität als gegeben betrachtet werden. Daraus kann geschlossen werden, dass die hier berichteten Ergebnisse auch bei einer anderen Auswahl von Ratern (die nach den gleichen Kriterien ausgewählt wurden) mit einer hohen Wahrscheinlichkeit zustande kommen würden.

Die hohe Zuordnungsrate (mind. 79%) der NEPS-Items zu den Kompetenzbereichen der Bildungsstandards bzw. den Wissensbereichen von PISA deutet darauf hin, dass der NEPS-Naturwissenschaftstest mit den PISA- und Bildungsstandards-Rahmenkonzeptionen kompatibel ist. Die im Vergleich dazu etwas schlechtere Passung der NEPS-Items (nur 64%) zu den Komplexitätsstufen der Bildungsstandards-Rahmenkonzeption kann durch eine starke Variation der Raterurteile erklärt werden. Dadurch konnte das für das Erkennen der Präferenz gesetzte Kriterium von vier Ratern pro Item und Komplexitätsstufe nicht erreicht werden. Diese Ergebnisse deuten auf eine hohe Vergleichbarkeit der theoretischen Rahmenkonzeption des NEPS für die Domäne Naturwissenschaften mit den Rahmenkonzeptionen der naturwissenschaftlichen Grundbildung von PISA und der Überprüfung des Erreichens der Bildungsstandards in den Fächern Biologie, Chemie und Physik hin.

Die inhaltliche Überschneidung der Rahmenkonzeptionen zwischen NEPS, PISA und den Bildungsstandards ist groß (vergleiche Abb.4 mit Abb.10). Insgesamt konnte hier für vier von fünf NEPS-Konzepten eine Übereinstimmung mit den Wissensbereichen von PISA und den Kompetenzbereichen der Bildungsstandards gefunden werden. Nur die Items des Konzepts *Systeme* des NEPS konnten den Inhalten der Bildungsstandards und von PISA nicht einheitlich zugeordnet werden. Diese Schwierigkeit in der Zuordnung kann bei Betrachtung der Teilbereiche des Kompetenzbereichs *Umgang mit Fachwissen* (Abb.10) dadurch begründet werden, dass der Begriff „Systeme“ in der Rahmenkonzeption der Bildungsstandards überkategorial verstanden wird. Man findet den System-Begriff sowohl im Kompetenzbereich *Umgang mit Fachwissen Physik* als auch im Kompetenzbereich *Umgang mit Fachwissen Biologie*. Ein ähnliches Bild zeigt sich in der PISA-Rahmenkonzeption. Hier werden vier Wissenssysteme unterschieden: *physikalische Systeme*, *lebende Systeme*, *Erd- und Weltraumsysteme* sowie *technologische Systeme* (Prenzel, Schöps et al., 2007). Der

System-Begriff wird somit in der PISA-Rahmenkonzeption für die Operationalisierung des *naturwissenschaftlichen Wissens* benutzt. Das erklärt möglicherweise die breite Streuung der NEPS-Items aus dem Konzept *Systeme* über die PISA- und Bildungsstandards-Rahmenkonzeptionen.

Die Ergebnisse des Experten-Reviews lassen den Schluss zu, dass die naturwissenschaftliche Kompetenz im NEPS nicht in der gleichen konzeptionellen Breite erfasst wird wie in den Bildungsstandards und bei PISA. Nur die Kompetenzbereiche der Bildungsstandards und die Teilkompetenzen der naturwissenschaftlichen Grundbildung bei PISA wurden von den NEPS-Items vollständig abgedeckt. Dieses Ergebnis kann mit dem Testdesign des NEPS erklärt werden. Die Testzeit liegt hier bei lediglich 30 Minuten, und die Items werden nicht in einem Multi-Matrix-Design angeboten. Dadurch enthält der Test deutlich weniger Items als die anderen Tests und ist also designbedingt in seiner Breite eingeschränkt.

Ein interessantes Ergebnis der vorliegenden Untersuchung ist die Erkenntnis, dass die Berücksichtigung des beruflichen und fachlichen Hintergrundes der Expertinnen und Experten auf den Dimensionen *kognitive Prozesse* und *Komplexität* (Abb. 8 und 9) nicht zur Erhöhung der Konsistenz ihrer Urteile beiträgt. Das heißt, dass es für die Konsistenz der Raterurteile irrelevant ist, welches Fach sie studiert haben und hinsichtlich welcher Studie sie ihre Erfahrung gesammelt haben. Eine mögliche Erklärung für das Zustandekommen dieser Ergebnisse kann die Unschärfe in der Beschreibung der *kognitiven Prozesse* bzw. der *Komplexitätsstufen* sein, die eine einheitliche Zuordnung der NEPS-Items unmöglich macht und zu einer unsystematischen Variation der Rater-Urteile führt.

Insgesamt zeigen die Ergebnisse, dass eine Äquivalenz der Studien-Konzepte angenommen werden kann. Sie deuten auf eine hohe Übereinstimmung der NEPS-Items mit den bei PISA und den Bildungsstandards gemessenen Inhalten sowie auf eine hohe inhaltliche Überschneidung hin. Die Ergebnisse haben nicht bestätigt, dass Unterschiede in der Spezifikationen von naturwissenschaftlicher Kompetenz einen Einfluss auf die Vergleichbarkeit der Tests haben. Stattdessen legen sie Unterschiede in der konzeptionellen Breite der Rahmenkonzeptionen offen. Es ist möglich, dass diese Unterschiede Implikationen auf die gemeinsame Interpretation der Testergebnisse haben. Dies würde bedeuten, dass die Ergebnisse der Tests nur eingeschränkt ineinander überführbar wären. Dieser Frage wird im Rahmen von weiterführenden Analysen nachgegangen.

Die hier beschriebenen Untersuchungen zur konzeptionellen Äquivalenz der Tests sind die Voraussetzung für eine inhaltliche Verknüpfung der Studien. In den sich nun

anschließenden Analysen zur dimensionalen und skalenbezogenen Äquivalenz wird in einem nächsten Schritt überprüft, ob und inwieweit die Kompetenzskalen für den Bereich Naturwissenschaften von PISA 2012 und den länderübergreifenden Bildungsstandards im NEPS verankert werden können. Nur wenn auch die dimensionale und skalenbezogene Äquivalenz gegeben sind, können Methoden entwickelt werden, um die Ergebnisse des NEPS in dem internationalen bzw. nationalen Referenzmaßstab von PISA oder den Bildungsstandards zu verorten.

2.6. Literatur

Bloch, R. & Norman, G. (2011). G String IV (Version 6.1.1). User Manual. Zugriff am 17.03.2014 unter http://fhsperd.mcmaster.ca/g_string/download/g_string_4_manual_611.pdf

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 Pro-posed Measurements. Unveröffentlichter BMBF-Antrag. Bamberg: Universität Bamberg.

Blossfeld, H.-P., Schneider, T. & Doll, J. (2009). Die Längsschnittstudie Nationales Bildungspanel: Notwendigkeit, Grundzüge und Analysepotential. *Pädagogische Rundschau, 63*(2*),* 249-259.

Brennan, R.L. (2001a). *Generalizability Theory*. New York: Springer.

Brennan, R.L. (2003). *Coefficients and Indices in Generalizability Theory*. Center for Advanced Studies in Measurement and Assessment. CASMA Research Report.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37-46.

Cronbach, L., Gleser, G. C., Nanda, H. & Rajaratnam, N. (1972). *The dependability of behavioral measurement. Theory of generalizability for scores and profiles.* New York: Wiley.

Drechsel, B., Prenzel, M. & Seidel, T. (2009). Nationale und internationale Schulleistungsstudien. In E. Wild, & J. Möller (Hrsg.). *Pädagogische Psychologie* (S. 353-380). Heidelberg: Springer.

Eisend, M. (2007). Methodische Grundlagen und Anwendungen der Generalisierbarkeitstheorie in der betriebswirtschaftlichen Forschung. *Diskussionsbeiträge des Fachbereichs Wirtschaftswissenschaft der Freien Universität Berlin.* Zugriff am 17.03.2014 unter http://edocs.fu-berlin.de/docs/servlets/MCRFileNodeServlet/FUDOCS derivate_000000000060/discpaper04_07.pdf

Hahn, I., Schöps, K., Rönnebeck, S.,Martensen, M., Hansen, S. Saß, S., Dalehefte, I.M. & 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.

Hughes, M. A. & Garrett, D. E. (1988). Inter-Coder Reliability Estimation Approaches in Marketing: A Generalizability Theory Framework for Quantitative Data. *Journal of Marketing Research, 27*, 185-195.

KMK (2005a). *Beschlüsse der Kultusministerkonferenz: Bildungsstandards im Fach Biologie für den Mittleren Schulabschluss. Beschluss vom 16.12.2004.* München: Luchterhand.

KMK (2005b). *Beschlüsse der Kultusministerkonferenz: Bildungsstandards im Fach Chemie für den Mittleren Schulabschluss. Beschluss vom 16.12.2004.* München: Luchterhand.

KMK (2005c). *Beschlüsse der Kultusministerkonferenz: Bildungsstandards im Fach Physik für den Mittleren Schulabschluss. Beschluss vom 16.12.2004.* München: Luchterhand.

KMK (2006). *Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring: siehe Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland*. München: Wolters Kluwer.

Kauertz, 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.

Kauertz, A. & Fischer, H. E. (2013). Die Operationalisierung naturwissenschaftlicher Kompetenzen im IQB-Ländervergleich 2012. In H. A. Pant, P. Stanat, U. Schroeders, A. Roppelt, T. Siegle & C. Pöhlmann (Hrsg.). *IQB-Ländervergleich 2012. Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I* (S. 50-51). Münster: Waxmann.

Kolen, M. J. & Brennan, R. L. (2004). *Test equating, scaling and linking: Methods and practices (2end ed.).* New York: Springer.

Landis J. R. & Koch G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33,* 159–174.

Li, M.N. & Lautenschlager, G.J. (1997). Generalizability theory applied to categorical data. *Educational and Psychological Measurement, 57,* 813-822.

Li, M.N. & Lautenschlager, G.J. (1999). IASGA: a SAS MACRO program for interrater agreement studies of qualitative data via a generalizability approach. *Educational and Psychological Measurement, 59,* 532-537.

Pant, H. A., Stanat, P., Pöhlmann, C. & Böhme, K. (2013). Die Bildungsstandards im allgemeinbildenden Schulsystem. In H. A. Pant, P. Stanat, U. Schroeders, A. Roppelt, T. Siegle & C. Pöhlmann (Hrsg.). *IQB-Ländervergleich 2012. Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I* (S. 13-21). Münster: Waxmann.

Pietsch, M. Böhme, K., Robitzsch, A. & T. C. Stubbe (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 und Basel: Belz Verlag.

Prenzel, M., Drechsel, B., Carstensen, C. H. & Ramm, G. (2004). PISA 2003- Eine Einführung. In M. Prenzel, J. Baumert, W. Blum, R. Lehmann, D. Leutner, M. Neubrand, R. Pekrun, H.-G. Rolff, J. Rost & U. Schiefele (Hrsg.). *PISA 2003. Der Bildungsstand der Jugendlichen in Deutschland-Ergebnisse des zweiten internationalen Vergleichs* (S. 14-46). Münster: Waxmann.

Prenzel, M., Carstensen, C. H., Frey A., Drechsel, B. & Rönnebeck, S. (2007). PISA 2006- Eine Einführung in die Studie. In M. Prenzel, C. Artelt, J. Baumert, W. Blum, M. Hammann, E. Klieme & R. Pekrun (Hrsg.). *PISA 2006. Die Ergebnisse der dritten internationalen Vergleichsstudie* (S. 31-59). Münster: Waxmann.

Prenzel, M., Schöps, K., Rönnebeck, S., Senkbeil, M., Walter, O., Carstensen, C. H. & Hammann, M. (2007). Naturwissenschaftliche Kompetenz im internationalen Vergleich.

In M. Prenzel, C. Artelt, J. Baumert, W. Blum, M. Hammann, E. Klieme & R. Pekrun (Hrsg.). *PISA 2006. Die Ergebnisse der dritten internationalen Vergleichsstudie* (S. 63-105). Münster: Waxmann.

Schiepe-Tiska, A., Schöps, K., Rönnebeck, S., Köller, O. & Prenzel, M. (2013). Naturwissenschaftliche Kompetenz in PISA 2012: Ergebnisse und Herausforderungen. In M. Prenzel, C. Sälzer, E. Klieme & O. Köller (Hrsg.). *PISA 2012. Fortschritte und Herausforderungen in Deutschland* (S. 189-215). Münster: Waxmann.

Siegle, T., Schroeders, U. & Roppelt, A. (2013). Anlage und Durchführung des Ländervergleichs. In H. A. Pant, P. Stanat, U. Schroeders, A. Roppelt, T. Siegle & C. Pöhlmann (Hrsg.). *IQB-Ländervergleich 2012. Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I* (S. 101-121). Münster: Waxmann.

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: University of Bamberg, National Educational Panel Study.

Shrout, P. E. & Fleiss, J. L. (1979). Intraclass correlation: Uses in assessing rater reliability. *Psychological Bulletin, 86*, 420-428.

Van de Vijver, F.J.R. (1998). Towards a Theory of Bias and Equivalence. In J. Harkness (Hrsg.). *ZUMA-Nachrichten Spezial, Band 3*, 41-65. Mannheim: ZUMA.

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, Nationales Bildungspanel.

Weinert, S., Artelt, C., Prenzel, M., Senkbeil, M., Ehmke, T. & Carstensen C.H. (2011). Development of competencies across the life span. In H. P. Blossfeld, H. G. Roßbach & J. v. Maurice (Eds.). *Education as a Lifelong Process: The German National Educational Panel Study (NEPS). Zeitschrift für Erziehungswissenschaft*, Sonderheft 14, S. 67-86. Wiesbaden: VS Verlag für Sozialwissenschaften.

3

Studie II:

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*

---

*Wagner, H., Hahn, I., Schöps, K., Ihme, J. M. & Köller, O. (under review). 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*.

Zusammenfassung

Verschiedene Schulleistungsstudien in Deutschland messen die Kompetenz von Schülerinnen und Schülern in den Naturwissenschaften am Ende der Sekundarstufe I, wie z.B. das *Programme for International Student Assessment* (PISA) und das *Nationale Bildungspanel* (NEPS). Die vorliegende Untersuchung beschäftigt sich mit der Überprüfung der dimensionalen und skalenbezogenen Äquivalenz der Testinstrumente von NEPS und PISA mit dem Ziel, die Skalen der beiden Studien miteinander zu verlinken. Zu diesem Zweck wurde eine Linking-Studie durchgeführt, in der 1,528 Schülerinnen und Schüler in einem Single-Group-Design die Aufgaben aus beiden Studien bearbeitet haben. Die Ergebnisse des Vergleichs zeigen einen hohen Zusammenhang zwischen NEPS und PISA sowie eine hohe Tendenz ihrer Testwerte zur Eindimensionalität. Ähnlich hoch ist die Vergleichbarkeit des NEPS- und des PISA-Tests hinsichtlich der Leistungsbewertung der getesteten Schülerinnen und Schülern. Somit belegen die Ergebnisse dieser Studie die Vergleichbarkeit der Konstrukte in den Testinstrumenten von NEPS und PISA sowie die Vergleichbarkeit ihrer Skalen. Weiterhin zeigen sich die Ergebnisse der Konstrukt-Äquivalenz-Prüfung beeinflussbar vom Umgang der Studien mit fehlenden Werten. Die Analysen legen nahe, dass sich die Vergleichbarkeit der untersuchten Testinstrumente erhöht, wenn fehlende Werte in beiden Studien ignoriert werden. Die Untersuchung der Konstrukt-Äquivalenz bildet die Basis für die Verknüpfung der naturwissenschaftlichen Skalen von NEPS und PISA. Das Linking erfolgte in der vorliegenden Studie mittels Equipercentile Equating und zeigt eine hohe Übereinstimmung hinsichtlich der Verteilung der Testpersonen auf die PISA-Kompetenzstufen. Allerdings gibt es einige kleinere Unterschiede zwischen den Studien bezüglich der individuellen Klassifikation zu den Kompetenzstufen. Insbesondere auf der ersten Kompetenzstufe kommen NEPS und PISA zu unterschiedlichen Ergebnissen. Folglich wird empfohlen, lediglich die Ergebnisse des Linkings für die Personen in den mittleren und hohen Kompetenzbereichen zu berücksichtigen.

3.1. Introduction

Several international large-scale studies, like the *Programme for International Student Assessment* (PISA; OECD, 2013) and the *Trends in International Mathematics and Science Study* (TIMSS; Martin, Mullis, Foy, & Stanco, 2012) measure science competency. These studies assess the extent to which students approaching the end of an education stage have acquired key knowledge and skills essential for their school success and later for full participation in a modern society (OECD, 2014a). Hanushek and Wößmann (2015) showed that the scientific literacy measure in PISA is a far better predictor for economic growth of a country than the results for mathematics and reading. Therefore, the assessment of scientific literacy at the end of secondary school is important not only with regard to the future careers of the students but also with regard to a nation's prosperity. Hence, Germany adopted National Educational Standards (NES; Neumann, Fischer, & Kauertz, 2010) and regularly takes part in PISA. However, these studies only allow cross-sectional analyses. Until recently no large-scale study measuring the development of competencies over the lifespan has been carried out in Germany. The National Educational Panel Study (NEPS; Blossfeld, 2008) which started in 2009 is the first German attempt to close this gap by assessing scientific literacy and the development of skills and competencies over the lifespan (Hahn et al., 2013). NEPS strives for connecting with national and international large-scale assessment studies to achieve a common interpretation of scores (Blossfeld, 2008). Linking the NEPS test to existing test procedures can be used to extend the interpretation of their test scores. This would allow classifying the NEPS test scores in a criterion-based international reference framework. This way the students whose performance lies below the baseline proficiency level or above the highest proficiency level in PISA could be examined longitudinally within NEPS in order to learn more about the conditional factors of competence acquisition and the development of educational careers.

The goal of this study is to link the Grade 9 science NEPS test with the PISA science test. According to Kolen and Brennan (2004) the linking of test scores from different studies requires a sufficient similarity of the tests with regard to (1) inferences, (2) target populations, (3) characteristics and conditions of the measurement, and (4) operationalized constructs.

3.2. Theoretical background

Kolen and Brennan (2004) suggest that the following four features in examining test similarity should be considered before the linking:

- Inferences: To what extent are the scores for the two tests used to draw similar types of inferences? In other words, to what extent do the two tests share common measurement goals?
- Populations: To what extent are the two tests designed for testing similar populations?
- Characteristics and conditions of the measurement: To what extent do the two tests share common measurement conditions, for example, with regard to test format, administration conditions, test length, etc.?
- Construct: To what extent do the two tests measure the same construct?

To closely examine these aspects the next section will look into the similarities and differences of the NEPS and PISA frameworks.

3.2.1. Comparing the scientific literacy tests of NEPS and PISA

Inferences

The studies PISA and NEPS differ in the type of inferences derived from their measurements. The aim of PISA is to monitor educational systems at the end of secondary school in terms of student performance (OECD, 2013). This goal is realized every three years by a cross-sectional overview of the educational level of 15-year-old students. The aim of NEPS is to provide longitudinal data of the competence development from early childhood to late adulthood in Germany. In order to achieve this goal the data collection in NEPS is embedded in a multicohort sequence design (Maurice, Sixt, & Blossfeld, 2011) which makes it possible to compare the educational level of 9th grade students from different cohorts. In other words, despite the different objectives the studies NEPS and PISA have a substantial overlap in their measurements as they assess the educational level of students at the end of secondary school.

Target populations

The target population of the NEPS test are 9th grade students (Maurice et al., 2011). PISA examines the competence of 15-year-old students (15 years and 3 months to 16 years and 2 months of age). Hence, the overlap of both target populations is high as most of the German students attend Grade 9 or 10 at the age of 15 (OECD, 2014b).

Characteristics and conditions of the measurement

In Germany data collection and processing for PISA 2012 and NEPS 2010 were coordinated by the IEA Hamburg. In 2012 both tests (NEPS and PISA) were administered as a paper pencil test. The majority of the items in NEPS 2010 and PISA 2012 have a closed-constructed response format (OECD, 2014a; Schöps & Saß, 2013). However, PISA 2012 also uses an open-constructed response format.

NEPS and PISA deal with the missing responses by the estimation of person parameters in a different way. PISA 2012 uses a two-stage procedure for handling missing responses (OECD, 2009): in the first step missing responses are ignored to estimate the item parameters. In the next step the estimated item parameters are used for the estimation of person parameters where missing responses are scored as incorrect. In contrast, NEPS ignores the missing responses for the estimation of item and person parameters (Pohl & Carstensen, 2012).

As Pohl, Gräfe, and Rose (2014) have pointed out different ways of handling missing responses can have an effect on the estimation of the person parameters. In the classical approaches the missing responses (a) may be ignored and handled as if the items were not administered; or (b) may also be scored as incorrect, assuming that the test person could not solve the item; or (c) may be scored as partially correct by taking the probability of guessing into account. Another possibility (d) is that the missing responses are handled differently depending on the parameters that need to be estimated: Item responses are ignored when estimating item parameters. In the next step the estimated item parameters are used for estimating person parameters where missing responses are scored as incorrect. A number of studies (De Ayala, Plake, & Impara, 2001; Finch, 2008; Lord, 1974; Ludlow & O'leary, 1999; Pohl et al., 2014; Rose, von Davier, & Xu, 2010) showed that the scoring of missing responses as wrong leads to a bias in the estimation of parameters and to the overestimation of the reliability. Therefore, it is important to examine to what extent the different handling of the missing responses by NEPS and PISA influences the comparability of their science scores.

The science tests of NEPS and PISA also differ in the number of items: in 2012 the PISA test included 53 items that are split in three clusters with each cluster representing 30 minutes of test time. The 28 items of the NEPS test are presented in 28 minutes and each person gets the same items in a fixed sequence. This difference in number of items can lead to a divergence of the conceptual width of the measured construct as pointed out by Wagner, Schöps, Hahn, Pietsch, and Köller (2014).

Operationalized Constructs: Comparing the contents of the science tests of NEPS and PISA

The definition of scientific literacy used by NEPS includes aspects of the *concept of competence* as defined by Weinert (2001), and of the *concepts of scientific literacy* developed by the American Association for the Advancement of Science (AAAS, 2009) and by PISA (OECD, 2006). Therefore the NEPS scientific literacy framework has a substantial overlap with the scientific literacy framework from PISA 2012 (Figure 1).

	NEPS	**PISA**
Knowledge and contexts	*Content-related components* matter, interactions, development, systems *Process-related components* scientific enquiry, scientific reasoning *Contexts* health, environment, technology	*Knowledge of science (KOS)* *physical systems* (structure of matter, properties of matter, chemical changes of matter, motions and forces, energy and its transformation, interactions of energy and matter); *living systems* (cells, humans, populations, ecosystems, biosphere); *Earth and space systems* (structures of Earth systems, energy in Earth systems, change in Earth systems, Earth's history, Earth in space); *technology systems* (role of science-based technology, relationships between science and technology, concepts, important principles) *Knowledge about science (KAS)* scientific enquiry, scientific explanations *Contexts* health, natural resources, environment, hazard, frontiers of science and technology
Competencies		identifying scientific issues, explaining phenomena scientifically, using scientific evidence
Source	(Hahn et al., 2013)	(OECD, 2013)

Fig.1. Framework of the scientific literacy by PISA and NEPS.

The frameworks of both studies differ in the number of components used for assessing scientific literacy: The framework of NEPS considers only the *content-related components* which are related to the *knowledge of science* in PISA, and *process-related components* which are related to the *knowledge about science* in PISA. The PISA framework differentiates further and also distinguishes between the competencies *identifying scientific issues*, *explaining phenomena scientifically* and *using scientific evidence*. At this point, it can be concluded that the frameworks of the two studies differ in their conceptual scope. But how different are they on the task level?

This question can be examined within the theory of bias and equivalence of van de Vijver (1998). The author suggests assessing the similarity of the operationalized constructs of two tests by regarding their conception, their dimensional structures and their scales. In order to analyze the conceptual equivalence of the scientific literacy in NEPS and PISA (Wagner et al., 2014) seven experts in the field of science didactics familiar with large-scale assessments classified the NEPS items according to the categories of KOS and KAS and to the competencies in PISA. The results showed that 79 percent of the NEPS items could be assigned to the contents of the PISA framework. However, according to five out of the seven raters some of the KOS components in PISA (earth and space systems and technology systems) were not covered by NEPS items. This is partly due to the strict time limitations of the NEPS test (28 minutes) which only allows for selected components.

After comparing the similarity of the operationalized constructs, the analysis of van de Vijvers two other aspects – the equivalence of dimensional structures and scales – will have to be completed in order to link the tests of NEPS and PISA. According to Kolen and Brennan (2004), strong linking of test scores from different studies requires a sufficient similarity of their test frameworks. The comparison of NEPS and PISA frameworks presented here shows that both studies make the same inference about the same target population. Also, the constructs in NEPS and PISA have a substantial overlap but differ in their conceptual width. Regarding the measurement conditions, NEPS and PISA use the same response formats but differ in the handling of missing values. In particular, the influence of this last aspect on the comparability of NEPS and PISA test scores and on the linking of their scales has to be investigated.

3.2.2. Linking-methods

Depending on the level of equivalence of the two tests, different methods of linking can be applied. Mislevy (1992) and Linn (1993) differentiate between five types of linking: *equating*,

vertical scaling, *concordance*, *projection* and *moderation*. The *moderation* is the weakest type of linking and can be used if the tests are dissimilar with regard to the inferences, constructs, population and measurement characteristics. This type of linking is usually used if the tests have different frameworks but similar constructs. The scores of the tests linked by the moderation method have different meanings and can be compared only with regard to their mean (like the z scores of two tests). For the linking with *projection* the tests have to be appropriate for the same population but may differ in inferences, constructs and measurement characteristics. If the linking is successful, it is possible to predict the score of Test A from the score of Test B, but not vice versa. The *concordance* method can be used to link tests with similar inferences, constructs and populations but with different measurement characteristics. This method is sensitive to the sample working on the test: different samples can lead to different linking results. For the linking with *vertical scaling/calibration* the tests must have similar inferences, constructs and measurement characteristics, but may differ in their populations. To link the two tests, IRT methods can be applied. The strongest type of linking is *equating*. It is often used for parallel tests. This method can only be applied, when the tests are similar in all features named in the Kolen and Brennan (2004) approach. The linking with equating makes it possible to predict the scores from one test to another and vice versa. However this method also has one disadvantage: It is sensitive to irregularities in the distribution of the test scores. For example, the equating relationship cannot be determined for scores ranges that exceed the highest observed score and scores that fall below the lowest observed score. One possibility of approaching the problem is by pre- or post-smoothing the equipercentile equivalents (Livingston, 2004).

3.2.3. Linking-studies to locate the outcomes of the national tests in an international reference

Worldwide there have been different approaches to link international data with national data since the first PISA survey in 2000. In Canada the results of 15-year-old students in the national test for reading comprehension of the Foundation Skills Assessment (FSA) were linked via equipercentile equating to the results of PISA 2000 (Cartwright et al., 2003). The goal of this study was to locate the outcomes of the national study on the international scale. The comparison of the linked FSA and PISA scores showed that the highest FSA performance level *exceeds expectations* is set above the threshold for the highest PISA proficiency Level 5. This means that if the PISA test were used to identify top readers in Canada, a greater number of students would be classified as exceeding provincial expectations as based on the FSA. The

comparison on the middle proficiency levels leads to similar results: students with reading scores categorized as *meets expectations* in FSA cover Levels 3, 4 and 5 in PISA. And finally, students who are classified as *not within expectations* performed at about Level 2 in PISA or below. It can be concluded that the FSA test uses stricter standards than PISA. Furthermore, the linking procedure showed that the linked scores of one test are not interchangeable with scores from the other test for individual students. This means that linking FSA and PISA scores is only valid when the results are used to make inferences about groups of students (about 30 or more) but linking individual scores is not possible.

Another attempt to connect a national assessment with an international assessment was undertaken in the USA in 2009 (Hambleton, Sireci, & Smith). The goal was to compare the outcomes of the National Assessment of Educational Progress in mathematics (NAEP) with the competencies of students in mathematics assessed by TIMSS and by PISA. The linking of the three achievement levels (basic, proficient, advanced) from the NAEP reporting scale was carried out via equipercentile equating. The study results show that the proficiency levels in NAEP are comparable with the standards of TIMSS and PISA.

Nissen, Ehmke, Köller, and Duchhardt (2015) compared different linking methods by applying them to the mathematics tests of 4th grade students in TIMSS 2011 and 5th grade students in NEPS 2010. The comparison of the equipercentile equating with the IRT linking method on the sample of 733 4th graders showed a satisfying classification consistency of the NEPS results to the TIMSS international benchmarks (44% and 36% respectively). Furthermore the authors found that the distribution of the linked NEPS and TIMSS score equivalents is approximately the same when the equipercentile equating is used. This offers an advantage compared to IRT linking. Due to minor differences in classification according to the international benchmarks the authors recommended inferences about groups of students but not about linking scores of individuals.

The overview of the linking-studies shows the relevance of the linkage between the national tests and the tests with international reference. Furthermore the last study found that the linking via equipercentile equating leads to a higher classification consistency between two tests. This result led the choice of the linking-method in the present study.

3.3. Research questions

The present paper examines the comparability of the science scores in NEPS 2010 and PISA 2012. For this purpose we investigate the following research questions:

To address our questions concerning the *dimensional equivalence,* we will investigate (1) to what extent the science scores in NEPS and PISA correlate, (2) whether the items of the NEPS and PISA tests measure the same construct of scientific literacy and (3) how different handling of missing responses in NEPS and PISA influences the comparability of their science scores.

To investigate the *scalar equivalence* of both tests, we will examine (4) to what extent the person parameters in NEPS and PISA are equivalent regarding the measures of data distribution before and after linking. (5) We will also gauge the classification consistency between the linked test scores in NEPS and PISA according to the PISA 2012 international benchmarks and (6) inspect whether and to what extent different handling of missing responses in NEPS and PISA influences the linking of the science scores.

3.4. Method

3.4.1. Data collection

An empirical study was conducted in four federal states in Germany using a single group design to assess the equivalence of the scientific literacy scores in PISA and NEPS. The study was carried out on two consecutive days parallel to the PISA survey in the spring of 2012. On the first day, every student completed the PISA science test (consisting of one of three clusters, overall 53 items). On the second day, the students took the NEPS science test which consisted of 28 items. The test booklets used in the presented study were the same as in the main studies of PISA 2012 and the NEPS 2010. The data collection and processing in the main NEPS and PISA studies as well as in the linking study was coordinated by the IEA Hamburg. The sample consisted of 1,528 9^{th} grade students from 65 schools that participated in the secondary school program *Increasing the Efficiency of Teaching in Mathematics and Science Education in Secondary School* (*SINUS*; Prenzel & Ostermeier, 2006; Prenzel, Stadler, Friedrich, Knickmeier, & Ostermeier, 2009). Overall N = 1,079 ninth grade students (50% female, age mean m = 15.5 years) took both the NEPS and the PISA science tests.

3.4.2. Scoring and data procedures

Since the year 2000 Germany has participated in PISA, which takes place every three years. Each PISA cycle examines a major domain in depth, so two-thirds of testing time is devoted to this domain; the other domains provide a summary profile of skills. Major domains were reading literacy in 2000 and 2009, mathematical literacy in 2003 and 2012 and scientific

literacy in 2006. PISA 2012 used a part of the items again that were developed in 2006, therefore the scoring of the PISA data in this study was conducted by applying the coding rules of PISA 2006 (OECD, 2009). The scoring of the NEPS data was carried out using the NEPS 2010 coding rules (Schöps & Saß, 2013).

In order to assess the comparability of scientific literacy scores in NEPS and PISA the software ConQuest (Wu, Adams, Wilson, & Haldane, 2007) was used to compute the person parameters of both studies by assigning the items to the corresponding tests as two dimensions. Our first research question (addressing the relation between the NEPS and PISA scores) was analyzed by correlating the two dimensions.

In order to examine the second research question regarding the extent in which the items of PISA and NEPS measure the same construct of *scientific literacy*, a factor analysis was conducted by using the software M*plus* (Muthén & Muthén, 2012). The goal of the factor analysis in our study was to find out whether a common factor *scientific literacy* can entirely explain the variance of both tests. Another possibility is to assume a second factor differentiates between tests. Additional analyses were carried out due to the different handling of missing data in NEPS and PISA in order to examine, which method of dealing with missings increases the comparability between the tests. For this reason the one factor model was compared with the two factor model with regard to the information criteria and the chi-square test.

The information criteria (AIC, BIC, SABIC) are measures of the goodness of fit of an estimated statistical model. The Akaike information criterion (AIC; Akaike, 1973) is the first developed model selection criterion which, in addition to the likelihood of the model to be analyzed, takes into account the number of estimated parameters in this model. Compared to the AIC, the Bayesian information criterion (BIC; Schwarz, 1974) strongly penalizes the number of parameters. The sample-size adjusted BIC (SABIC; Sclove, 1984) places a penalty for adding parameters based on sample size but less strongly than the BIC. Several simulation studies (Enders & Tofighi, 2008; Tofighi & Enders, 2007) showed that the SABIC is a useful tool for comparing models. Therefore, the results of the factor analysis are reported for AIC, BIC and SABIC but the SABIC gets a higher weight when selecting the model. Given a set of candidate models for the data, the preferred model is the one with the minimum information criteria value. Raftery (1995) proposes to interpret the difference of $\Delta BIC \geq 10$ as very strong evidence for a better model fit.

The chi-square test (χ^2) evaluates the null hypothesis H_0 (observed values = expected values) against the alternative H_1 (observed values ≠ expected values). If the chi-square test

value is significant, the null hypothesis will be rejected. Because the model will usually be rejected by using the likelihood ratio chi-square test if the sample size is large (>400; Bagozzi, 1981; Bentler & Bonett, 1980), the interpretation of this result must be regarded with caution.

Another method of assessing the unidimensionality of the tests is to examine the assumption of local item independence (LII), one of the cornerstones of the item response theory model (Embretson & Reise, 2000). The LII means that the observed items are conditionally independent of each other given an individual score on the latent variable (Henning, 1989). If the items of different tests measure different constructs, the LII is violated. Any statistical analysis based on the Rasch model is unjustified if the assumption of local item independence is violated. Moreover the violation of the local item independence is the evidence that the tests measure different constructs and are not equal.

One of the statistical procedures to identify the local item dependency is the computation of the partial correlation index (PRT) from Huynh, Michaels, and Ferrara (1995): the predicted value from the linear regression on the raw score of the total test (in our study the weighted likelihood estimates (WLE)) is subtracted from the item raw score. The correlations between the residuals of items are partial correlations and the mean of the partial correlations (PRT) is the index of the local item dependency. The residuals should be uncorrelated and generally close to zero. The critical value of 0.2, as proposed by Chen and Thissen (1997) indicates a violation of the local independence.

3.4.3. Linking procedures

To assess the scalar equivalence, the raw scores of NEPS and PISA were analyzed based on the 1PL Rasch model with fixed item parameters taken from NEPS 2010 (Schöps & Saß, 2013) and PISA 2012 (OECD, 2014a). Five plausible values were drawn per student and test and linearly transformed to the mean of 500 and a standard deviation of 100. The mean, skewness and kurtosis were compared to assess the scalar equivalence of the NEPS and PISA science tests. Nissen et al. (2015) showed that the classification consistency of linking via equipercentile equating is higher than via IRT linking. Therefore the linking of the NEPS and PISA scores in the present study was conducted via equipercentile equating (e.g. Cartwright, 2012; Nissen et al., 2015; van den Ham, Ehmke, Nissen, & Roppelt, 2016). This procedure is based on the idea that the scores of the two tests with the same percentile rank are declared as equivalent (Kolen & Brennan, 2004). For example, if a score 425 is equal to percentile rank of 10 on the NEPS scale, and a score 496 is equal to percentile rank of 10 on the PISA scale,

then the scores 425 and 496 are considered equivalent. Due to the sensitivity of equipercentile equating to irregularities in the distribution of the test scores (Livingston, 2004) the NEPS equivalents were post smoothed with a value of 0.3. Post smoothing means that equipercentile equating is performed on the basis of observed distributions and the equating relation was smoothed. This step produces a smoothed distribution with nonzero probability at the highest and lowest score levels (Kolen & Brennan, 2004). In our study equipercentile equating was carried out for each plausible value by using the computer software LEGS (Brennan, 2003b). Afterward the linking results were averaged.

In the last step, the students were classified according to the international benchmarks of PISA 2012 (OECD, 2013) based on the PISA scores and the NEPS equivalents on the PISA metric. To analyze the scalar equivalence the distribution of the classification was compared and the percentage matching (the percentage of the students that are assigned to the same benchmark) was calculated. To measure the concordance between the tests of NEPS and PISA the *Cohen's coefficient Kappa (1960)* was calculated. Landis and Koch (1977) propose the following labels assigned to the corresponding ranges of kappa: "0.41-0.60" moderate, "0.61-0.80" substantial and "0.81-1.00" almost perfect agreement.

Due to the different handling of the missing values in NEPS and PISA the research questions described earlier were analyzed in two ways in order to answer research questions three and six: in the first case the missing values were handled study-specific: the missings in PISA were scored as incorrect and missings in NEPS were ignored. In the second case the missings in both tests were ignored.

3.5. Results

3.5.1. Assessing the dimensional equivalence

The aim of this section is to examine the comparability of the scientific literacy scores of NEPS and PISA including different ways of dealing with the missing values. Figures 2 and 3 provide the average percentage distribution of missing and invalid values in NEPS and PISA. They show that the PISA test has significantly ($p < .05$) more missing values than the NEPS test. In PISA these values are scored as incorrect when estimating the person parameters (OECD, 2009). NEPS ignores all missing values when estimating the item and person parameters (Pohl & Carstensen, 2012).

The dimensional structure of the test data in NEPS and PISA was investigated by correlating the two test dimensions. Our analyses show that the relation between the two

dimensions seems to depend on the handling of missing data: if the missing data are scored study-specific the correlation between NEPS and PISA is r = .84. This correlation increases to r = .90 if the missing values are consistently ignored. The difference in the height of the correlation is statistically significant (p < .05). The NEPS and PISA tests share 71% (if the missings are scored study-specific) to 81% (if the missings are ignored) of the variance. However, how the missing values are handled has no influence on the reliability of the data (PV Reliability NEPS = .84, PV Reliability PISA = .80).

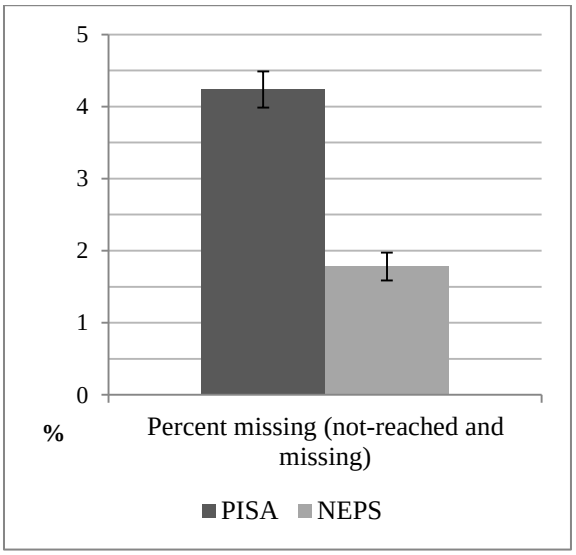

Fig. 2. Average percentage distribution of missing values in NEPS and PISA.

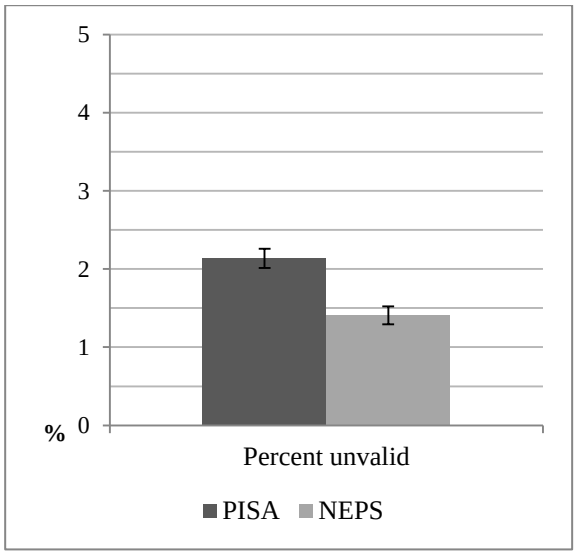

Fig. 3. Average percentage distribution of invalid values in NEPS and PISA.

The second research question asking to what extent the tests of NEPS and PISA are measuring the same *scientific literacy* construct was examined with a factor analysis taking into account the different handling of missing values. Table 1 shows the results if the missings are scored study-specific.

Table 1
Summary of factor analysis of the NEPS and PISA data (the missings are scored study-specific).

	N of parameters	Log likelihood	AIC	BIC	SABIC	χ^2 (df)
1 factor	243	60300	60786	61997	61225	3514 (3159)*
2 factors	323	59924	60571	62181	61155	3138 (3079)

AIC: Akaike information criterion; BIC: Bayesian information criterion; SABIC: Sample-Size Adjusted BIC; * = p < .05.

The AIC index and the SABIC index of the one factor model are smaller compared to the AIC index and the SABIC index of the two factor model. Also the chi-square test value of the one factor model is significant and shows that the observed values are not equivalent to the expected values. However the BIC index of the one factor model is smaller compared to the BIC index of the two factor model.

If the missings are ignored, the results of the factor analysis (Table 2) show a different picture for the unidimensionality. The BIC index and the SABIC index of the one factor model are smaller compared to the two factor model. Also the chi-square test value of the one factor model is not significant and shows that the observed data does not contradict the expected model. Only the AIC index of the two factor model is smaller compared to the one factor model. Following the results of assessing the dimensional equivalence, the comparability of the science values in NEPS and PISA is highest when the missings are ignored in both tests.

Table 2
Summary of factor analysis of the NEPS and PISA data (the missings are ignored).

	N of parameters	Log likelihood	AIC	BIC	SABIC	χ^2 (df)
1 factor	243	56156	56643	57854	57082	3292 (3159)
2 factors	323	55922	56568	58178	57152	3057 (3079)

AIC: Akaike information criterion; BIC: Bayesian information criterion; SABIC: Sample-Size Adjusted BIC; * = $p < .05$.

The next step examined the assumption of local item independence for the items of NEPS and PISA. Figure 4 provides information about the correlations between the item residuals in NEPS and PISA after the individual score on the latent variable (*scientific literacy*) is subtracted from the item raw score. The diagram shows that only few correlations exceed the critical value of .2. The mean of the partial correlations is - .02 irrespective of how the missing values are handled.

Whereas few correlations in Figure 4 exceed the cut score of 0.2, the assumption was tested, whether the high partial correlations are study-specific and can be obtained even in a two-dimensional scaling of data. For this purpose, the data from NEPS and PISA were scaled two-dimensionally and the correlations between the item residuals were calculated separately for each test. Figure 5 shows the partial correlations for NEPS and PISA data when missings were handled study-specific and when missings were consistently ignored. The results suggest that the high partial correlations in PISA are stable despite the two-dimensional scaling. Compared to the partial correlations of PISA, the NEPS correlations range between -0.13 and 0.14. These correlations can be considered as normal. A closer look at the PISA items in Figure 5 shows that the study-specific handling of missings leads to a higher number of critical partial correlations than when missings were ignored (25 to 17).

The results of assessing the dimensional equivalence indicate a better comparability of the NEPS and PISA science scores under the condition of ignoring the missing data. Therefore the scalar equivalence was assessed under this condition.

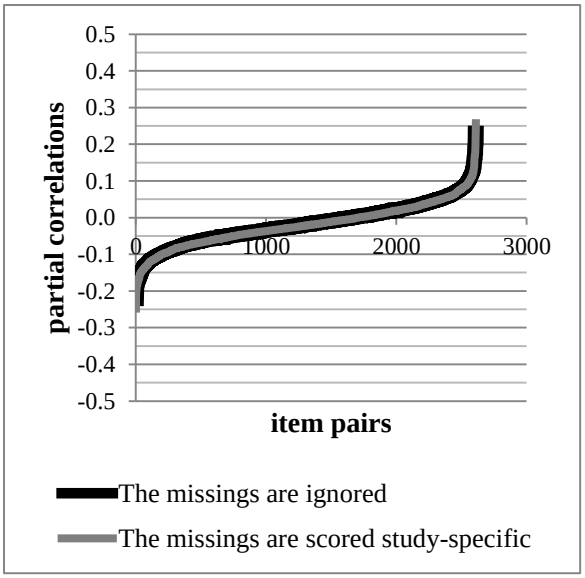

Fig. 4. The partial correlations of item pairs in NEPS and PISA from the one-dimensional scaling.

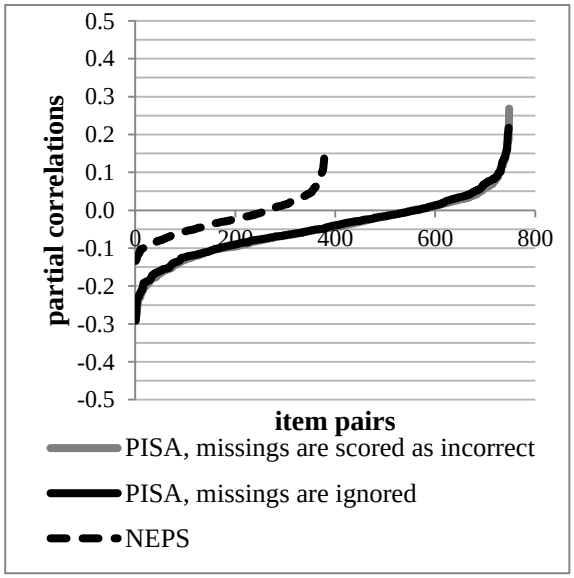

Fig. 5. The partial correlations of item pairs in NEPS and PISA from the two-dimensional scaling.

3.5.2. The distribution of the person parameters in NEPS and PISA

Figure 6 shows the linear relationship between the person parameters in NEPS and PISA measured by the first plausible value when the missings were ignored. The average correlation of the competency scores in NEPS and PISA is .91. The graph shows that the performance of most of the students is located in the higher range (above the ability mean of 0 logits). This might be caused by their participation in the primary school program *SINUS* (Dalehefte et al., 2014), which might lead to a higher average ability than the ability of the reference sample.

The descriptive statistics for the NEPS and PISA scores are provided in the upper part of Table 3. The statistics show significant differences ($p < .05$) between the means of person parameters in NEPS and PISA. The students achieved higher proficiency values in the PISA test than in the NEPS test. The data of both tests are normally distributed and do not differ in skewness and kurtosis.

3.5.3. The linking

Figure 7 provides a graphic representation of the equipercentile equating of NEPS scores. The results of the equipercentile equating are provided in the lower part of Table 3. The descriptive statistics show that the mean as well as the skewness and kurtosis of the NEPS equivalents on the PISA metric strongly resemble the PISA scale statistics.

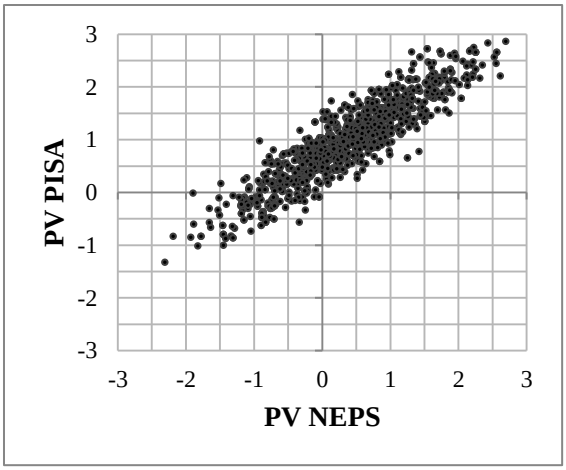

Fig. 6. Relationship between the NEPS and PISA person parameters.

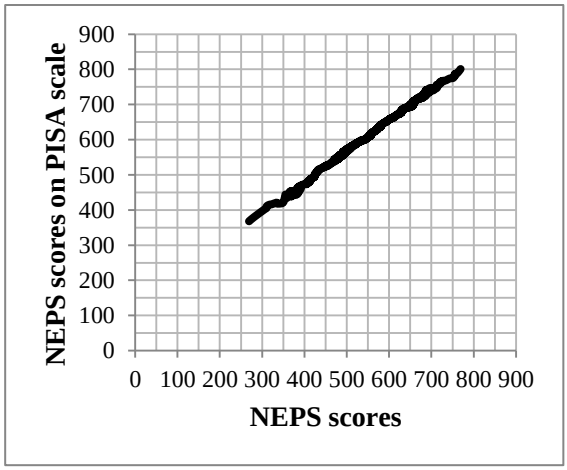

Fig. 7. Matching of NEPS person parameters and the equivalents on the PISA scale.

Table 3

Descriptive statistics for NEPS and PISA science scores and results of the equipercentile linking.

	Mean (SD)	*SE*	Skewness	*SE*	Kurtosis	*SE*
NEPS scores	540.16 (86.67)	2.64	-0.04	.07	-0.21	.15
PISA scores	600.87 (75.12)	2.29	-0.06	.07	-0.13	.15
$NEPS_{PISA}$	600.86 (75.05)	2.28	-0.06	.07	-0.13	.15

SD: standard deviation; SE: standard error.

In the next step the students were classified according to the PISA international benchmarks in science (Figure 8). The classification of students based on the linked PISA and NEPS scores is equivalent in percentage ($\chi^2 = 0.95$; $df = 5$, $p > .05$). The Cohen's coefficient Kappa (a measure of consistency) is $k = 0.55$, so the matching between the tests can be rated as *moderate* (Landis & Koch, 1977).

The matching of the students according to the PISA 2012 international benchmarks based on the linked PISA and NEPS scores is provided in Table 4. In contrast to the classification in Figure 8, the individual assignment of students differs between the tests. From 29% to 70% of the students are allocated to the same proficiency level irrespective of the test used for the classification. The consistency between the tests on proficiency Level 1 is particularly low. Most of the 3 to 9 persons (depending on the choice of the PV) assigned to the proficiency Level 1 in PISA are assigned to the proficiency Level 2 in NEPS. Overall, 60% of the students were assigned to the same proficiency level. If the missing values were treated study-specific the percentage matching between NEPS and PISA drops to 55%.

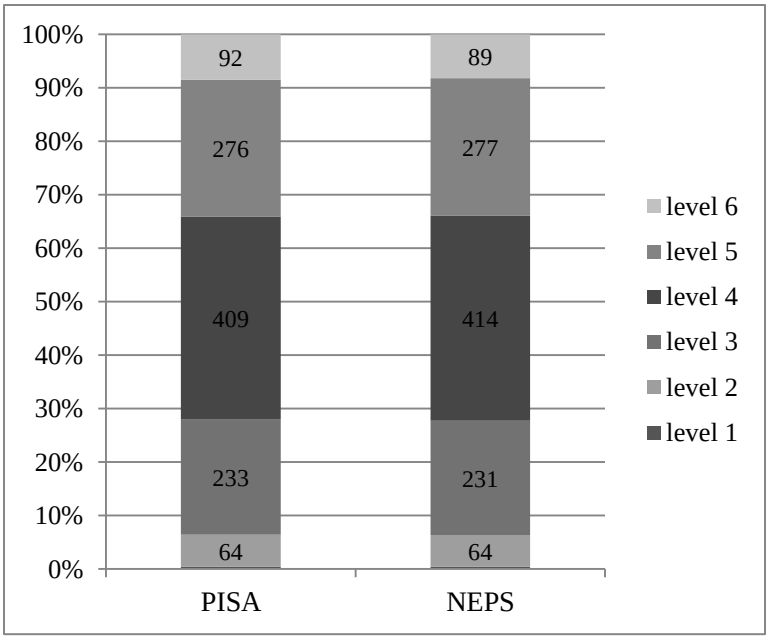

Fig. 8. Classification of students according to PISA 2012 international benchmarks in science.

Table 4
Percentage of students classified according to the PISA 2012 international benchmarks in science.

		PISA 2012 international benchmarks					
		1	2	3	4	5	6
$NEPS_{PISA}$	1	**29**	4				
	2	71	**64**	8			
	3		31	**65**	14		
	4			26	**70**	25	1
	5			1	16	**66**	32
	6					9	**67**
Total		100	100	100	100	100	100
Total number of persons		5	64	233	409	276	92

3.6. Discussion

The aim of this study was to link the NEPS 2010 science test with the PISA 2012 science test. As a basis for this analysis, the comparability of the scientific literacy scores has been examined for both tests, taking into account the study-specific way of dealing with missing values. For this purpose the dimensional equivalence and the scalar equivalence of the test scores were assessed.

3.6.1. Dimensional equivalence

The first research question pertained to the latent relation between the science scores of NEPS and PISA. The analyses show a high correlation between the tests, and the amount of the

shared variance is high. However, both the correlation and the shared variance depend on the handling of missing data.

The handling of missing data also influences the dimensionality of the tests. When the missings are ignored in both tests most of the tested fit indices favor the one factor model. In contrast to this result most of the tested fit indices prefer the two factor model when the missings were treated study-specific. It can be concluded that the study-specific handling of missings makes the assumption of an additional factor necessary. The factor of *scientific literacy* cannot explain the systematic item variance completely.

This finding is related to the examination of local item independence of NEPS and PISA items. When the missings are handled study-specific the PISA test shows more high partial correlations compared to the case where the missings are ignored. These results can be caused by the bias in the estimation of parameters through the handling of missing responses as wrong (De Ayala et al., 2001; Finch, 2008; Lord, 1974; Ludlow & O'leary, 1999; Pohl et al., 2014; Rose et al., 2010).

The examination of local item independence also shows that the PISA test has more partial correlations compared to the NEPS test. This may be a consequence of the violation of local item independence of the items within a unit (Monseur, Baye, Lafontaine, & Quittre, 2011). The common scaling of NEPS and PISA data however shows no evidence for multidimensionality. Only few partial correlations exceed the critical value and the mean of the partial correlations is low.

Assessing the dimensional equivalence of the NEPS and PISA data does not contradict the assumption that these tests measure the same construct *scientific literacy*. Furthermore the results show the importance of the measurement conditions (i.e. the handing of missings) for examining the similarity of two tests. The results of our study show that ignoring the missings in both tests creates favorable conditions for the comparability of their scores. The last PISA assessment (OECD, 2016) partially eliminates the differences to the NEPS assessment in the handling of missing values by ignoring the not-reached missing values. This change in the framework of PISA 2015 increases the comparability of NEPS and PISA results.

The assessment of the conceptual equivalence between NEPS and PISA (Wagner et al., 2014) showed a substantial overlap between the science tests of the two studies. The results of the dimensional analyses support this finding. Both steps in assessing the equivalence make it clear that the tests of NEPS and PISA are very similar but not exchangeable, especially with regard to their inferences.

3.6.2. Scalar equivalence and linking

Due to the better comparability of the NEPS and PISA science scores under the condition of ignoring the missing data the scalar equivalence was assessed under this condition. The distributions of the NEPS and PISA test scores in our study show different means, but no differences in skewness and kurtosis. Similar to the study results of Nissen et al. (2015), the students in our study achieved higher test scores on the PISA test than they do on the NEPS test. To compensate for these differences, the test scores in NEPS and PISA were linked by equipercentile equating. The distribution of the linked NEPS scores is equivalent to the distribution of the PISA scores.

The student classification according to the PISA 2012 international benchmarks on basis of the linked NEPS and PISA scores leads to very similar results. Also, the classification consistency by the Cohen's Kappa $k = 0.55$ and the percentage matching of 60% show a good approximation. The study-specific handling of missings also leads to a comparatively high consistency, but loses about 5% of the agreement compared to the ignoring of missings due to the lower correlation between the science scores.

The classification consistency is influenced by the correlation between the tests, their reliabilities and the number of proficiency levels (Ercikan & Julian, 2002; Pietsch, Böhme, Robitzsch, & Stubbe, 2009). Pietsch et al. (2009) showed that the expected consistency between two tests with five proficiency levels is 42% if the tests have the reliability of Rel_1=1 and Rel_2=.8 and the correlation between the tests is r= .9. Furthermore the classification accuracy decreases by 10% for an increase by one proficiency level (Ercikan & Julian, 2002). PISA has six proficiency levels. Therefore, the percentage matching of 60% exceeds the expected value. Compared with the study results of Nissen et al. (2015) the percentage matching calculated in our study can be rated as very satisfactory. However, similar to the study results of Cartwright et al. (2003) and Nissen et al. (2015) there are some minor differences between NEPS and PISA regarding the individual classification to the benchmarks. Thus, inferences from the NEPS and PISA results can be drawn for groups of students but not for individual students.

The NEPS study does not use proficiency levels to classify test persons. Therefore the present study offers no opportunity to compare the educational standards in NEPS and PISA. However, the individual student classification according to the PISA 2012 international benchmarks show the tendency of the NEPS test to classify a high proportion of students assigned to the first three levels by PISA to the next higher level. On the other hand, students

assigned to the last two levels by PISA fall to the next lower level by NEPS. Thus, the presented study provides no evidence that the NEPS test systematically overestimates or underestimates the abilities of individuals in comparison to the PISA test. However, the mean difference of the two studies shows that the students achieved higher test scores on the PISA test than they did on the NEPS test.

3.6.3. Limitations

This study examined the equivalence of the NEPS test from 2010 and the PISA test from 2012. From 2012 to 2015 (the last PISA investigation) the tests of NEPS and PISA have been changed in many ways: for the NEPS assessment in 2014 new items were developed in addition to the items of 2010 so that a three-stage (easy, average and difficult) test could be administered and could still be linked to the test from 2010 by using link items. PISA 2015 introduced the following changes in the test administration and scaling (OECD, 2016): the assessment mode (computer-based instead of paper-pencil), the scaling model (two-parameter model instead of one-parameter model), the handling of differential item functioning across countries (calibration for a number of country-by-cycle-specific deviations from the international item parameters instead of ignoring of the "dodgy" items for some countries), the handling of non-reached items (dealing as not administered instead as wrong answers when estimating the person parameters) and finally the changes in the framework of the scientific literacy (e.g. the KOS component technology systems was excluded from the framework). These changes are not relevant for the linking carried out in this work, especially considering the question of the development of scientific literacy of high performers. However, they must be taken into account in the future implementation of linking.

According to the PISA 2015 report (OECD, 2016) the scientific literacy tests from 2006, 2009, 2012 and 2015 use the same science performance scale and therefore the comparison of the scores across time are possible. Robitzsch et al. (2016) however showed in their work, that the change from paper pencil to computer tests could have biased the trend estimation of the German data in science. They suggest using the field test data for the trend estimation. We would follow this recommendation and suggest including field test data for the future linking of NEPS and PISA.

The authors of the above-mentioned study have also investigated the question of how far the new interactive PISA 2015 tasks could have biased the trend estimation of the German data in science. For this purpose, the German trend in science was examined based only on the old (less interactive) items from PISA 2012. The analyses show no change in the trend

estimation compared to the trend estimation based on all items. Furthermore, the dimension formed by the old items correlates closely to one with the dimension consisting of new items. It can be concluded that the science performance scale from NEPS 2010 could be linked with the science performance scale from PISA 2015.

A second limitation of the study concerns the selectivity of the tested sample. The choice of the sample can be justified by the second goal of this study, namely the investigation of the long-term effect of the SINUS program. The aim of this school program is to support teachers in teaching mathematics and science more efficiently. This could lead to a higher performance of students at these schools in science. At the same time, the choice of the sample offers the possibility to examine, which contents of this program have an influence on the students' performance. Another limitation concerns the test design, namely that the NEPS and PISA tests were administered on two days. A training effect or a decrease in motivation might be a possible consequence.

3.6.4. Practical implications

The linking of science scores from NEPS and PISA offers the possibility to classify the NEPS test values in a criterion-based reference framework of PISA. This and the longitudinal assessment in NEPS could make it possible to learn more about the conditional factors of competency acquisition and development for certain subgroups. Due to the low classification consistency on Level 1 in our data it does not seem possible to examine the factors of competency acquisition for the students on this level. However the classification consistency is high at the levels 5 and 6. This result makes it possible to learn more about the development of scientific literacy of high performers. This is particularly interesting regarding students that participated in school development programs such as SINUS.

The goal of this study was to find a linking function between the NEPS and PISA science scores. This result can be used to classify the NEPS scores within the international benchmarks of PISA. Until now, no proficiency levels are defined in NEPS. Haschke, Kampa, Hahn, and Köller (2017) developed educational standards for adults based on the NEPS scientific literacy test using the Item Descriptor Matching method. The linking function presented here provides a basis for the educational standards in NEPS with regard to the outcome that students should have achieved in secondary school. In connection with the longitudinal design of NEPS, the educational standards can be used to analyze the influence of the educational development in Germany on the competencies which are internationally considered as crucial for secondary school.

3.7. References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Second international symposium on information theory* (pp. 267-281). Budapest, Hungary: Akademai Kiado.

American Association for the Advancement of Science (2009). *Benchmarks for science literacy*. Project 2061. New York: Oxford University Press.

Bagozzi, R. P. (1981). Evaluating structural equation models with unobservable variables and measurement error: a comment. *Journal of Marketing Research, 18*(3), 375.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88*, 588–606.

Blossfeld, H.-P. (2008). *Education as a lifelong process. A proposal for a national educational panel study (NEPS) in Germany.* Part B: Theories, operationalization and piloting strategies for the proposed measurements. Bamberg: Universität Bamberg.

Brennan, R. L. (2003b). *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.

Cartwright, F. (2012). *Linking the British Columbia English examination to the OECD combined reading scale*. Prepared for the British Columbia Ministry of Education.

Cartwright, F., Lalancette, D., Mussio, J., & Xing, D. (2003). Linking provincial student assessments with national and international assessments. *Education, skills and learning, research papers,* Bd. 005.Ottawa: Statistics Canada.

Chen, W.-H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22,* 265–289.

Cohen, J. (1960). *A coefficient of agreement for nominal scales*. Educational and Psychological Measurement, *20,* 37-46.

De Ayala, R. J., Plake, B. S., & Impara, J. C. (2001). The impact of omitted responses on the accuracy of ability estimation in item response theory. *Journal of Educational Measurement, 38,* 213–234.

Dalehefte, I. M., Wendt, H., Köller, O., Wagner, H., Pietsch, M., Döring, B., . . . Bos, W. (2014). Bilanz von neun Jahren SINUS in deutschen Grundschulen: Evaluation im Rahmen der TIMSS 2011-Erhebung. *Zeitschrift für Pädagogik, 60*, 245–263.

Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for Psychologists.* Multivariate Applications Books Series. Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers (371).

Ercikan, K., & Julian, M. (2002). Classification accuracy of assigning student performance to proficiency levels: Guidelines for assessment design. *Applied Measurement in Education, 15*(3), 269–294.

Enders, C.K.,& Tofighi, D. (2008). The impact of misspecifying class-specific residual variances in growth mixture models. *Structural Equation Modeling*: *A Multidisciplinary Journal*, *15*, 75-95.

Finch, H. (2008). Estimation of item response theory parameters in the presence of missing data. *Journal of Educational Measurement, 45*, 225–245.

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.

Hambleton, R. K., Sireci, S. G., & Smith, Z. R (2009). How Do Other Countries Measure Up to the Mathematics Achievement Levels on the National Assessment of Educational Progress? *Applied Measurement in Education*, *22*(4), 376-393.

Hanushek, E. A., & Wößmann, L. (2015). *The knowledge capital of nations: education and the economics of growth*. Cambridge, MA: MIT Press.

Haschke, L., Kampa, N., Hahn, I. & Köller, O. (2017). Setting standards to a scientific literacy test for adults using the item-descriptor (ID) matching method. In S. Blömeke & J.-E. Gustafsson (Eds.), *Standard Setting in Education. The Nordic countries in an international perspective* (pp. 319 –339). Cham: Springer.

Henning, G. (1989). Meanings and implications of the principle of local independence. *Language Testing, 6*(1), 95–108.

Huynh, H., Michaels, H. R., & Ferrara, S. (1995). *Comparison of three statistical procedures to identify clusters of items with local dependency*. Annual meeting of National Council on Measurement in Education. San Francisco, CA.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: methods and practices* (2nd ed.). New York: Springer.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33,* 159–174.

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, NJ: ETS Educational Testing Service.

Lord F. M. (1974). Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika, 39,* 247–264.

Ludlow, L. H., & O'leary, M. (1999). Scoring omitted and not-reached items: Practical data analysis implications. *Educational and Psychological Measuremant, 59,* 615–630.

Martin, M. O., Mullis, I. V. S., Foy, P., & Stanco, G. M. (2012). *TIMSS 2011. International Results in Science.* Chestnut Hill MA: IEA TIMSS and PIRLS International Study Center Lynch School of Education Boston College.

Maurice, J. von, 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, Nationales Bildungspanel.

Mislevy, R. J. (1992). *Linking educational assessments: concepts, issues, methods, and prospects.* Princeton, NJ: ETS Policy Information Center.

Monseur, C., Baye, A., Lafontaine, D., & Quittre, V. (2011). *PISA test format assessment and the local independence assumption.* In IERI monograph series: Issues and methodologies in large-scale assessments, 4. Hamburg, Germany: IEA/ETS Research Institute (IERI).

Muthén, L. K., & Muthén, B. O. (1998-2012). *Mplus User's Guide.* Seventh Edition. Los Angeles, CA: Muthén and Muthén.

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

equipercentile and IRT methods. *Studies in Educational Evaluation, 47,* 58–67. doi: 10.1016/j.stueduc.2015.07.003

Neumann, K., Fischer, H. E., & Kauertz, A. (2010). From PISA to Educational Standards: the impact of large-scale assessments on science education in Germany. *International Journal of Science and Mathematics Education, 8,* 545-563.

OECD (2006). *Assessing scientific, reading and mathematical literacy: A framework for PISA 2006.* Paris: OECD.

OECD (2009). *PISA 2006 technical report.* Paris: OECD.

OECD (2013). *PISA 2012 Assessment and Analytical Framework: Mathematics, Reading, Science, Problem Solving and Financial Literacy.* Paris: OECD Publishing.

OECD (2014a). *PISA 2012 Results.* Paris: OECD Publishing.

OECD (2014b). *PISA 2012 Technical Report.* Paris: OECD Publishing.

OECD (2016). *PISA 2015 Results (Volume I): Excellence and Equity in Education.* Paris: OECD Publishing.

Pietsch, M., Böhme, K., Robitzsch, A., & Stubbe, T. C. (2009). Das Stufenmodell zur Lesekompetenz der längerübergreifenden Bildungsstandards im Vergleich zu IGLU 2006. In D. Granzer, O. Köller, A. Bremerich-Vos, M. van den Heuvel-Panhuizen, K. Reiss & G. Walther (Eds.), *Bildungsstandards Deutsch und Mathematik. Leistungsmessung in der Grundschule* (pp. 393-416). Weinheim und Basel: Beltz Verlag.

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

Pohl, S., Gräfe, L., & Rose, N. (2014). Dealing with omitted and not reached items in competence tests: Evaluating approaches accounting for missing responses in IRT models. *Educational and Psychological Measurement, 74,* 423–452.

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 (Eds.), *Competence oriented teacher training. Old research demands and new pathways* (pp. 79-96). Rotterdam: Sense Publisher.

Prenzel, M., Stadler, M. Friedrich, A., Knickmeier, K., & Ostermeier, C. (2009). *Increasing the efficiency of mathematics and science instruction (SINUS) – A large scale teacher*

professional development programme in Germany. Kiel: Leibniz-Institute for Science and Mathematics Education.

https://www.ntnu.no/wiki/download/attachments/8324749/SINUS_en_fin.pdf?version=1andmodificationDate=1251384255000

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology, 25*, 111-163.

Robitzsch, A., Lüdtke, O., Köller, O., Kröhne, U., Goldhammer, F., & Heine, J.-H. (2016). Herausforderungen bei der Schätzung von Trends in Schulleistungsstudien. *Diagnostica*. doi:10.1026/0012-1924/a000177

Rose, N., von Davier, M., & Xu, X. (2010). *Modeling nonignorable missing data with item response theory (IRT).* ETS Research Rep. no. RR-10-11. Educational Testing Service. Princeton, NJ.

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: University of Bamberg, National Educational Panel Study.

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.

Tofighi, D., & Enders, C. K. (2007). Identifying the correct number of classes in mixture models. In G. R. Hancock & K. M. Samulelsen (Eds.), *Advances in latent variable mixture models* (pp. 317-341). Greenwich, CT: Information Age.

van den Ham, A.-K., Ehmke, T., Nissen, A., & Roppelt, A. (2016). Assessments verbinden, Interpretationen erweitern? *Zeitschrift für Erziehungswissenschaft, 20*(1), 89-111. doi: 10.1007/s11618-016-0686-2

van de Vijver, F.J.R. (1998). Towards a Theory of Bias and Equivalence. In J. Harkness (Eds.), *ZUMA-Nachrichten Spezial*, *3*, pp. 41-65. Mannheim: ZUMA.

Wagner, H., Schöps, K., Hahn, I., Pietsch, M., & Köller, O. (2014). Konzeptionelle Äquivalenz von Kompetenzmessungen in den Naturwissenschaften zwischen NEPS, IQB-Ländervergleich und PISA. *Unterrichtswissenschaft*, *42*(4), 301–320.

Weinert, F. E. (2001). Concept of competence: A conceptual clarification. In D. S. Rychen & L. H. Salganik (Eds.), *Defining and Selecting Key Competencies.* Göttingen: Hogrefe and Huber Publishers.

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). *ACERConQuest Version 2: Generalised item response modelling software.* Camberwell: Australian Council for Educational Research.

4

Studie III:

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

---

*Wagner, H., Hahn, I., Schöps, K., Haag, N. & Köller, O. (unter Begutachtung). Vergleichbarkeit der naturwissenschaftlichen Kompetenz in der neunten Klasse im Nationalen Bildungspanel und im IQB-Ländervergleich 2012. *Zeitschrift für Erziehungswissenschaft*.

4.1. Einleitung

Studien wie die *Third International Mathematics and Science Study* (TIMSS; Bos, Wendt, Köller, & Selter, 2012), das *Programme for International Student Assessment* (PISA; Reiss, Sälzer, Schiepe-Tiska, Klieme, & Köller, 2016) und die *Überprüfung des Erreichens der Bildungsstandards in den Fächern Biologie, Chemie und Physik für den Mittleren Schulabschluss* (Pant, Stanat, Schröders, Roppelt, Siegle, & Pöhlmann, 2013) untersuchen regelmäßig den Stand der naturwissenschaftlichen Kompetenz der Schülerinnen und Schüler in Deutschland. Neben einem Vergleich der Kompetenzen von Schülerinnen und Schülern der verschiedenen (Bundes-)Länder stehen dabei auch die unterschiedlichen Bildungssysteme auf dem Prüfstand.

Wie die Studie von Hanushek und Wößmann (2015) zeigt, steht der in PISA gemessene Bildungserfolg der Schülerinnen und Schüler am Ende der Sekundarstufe I im Zusammenhang mit dem wirtschaftlichen Erfolg des jeweiligen Landes. Dabei stellt die naturwissenschaftliche Kompetenz im Vergleich zur mathematischen Kompetenz und zur Lesekompetenz den stärksten Prädiktor des Wirtschaftswachstums eines Landes dar. 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; Blossfeld, Schneider, & Doll, 2009) 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). Gelänge beispielsweise das Linking der NEPS-Kompetenzwerte mit den Kompetenzskalen der LV-Tests, so könnten die NEPS-Testwerte in ihren kriterialen Bezugsrahmen eingeordnet werden. Auf diese Weise wäre es möglich, Schülerinnen und Schüler zu identifizieren, die unterhalb des Mindeststandards bzw. oberhalb des Regelstandards liegen, und diese längsschnittlich im NEPS zu begleiten. So könnte man detailliertere Informationen über Bedingungs- und Risikofaktoren des Kompetenzerwerbs erhalten (vgl. van den Ham, Ehmke, Nissen, & Roppelt, 2016; Nissen, Ehmke, Köller, & Duchhardt, 2015).

In der vorliegenden Arbeit soll geprüft werden, inwiefern die im LV verwendeten Kompetenzstufenmodelle für den Mittleren Schulabschluss in den Naturwissenschaften auf die Ergebnisse des NEPS-Naturwissenschaftstests für die neunte Klassenstufe übertragen

werden können. Dafür wird als erstes die Vergleichbarkeit der im NEPS in der neunten Klasse (K9) erhobenen naturwissenschaftlichen Kompetenz mit den Testwerten des LV 2012 zur Überprüfung des Erreichens der Bildungsstandards in den Fächern Biologie, Chemie und Physik für den Mittleren Schulabschluss untersucht. Anschließend werden die Skalen des NEPS und des LV miteinander verlinkt und das erfolgte Linking wird auf Robustheit geprüft.

4.2. 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 oder 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. Im Folgenden werden in diesem Zusammenhang zunächst die Rahmenkonzeptionen der Studien NEPS und LV miteinander verglichen.

4.2.1. 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 auf Basis der Kriterien von Kolen und Brennan (2004) einander gegenübergestellt.

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 Schülerinnen und Schülern in den Fächern Biologie, Chemie und Physik am Ende der Sekundarstufe I (neunte Klassenstufe) untersucht. Auf diese Weise ist es zwar möglich, einen Trend in den erreichten Kompetenzen abzubilden, allerdings kann mit Hilfe querschnittlicher Analysen keine Aussage darüber gemacht werden, welche Ursachen für diese Leistung verantwortlich sind.

Diese Forschungslücke versucht seit 2009 die NEPS-Studie (Blossfeld, von Maurice, & Schneider, 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 Schülerinnen und Schüler unterschiedlicher Startkohorten zum Messzeitpunkt der neunten Klasse. Somit können aus den Testwerten des NEPS und des LV trotz unterschiedlicher Zielsetzungen ähnliche Schlussfolgerungen abgeleitet werden, nämlich, über welche naturwissenschaftlichen Fähigkeiten und Fertigkeiten Schülerinnen und Schüler am Ende der Sekundarstufe I verfügen.

Zielpopulationen

Die in diesem Artikel zu vergleichenden Tests der Studien NEPS (Maurice, Sixt, & Blossfeld, 2011) und LV (Siegle, Schroeders, & Roppelt, 2013) haben dieselbe Zielpopulation, nämlich Schülerinnen und Schüler der neunten Klasse.

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 Naturwissenschaftstest von NEPS für die neunte Jahrgangsstufe werden ausschließlich geschlossene Antwortformate (Schöps & Saß, 2013) verwendet, in welchen aus vier Antworten eine richtige ausgewählt werden soll (*simple multiple choice*) oder bei jeder Antwort angegeben werden soll, ob die darin enthaltene Information richtig oder falsch ist (*multiple true false*). 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 & 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 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 Minuten bearbeitet werden (Schöps & 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, Schöps, Hahn, Pietsch, & Köller, 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 Kalibrierung der Items zur Bestimmung von Itemschwierigkeiten als auch bei der Schätzung von Personenparametern als fehlend betrachtet (Pohl & Carstensen, 2012). Im LV werden dagegen alle fehlenden Werte in beiden genannten Schritten als falsch kodiert. Diesem Vorgehen liegt die Annahme zugrunde, dass Schülerinnen und Schüler vor allem solche Aufgaben nicht bearbeiten, die sie nicht lösen können. Wie eine Reihe von Studien (De Ayala, Plake, & Impara, 2001; Finch, 2008; Lord, 1974; Ludlow & O'leary, 1999; Pohl, Gräfe, & Rose, 2014; Rose, von Davier, & Xu, 2010) gezeigt hat, kann die Behandlung der fehlenden Werte als falsch die Schätzung der Parameter verzerren und die Reliabilität eines Tests überschätzen, wenn diese Annahme nicht zutrifft. Folglich kann sich der Unterschied im Umgang mit fehlenden Werten auf die Vergleichbarkeit der Testwerte aus NEPS und dem LV auswirken.

Ein weiterer Unterschied zwischen den zu vergleichenden Tests, der sich aus den Rahmenkonzeptionen der Studien ergibt, betrifft die Dimensionalität. Da die Tests des LV die Kompetenzen fächerspezifisch (Biologie, Chemie oder Physik) und inhaltsspezifisch (Fachwissen oder Erkenntnisgewinnung) erheben, werden die Personenparameter hier in einem mehrdimensionalen Rasch-Modell geschätzt. NEPS versteht die naturwissenschaftliche Kompetenz als ein eindimensionales Konstrukt. Dementsprechend wird die Fähigkeit der Personen in einem eindimensionalen Schätzverfahren ermittelt.

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

Der Vergleich der in der Abbildung 1 dargestellten Rahmenkonzeptionen des NEPS-Naturwissenschaftstests und der Tests zur Überprüfung des Erreichens der Bildungsstandards in Biologie, Chemie und Physik zeigt eine große Überschneidung der Studien hinsichtlich ihrer Inhaltsbereiche.

	NEPS	**LV 2012**
Inhaltsbereiche	*Inhaltsbezogene Komponenten:* Entwicklung → Systeme → Wechselwirkungen → Stoffe ⇢ *Prozessbezogene Komponenten:* Naturwissenschaftliche Denk- und Arbeitsweisen; ⇢	*Umgang mit Fachwissen:* Entwicklung, Struktur und Funktion, System (Biologie); System, Materie, Wechselwirkung, Energie (Physik); Struktur der Materie, Struktur-Eigenschafts-Bez., chemische Reaktion (Chemie); *Erkenntnisgewinnung:* Naturwissenschaftliche Untersuchungen, Naturwissenschaftliche Modellbildung, Wissenschaftstheoretische Reflexion; *Kommunikation und Bewertung* (sind keine Bestandteile des LV 2012)
Kontexte	Gesundheit, Umwelt, Technologie	Nicht vorhanden
Teil-kompetenzen	Nicht vorhanden	*Komplexität:* ein Fakt, zwei Fakten, ein Zusammenhang, zwei Zusammenhänge, übergeordnetes Konzept; *Kognitive Prozesse:* Reproduzieren, Selegieren, Organisieren, Integrieren.
Quelle	(Blossfeld, 2008); (Blossfeld et al., 2009); (Hahn et al., 2013)	(Kauertz, Fischer, Mayer, Sumfleth & Walpuski, 2010); (KMK, 2005a, 2005b, 2005c)

Abb. 1 Übersicht über die Rahmenkonzeptionen des NEPS-Naturwissenschaftstests und der LV-Tests zur Überprüfung des Erreichens der Bildungsstandards in Biologie, Chemie und Physik. Ein durchgezogener Pfeil bedeutet eine Zuordnung zum entsprechenden Äquivalent, ein gestrichelter Pfeil bedeutet eine Zuordnung auf Grundlage der inhaltlichen Überschneidung.

So finden sich für die NEPS-Konzepte *Entwicklung*, *Systeme* und *Wechselwirkungen* jeweils ein oder mehrere entsprechende Äquivalente in der Rahmenkonzeption der Bildungsstandards (unterstrichen und mit durchgezogenen Pfeilen gekennzeichnet). Das Konzept *Stoffe* wird zwar als fächerbindend bezeichnet, schwerpunktmäßig wird dieser Inhaltsbereich im NEPS jedoch der *Chemie* zugeordnet (gestrichelter Pfeil).

Außer den inhaltsbezogenen Komponenten unterscheiden die Rahmenkonzeptionen des NEPS und der Bildungsstandards die prozessbezogenen Komponenten. Hier ähneln sich die *naturwissenschaftlichen Denk- und Arbeitsweisen* des NEPS und die *Erkenntnisgewinnung* der Bildungsstandards. Allerdings zeigt der Vergleich der Rahmenkonzeptionen neben einer großen inhaltlichen Überschneidung auch einen Unterschied in der Definition der operationalisierten Konstrukte, nämlich in der Anzahl der Komponenten zur Erfassung der naturwissenschaftlichen Kompetenz. Während die NEPS-Rahmenkonzeption *nur* die Wissenskomponenten betrachtet, werden bei den Bildungsstandards (zusätzlich zu den Wissensbereichen) die Bereiche *Komplexität* und *Kognitive Prozesse* unterschieden. Dagegen verzichten die Bildungsstandards - anders als das NEPS - auf eine explizite Formulierung der Kontexte, in denen die Inhalte der Items abgefragt werden.

Ein weiterer Unterschied zwischen den Rahmenkonzeptionen besteht in der fachlichen Spezifität des zu untersuchenden Konstrukts. Mit den Tests des LV soll 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 Rahmenkonzeption des NEPS-Naturwissenschaftstests hingegen orientiert sich bei der Untersuchung der Kompetenz am Literacy-Konzept (Bybee, 1997; OECD, 2006) im Sinne der funktionalen Grundbildung, die in den Alltagskontext eingebunden ist. Dementsprechend verzichtet NEPS bei der Definition der Inhalte des Naturwissenschaftstests auf eine enge Orientierung an Lehrplänen. Ähnlich dem Vorgehen in der PISA-Studie wird die naturwissenschaftliche Kompetenz in der NEPS-Rahmenkonzeption als ein eindimensionales Konstrukt mit zwei Facetten *inhaltsbezogene Komponenten* und *prozessbezogene Komponenten* definiert, die in der PISA-Rahmenkonzeption dem *naturwissenschaftlichen Wissen* und dem *Wissen über die Naturwissenschaften* entsprechen. An dieser Stelle kann also festgehalten werden, dass sich die Rahmenkonzeptionen der Studien in der Breite der Definition ihrer Konstrukte unterscheiden. Aber wie verhalten sich die Tests beider Studien zueinander auf der Aufgabenebene?

Der Aspekt der 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. Nissen et al., 2015; Pietsch, Böhme, Robitzsch & Stubbe, 2009; van den Ham et al., 2016). Untersuchungen zur konzeptionellen Äquivalenz beider Tests konnten zeigen, dass der NEPS-Test auf der Aufgabenebene große Gemeinsamkeiten mit der Rahmenkonzeption der Bildungsstandards aufweist. Gleichzeitig zeigte sich, dass die Aufgaben des NEPS-Tests nicht alle Inhalte der Bildungsstandards-Rahmenkonzeption abdecken (Wagner et al., 2014).

Die Untersuchung der dimensionalen Äquivalenz und der Skalenäquivalenz soll nun zeigen, wie stark die naturwissenschaftliche Kompetenz im NEPS-Test mit den fächerspezifischen Kompetenzwerten im LV zusammenhängt. Können die Testwerte durch eine gemeinsame Dimension abgebildet werden oder ist es notwendig, für jeden Test oder gar jeden Kompetenzbereich eine eigene Dimension anzunehmen? Haben die Werte im NEPS-Test und in den LV-Tests eine ähnliche Verteilung hinsichtlich des Mittelwertes, des Exzesses und der Schiefe?

4.2.2. 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*. Zur Übertragung der Kompetenzstufenmodelle des LV auf die NEPS-Kompetenzwerte kommen lediglich das Vertical Scaling oder das Equating in Frage, weil nur diese Methoden den Vergleich der individuellen Zuordnungen zu den Kompetenzstufen ermöglichen. Allerdings setzen diese Methoden die höchsten Standards hinsichtlich der Vergleichbarkeit der zu verlinkenden Skalen. Zum Beispiel müssen die Tests für die Verknüpfung anhand des Vertical Scaling ähnliche Schlussfolgerungen, Konstrukte und Messbedingungen aufweisen, können sich aber in ihren Populationen unterscheiden. Das Linking zweier Tests mit dem Vertical Scaling wird oft im Rahmen der *Item-Response-Theorie* (IRT) durchgeführt. Für die Verknüpfung mit Hilfe des Equating müssen die Tests in Bezug auf alle Merkmale des Ansatzes von Kolen und Brennan (2004) ähnlich sein. Wenn die Skalen zweier Tests (A und B) mit der Equating-Methode verknüpft sind, können die Werte des Tests A aus den Werten des Tests B vorhergesagt werden und umgekehrt. Allerdings hat diese Methode auch einen Nachteil: die Sensitivität gegenüber den Unregelmäßigkeiten in der Verteilung der Testergebnisse. Das

Equating kann beispielsweise nicht für den Wertebereich über dem höchsten sowie unter dem niedrigsten beobachteten Wert bestimmt werden. Dieses Problem kann durch eine Prä- oder Postglättung der Skalen, die mit der Equating-Methode verlinkt wurden, gelöst werden (Livingston, 2004).

Seit PISA 2000 gibt es Bestrebungen nationale und internationale Assessments zu verlinken (Cartwright, 2012; Cartwright, Lalancette, Mussio & Xing, 2003; Hambleton, Sireci & Smith, 2009; National Center for Educational Statistics, 2013; Nissen et al., 2015; Pietsch et al., 2009). Jedoch kann auch die Verlinkung der nationalen Assessments gegenseitig zu einer Erweiterung der Interpretationen beitragen. Van den Ham et al. (2016) übertrugen beispielsweise die Ergebnisse des NEPS-Mathematiktests für die neunte Klassenstufe auf die Kompetenzstufen des LV. Die gemeinsame Skalierung der Testwerte zeigt, dass beide Skalen trotz des hohen latenten Zusammenhangs von $r = 0{,}92$ nicht ohne Weiteres austauschbar sind. Das Linking erfolgte via Equipercentile Equating und zeigte für die Gesamtpopulation vergleichbare Verteilungen der Schülerinnen und Schüler auf die Kompetenzstufen des LV. Dagegen lag die prozentuale Übereinstimmung der Studien (*PÜ* = Prozentsatz der Schüler, die derselben Kompetenzstufe zugeordnet sind) auf der Individualebene bei *PÜ* = 48% und wies auf beträchtliche Unterschiede in der Individualzuordnung hin. Auch Cohens Kappa lag nur bei $k = 0{,}31$. Aus diesem Grund raten die Autoren von studienübergreifenden Schlussfolgerungen auf Individualniveau ab.

In einer weiteren Studie wurden verschiedene Linking-Methoden verglichen, indem die Mathematik-Werte des TIMSS 2011-Tests für die vierte Klasse mit den entsprechenden Werten des NEPS 2010-Tests für die fünfte Klasse verknüpft wurden (Nissen et al., 2015). Die Verlinkung der beiden Skalen mit dem Equipercentile Equating und dem IRT-Linking zeigte eine zufriedenstellende Klassifizierungskonsistenz gegenüber den internationalen TIMSS-Benchmarks (44% bzw. 36%). Darüber hinaus stellten die Autoren 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. Aufgrund der geringfügigen Unterschiede in der Klassifizierung zu den internationalen *Benchmarks* (Kompetenzstufen) empfehlen die Autoren lediglich die Schlussfolgerungen über die Gruppen von Schülern und nicht über die Einzelpersonen.

Die Übersicht der Linking-Studien zeigt die Relevanz dieser Arbeiten im Bereich der Kompetenzmessung. Darüber hinaus liefert die letzte Studie Erkenntnisse über die Vorteile des Equipercentile Equating gegenüber dem IRT-Linking.

4.3. Fragestellungen

Das Ziel dieser Arbeit liegt in der Verlinkung der Naturwissenschaftsskalen 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) Wie hoch ist der Zusammenhang zwischen den Naturwissenschaftsskalen des NEPS und des LV?
(2) Inwiefern messen die Naturwissenschaftsskalen des NEPS und des LV dasselbe Konstrukt?

Skalenäquivalenz

(3) Inwiefern zeigen sich bei den Naturwissenschaftsskalen des NEPS und des LV vor und nach dem Linking ähnliche Verteilungen?
(4) Wie hoch ist die Klassifikationskonsistenz der verlinkten Werte hinsichtlich der Zuordnung zu den Kompetenzstufen der LV-Tests?

4.4. Methode

4.4.1. 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 & Ostermeier, 2006; Prenzel, Stadler, Friedrich, Knickmeier, & Ostermeier, 2009). Aus diesem Grund besteht die Stichprobe aus 80 SINUS-Schulen (1728 Schülerinnen und Schülern). 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 in der vorliegenden Studie ähnlich wie in den Haupterhebungen der Studien vom DPC durchgeführt.

Insgesamt haben 678 Schülerinnen und Schüler (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 bearbeitet, die auf 12 Blöcke verteilt und im Multi-Matrix-Design dargeboten wurden, so dass jede Schülerin und jeder Schüler nur einen Teil der Aufgaben (3 Blöcke à 20 Minuten) bearbeitete. Die Aufgabenblöcke entstammen 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 Schülerinnen und Schüler 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 Minuten dargeboten. Zur Reduktion des Ermüdungseffekts wurden die NEPS- und LV-Testblöcke in ihrer Position rotiert.

4.4.2. Scoring und Umgang mit den Daten

Die erhobenen Daten wurden in Anlehnung an die Manuale der Studien NEPS (Schöps & Saß, 2013) und LV (Hecht, Roppelt, & Siegle, 2013) kodiert und im Rahmen der IRT mit der Software ConQuest (Wu, Adams, Wilson, & Haldane, 2007) analysiert. Zur Sicherstellung der Vergleichbarkeit der Linking-Ergebnisse mit den Hauptstudien wurden fehlende Werte sowohl bei der Untersuchung der Äquivalenz der Tests als auch beim Linking studienspezifisch kodiert. Zur Veranschaulichung der Unterschiede in den Anteilen fehlender Werte am Gesamttest wurde eine Analyse der fehlenden Werte durchgeführt.

Die Daten des LV wurden, angelehnt an das Vorgehen in der Hauptstudie (Biologie: Mayer, Wellnitz, Klebba, & Kampa, 2013; Chemie: Walpuski, Sumfleth, & Pant, 2013; Physik: Kauertz, Fischer, & Jansen, 2013), unter Berücksichtigung des Inhaltsbereichs (Fachwissen oder Erkenntnisgewinnung) und des Faches (Biologie, Chemie oder Physik) mehrdimensional skaliert. Die Skalierung der NEPS-Daten erfolgte im Rahmen des (eindimensionalen) Rasch-Ansatzes (Schöps & Saß, 2013).

Die erste Fragestellung bezüglich des Zusammenhangs der Naturwissenschaftswerte des NEPS und des LV wurde mit Hilfe der messfehlerkorrigierten Korrelationen des NEPS-Tests mit den einzelnen Inhaltsbereichen des LV analysiert.

Die zweite Fragestellung zur Äquivalenz der Konstrukte, die dem NEPS-Test und den LV-Tests zugrunde liegen, wurde anhand einer konfirmatorischen Faktorenanalyse mit der Software Mplus (Muthén & Muthén, 2012) untersucht. Aus theoretischer Sicht sind zur Abbildung der Daten aus NEPS und dem LV folgende Faktorenstrukturen denkbar:

- 7 Dimensionen: 6 Dimensionen zur Abbildung der Inhaltsbereiche im LV (Umgang mit Fachwissen bzw. Erkenntnisgewinnung in Biologie, Chemie oder Physik) & NEPS
- 4 Dimensionen: 3 Dimensionen zur Abbildung des Faches im LV (Biologie, Chemie oder Physik) & NEPS
- 3 Dimensionen: 2 Dimensionen zur Abbildung des Kompetenzbereiches im LV (Fachwissen bzw. Erkenntnisgewinnung) & NEPS
- 2 Dimensionen: 2 Dimensionen zur Abbildung des Kompetenzbereiches im LV und NEPS unabhängig der Studienzugehörigkeit der Items (inhaltsbezogene Komponente & prozessbezogene Komponente)
- 2 Dimensionen: 2 Dimensionen zur Abbildung der Studienzugehörigkeit der Items (LV & NEPS)
- 1 Dimension: 1 Dimension zur Abbildung des Konstrukts *Naturwissenschaftliche Kompetenz*.

Die Ergebnisse konfirmatorischer Faktorenanalysen können mit Hilfe informations-theoretischer Indizes bewertet und interpretiert werden. 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 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 und eignet sich insofern besonders gut für die Auswahl der Modelle bei größeren Stichproben ($N > 300$; Dziak, Coffman, Lanza, & Li, 2012). 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 wird das Modell mit dem kleinsten Wert des betreffenden Maßes ausgewählt. Raftery (1995) schlägt vor, den Unterschied von $\Delta BIC \geq 10$ als sehr starken Beleg für eine bessere Modellanpassung zu interpretieren.

4.4.3. Linking

Für die Untersuchung der Skalenäquivalenz 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. Der Vergleich der Verteilungen aus NEPS und dem LV erfolgte getrennt für jeden Inhaltsbereich des LV anhand des Mittelwertes, der Schiefe und des Exzesses.

Nissen et al. (2015) haben gezeigt, dass die Klassifikationskonsistenz der mit dem Equipercentile Equating verlinkten Mathematik-Werte des TIMSS 2011-Tests für die vierte Klasse mit den entsprechenden Werten des NEPS 2010-Tests für die fünfte Klasse höher ist als auf Grundlage des IRT-Linkings. Daher wurde die Verknüpfung der NEPS- und LV-Werte in der vorliegenden Studie mittels des Equipercentile Equating (vgl. Cartwright, 2012, Nissen et al., 2015, van den Ham et al., 2016) durchgeführt. Dieses Verfahren basiert auf der Idee, dass die Testwerte aus zwei Tests mit dem gleichen Perzentilrang als gleichwertig deklariert werden (Kolen & Brennan, 2004). Wenn zum Beispiel 9,74% der Schülerinnen und Schüler im NEPS-Test 427 Punkte oder weniger erreichten und 9,74% der Schülerinnen und Schüler im LV-Test 461 Punkte oder weniger erreichten, dann werden die Punktwerte 427 und 461 als äquivalent erklärt. Aufgrund der Sensitivität des Equipercentile Equating für Unregelmäßigkeiten in der Verteilung von Testergebnissen (Livingston, 2004), wurden die NEPS-Äquivalente in der LV-Metrik mit einem Wert von 0,3 nachgeglättet. Das Linking erfolgte getrennt für jeden der fünf PVs mit der Computer-Software LEGS (Brennan, 2003).

Im letzten Schritt wurden die untersuchten Schülerinnen und Schüler 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 (Fachwissen bzw. 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 Schülerinnen und Schüler auf die Kompetenzstufen des LV auf Grundlage der Tests aus beiden Studien gleich ist.

In einem weiteren Schritt wurde geprüft, wie hoch die Übereinstimmung der Tests in der Zuordnung der Schülerinnen und Schüler zu den Kompetenzstufen auf der individuellen Ebene ist. Dazu wurde die prozentuale Übereinstimmung zwischen den zu vergleichenden Tests des NEPS und des LV bestimmt. Als weiteres Maß für die Beurteilung der Übereinstimmung in der Kompetenzstufenzuordnung wurde Cohens Kappa berechnet. Für die

Interpretation der Werte von Cohens Kappa schlagen Landis und Koch (1977) folgende Kriterien vor: „0,21-0,40“ ausreichende Übereinstimmung, "0,41-0,60" moderate Übereinstimmung, "0,61-0,80" substantielle Übereinstimmung und "0,81-1,00" nahezu perfekte Übereinstimmung.

4.5. Ergebnisse

4.5.1. Dimensionale Äquivalenz

Beim Vergleich der Rahmenkonzeptionen des NEPS und des LV wurde ein Unterschied im Umgang mit fehlenden Werten festgestellt. Im NEPS werden nur die bearbeiteten Aufgaben zur Schätzung der Personenfähigkeit herangezogen. Im LV gehen dagegen alle zur Bearbeitung vorgelegten Aufgaben in die Schätzung der Personenfähigkeit ein. Fehlende Werte werden dabei als falsch kodiert, da angenommen wird, dass Schülerinnen und Schüler die ausgelassenen Items nicht korrekt beantworten könnten. Dieser Umstand kann sich auf die Vergleichbarkeit der Werte der beiden Studien auswirken.

Analyse der fehlenden Werte

Um den Effekt des unterschiedlichen Umgangs mit fehlenden Werten im NEPS und LV abzuschätzen, wurde eine Analyse der fehlenden Werte durchgeführt. Die in Abbildung 2 dargestellten Anteile der fehlenden Antworten am Gesamttest zeigen, dass es in den Kategorien *missing not reached* (Aufgaben, die am Ende des Tests nicht bearbeitet wurden) und *nicht interpretierbar* (Aufgaben mit nicht interpretierbaren Antworten) signifikante Unterschiede ($p < 0,05$) zwischen den Studien bestehen.

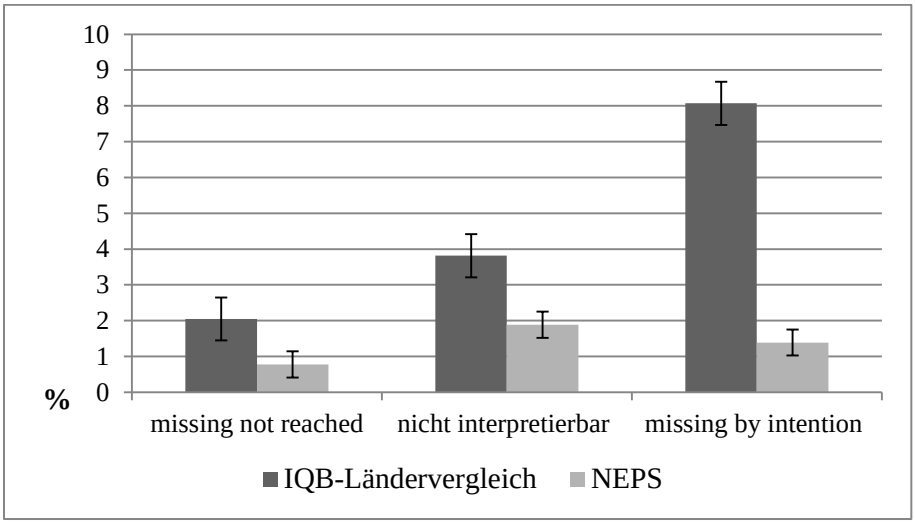

Abb. 2 Prozentualer Anteil fehlender Antworten am Gesamttest im NEPS und LV

Dieser Unterschied fällt bei der Betrachtung der übersprungenen Aufgaben (*missing by intention*) noch größer aus. Eine Möglichkeit für die Erklärung dieses Unterschiedes liegt im eingesetzten Antwortformat: In den LV-Tests gibt es viele offene Aufgaben, während im NEPS-Test nur geschlossene Formate vorkommen.

Tatsächlich zeigt Abbildung 3, dass sich der Anteil fehlender Werte in den untersuchten LV-Aufgaben abhängig vom eingesetzten Antwortformat signifikant unterscheidet. Während lediglich 1,6% der Aufgaben mit geschlossenem Antwortformat keine gültige Antwort aufweisen, liegt dieser Anteil in den halboffenen und offenen Aufgaben bei 14,3% und 21,6%. Da die halboffenen und offenen Aufgaben sowohl am Gesamttest des LV 2012 als auch an der Gesamtzahl der in dieser Studie eingesetzten Aufgaben über 40 % ausmachen, kann in Verbindung mit dem unterschiedlichen Umgang mit fehlenden Werten der Studien NEPS und LV ein starker Effekt auf die Vergleichbarkeit ihrer Ergebnisse sowohl in der vorliegenden Studie als auch in den Hauptstudien angenommen werden.

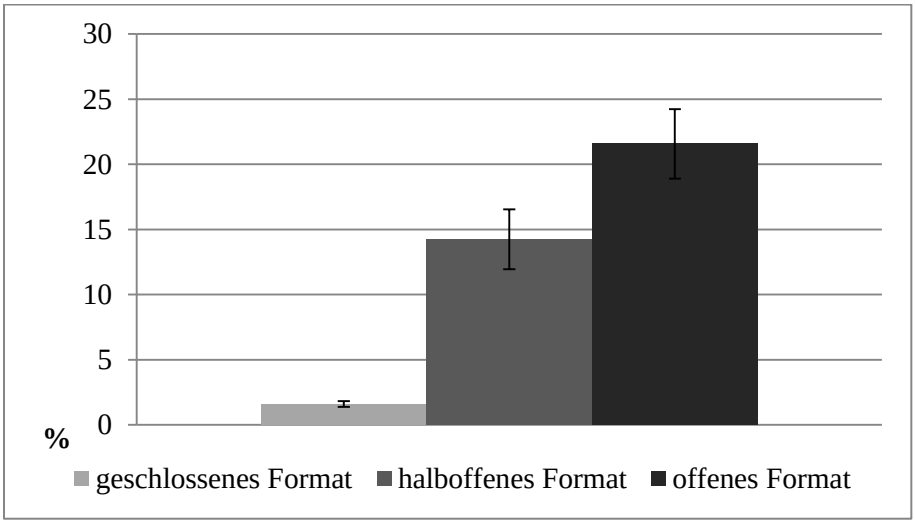

Abb. 3 Prozentualer Anteil fehlender Werte in den Aufgaben des LV in der Abhängigkeit vom Antwortformat

Korrelationen

Die erste Fragestellung zielt auf die Untersuchung der messfehlerkorrigierten Zusammenhänge der NEPS-Testwerte mit den Bildungsstandards-Inhaltsbereichen ab. Die in Tabelle 1 abgebildeten korrigierten Korrelationen liegen im Bereich von 0,78 bis 0,82 und unterscheiden sich nicht signifikant voneinander. Dementsprechend teilt der NEPS-Test 61%-67% seiner Varianz mit den jeweiligen Bildungsstandards-Tests. Messfehlerkorrigierte Korrelationen der Inhaltsbereiche des LV-Tests liegen im Bereich von 0,8 bis 0,91 und weisen somit nur eine geringe Abweichung von den Korrelationen der LV-Tests mit dem NEPS-Test auf. Die PV-Reliabilität liegt für den NEPS-Test bei 0,85 und für die LV-Tests

zwischen 0,75 und 0,79. Somit verfügen die Testwerte beider Studien über eine ausreichende Reliabilität.

Tab. 1 Latente Korrelationen der NEPS-Testwerte mit den Testwerten der Bildungsstandards-Inhaltsbereiche

	Fachwissen	Erkenntnisgewinnung
Biologie	0,78	0,78
Chemie	0,82	0,79
Physik	0,79	0,78

Dimensionale Struktur

Das Ziel der zweiten Fragestellung liegt in der Untersuchung der dimensionalen Struktur der NEPS- und LV-Daten. Hierfür wurden mehrere Faktorenanalysen mit einer unterschiedlichen Anzahl von Dimensionen gerechnet (s. Abschnitt 3.3).

Die informationstheoretischen Indizes (AIC, BIC und SABIC) in der Tabelle 2 fallen für das Modell mit drei Dimensionen am kleinsten aus und zeigen somit im Vergleich mit den anderen Modellen die beste Passung. Die drei Dimensionen dieses Modells sind Fachwissen und Erkenntnisgewinnung des LV sowie eine NEPS-Dimension. Alle drei Dimensionen korrelieren hoch miteinander: Die messfehlerkorrigierte Korrelation zwischen Fachwissen und Erkenntnisgewinnung im LV liegt bei $r = 0{,}86$. Weiterhin korrelieren das Fachwissen (FW) und die Erkenntnisgewinnung (EG) im LV mit der NEPS-Dimension zu $r_{FW/NEPS} = 0{,}79$ und $r_{EG/NEPS} = 0{,}82$.

Die Ergebnisse der Faktorenanalyse zeigen außerdem, dass die im Rahmen der vorliegenden Studie erhobenen Daten der Unterscheidung von sechs Inhaltsbereichen des LV und einer zusätzlichen NEPS-Dimension widersprechen. Dieses Modell hat von den getesteten Modellen zwar die kleinste Devianz und weist somit die geringste Abweichung von den erhobenen Daten auf, allerdings macht die Anzahl der Parameter dieses Modells im Vergleich zum Modell mit drei Dimensionen weniger geeignet zur Beschreibung der Datenstruktur. Gleichzeitig erweist sich die Vereinfachung der Faktorenstruktur auf zwei oder gar eine Dimension als mit den erhobenen Daten nicht vereinbar. Insgesamt zeigt die Differenz des BIC mit $\Delta BIC \geq 10$ eine starke Evidenz für die bessere Anpassung des dreidimensionalen Modells und widerspricht somit der Annahme, dass die Testwerte von NEPS und LV durch eine Dimension abgebildet werden können.

Tab. 2 Ergebnisse konfirmatorischer Faktorenanalysen der NEPS- und LV-Daten

	$N_{Parameter}$	Devianz	AIC	BIC	SABIC
7 Dimensionen (6 Inhaltsbereiche des LV & NEPS)	225	*51.712*	52.162	53.179	52.466
4 Dimensionen (Biologie, Chemie, Physik & NEPS)	207	51.738	52.152	53.088	52.431
3 Dimensionen (Fachwissen, Erkenntnisgewinnung & NEPS)	203	51.737	*52.143*	*53.061*	*52.417*
2 Dimensionen (inhaltsbezogene & prozessbezogene Komponenten)	200	51.873	52.273	53.177	52.543
2 Dimensionen (LV & NEPS)	200	51.778	52.178	53.082	52.448
1 Dimension (Naturwissenschaftliche Kompetenz)	*198*	51.934	52.330	53.225	52.597

Legende: AIC: Akaike information criterion; BIC: Bayesian information criterion; SABIC: Sample-Size Adjusted BIC.

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

Die Untersuchung der dimensionalen Äquivalenz hat gezeigt, dass den NEPS- und LV-Testwerten ein dreidimensionales Modell zugrunde liegt, was einen Vergleich der Verteilungen der Personenfähigkeiten im NEPS und im LV anhand dieses Modells nahelegt. Da allerdings die Untersuchung der Skalenäquivalenz und insbesondere das Linking auf den Vorgaben der Hauptstudien basieren, welche im LV ein sechsdimensionales Modell vorsehen, wurden die nachfolgenden Schritte an dieses Vorgehen angelehnt. Dies geschah, obwohl die Dimensionsanalysen die sechsdimensionale Behandlung der Testwerte im LV in Frage stellen.

Skalenäquivalenz im Bereich des Umgangs mit Fachwissen des LV

Im ersten Schritt der Untersuchung der Skalenäquivalenz und damit der dritten Fragestellung werden die Verteilungen der Personenschätzungen im NEPS und im LV zunächst miteinander verglichen. In Tabelle 3 werden die Indizes zur Beschreibung der Verteilung von Personenfähigkeiten im NEPS und den Personenfähigkeiten im Umgang mit Fachwissen im

LV zusammengefasst. Die Ergebnisse zeigen einen signifikanten Unterschied ($p < 0,05$) in den NEPS- und LV-Mittelwerten bei jeweils gleich ausgeprägter Schiefe und Exzess.

Tab. 3 Verteilung der Personenfähigkeit im NEPS, den Tests zum Umgang mit Fachwissen im LV und im NEPS in der LV-Metrik

	M (*SE*)	*SD*	Schiefe	*SD*	Exzess	*SD*
NEPS	540 (3,13)	81	- 0,01	0,09	- 0,37	0,19
Biologie Fachwissen (BF)	589 (3,65)	95	- 0,11	0,09	- 0,38	0,19
NEPS in der BF-Metrik	589 (3,82)	99	- 0,10	0,09	- 0,39	0,19
Chemie Fachwissen (CF)	601 (3.13)	81	- .07	0.09	- .07	0.19
NEPS in der CF-Metrik	601 (3.21)	83	- .09	0.09	- .24	0.19
Physik Fachwissen (PF)	610 (3.08)	80	- .17	0.09	- .21	0.19
NEPS in der PF-Metrik	608 (3.20)	83	- .12	0.09	- .24	0.19

Legende: *M*: Mittelwert; *SE*: Standardfehler; *SD*: Standardabweichung.

Der Unterschied in den Mittelwerten der Studien weist auf eine unterschiedliche Bewertung der Leistung von Schülerinnen und Schülern in Abhängigkeit vom bearbeiteten Test hin. Werden den Personen beide Tests vorgelegt, erreichen sie im Durchschnitt im LV-Test einen höheren Fähigkeitswert als im NEPS-Test, wenn die Items auf die Schwierigkeiten der jeweiligen Hauptstudie fixiert sind. Durch die Verlinkung werden die beiden Verteilungen einander angeglichen. Auch der Unterschied von 49 bis 70 Punkten (je nach Fach) zwischen den Mittelwerten verschwindet nach der Verlinkung von beiden Skalen, wie Tabelle 3 zeigt.

Im nächsten Schritt erfolgte die Einordung der Schülerinnen und Schüler auf die Kompetenzstufen des LV auf der Grundlage der LV-Werte und der NEPS-Testwerte in der LV-Metrik. Wie Abbildung 4 zeigt, sind die verlinkten NEPS- und LV-Skalen untereinander äquivalent. 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.*). In beiden Studien erreichen mindestens 97% der Schülerinnen und Schüler den Mindeststandard (Kompetenzstufe 2). Weiterhin erfüllen mindestens 85% der gesamten Stichprobe die Anforderungen des Regelstandards (Kompetenzstufe 3). Bei mindestens 16 % (im Fach Physik sogar 28%) der untersuchten Schülerinnen und Schüler reicht die Leistung sogar für den Optimalstandard (Kompetenzstufe 5) aus.

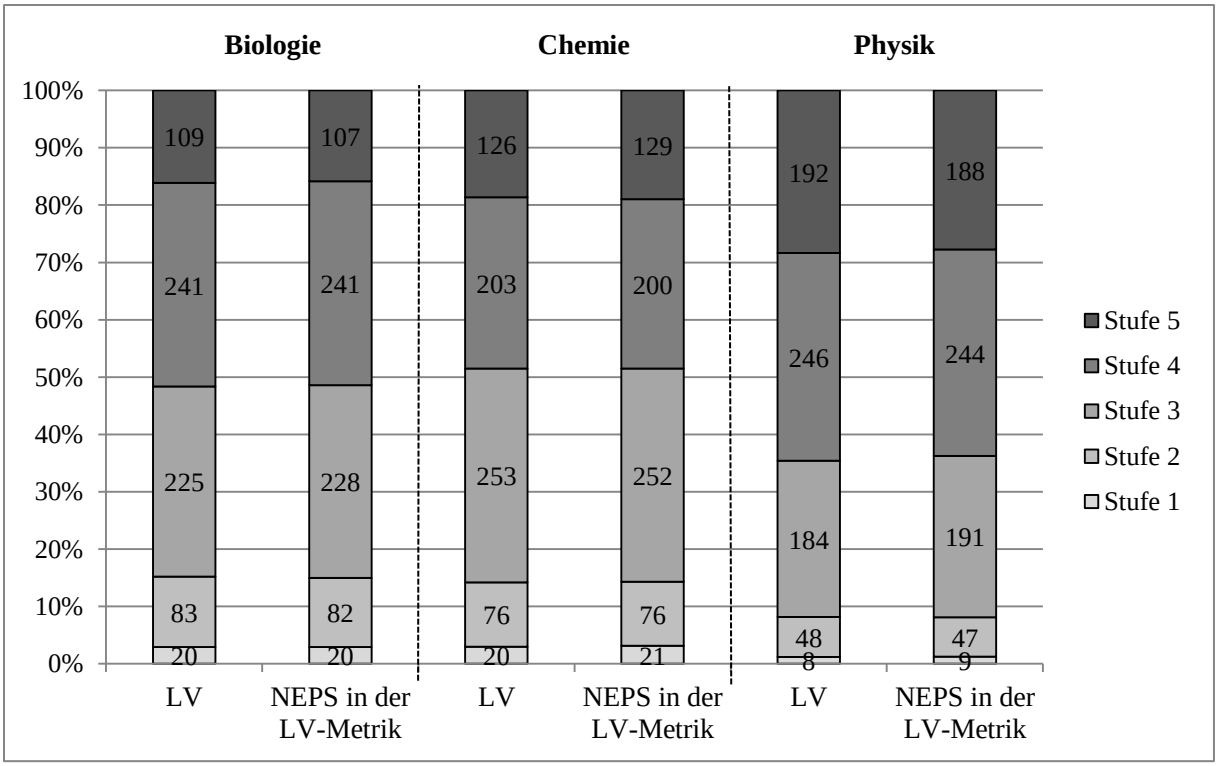

Abb. 4 Einordnung der Schülerinnen und Schülern auf die Kompetenzstufen der Tests zum Umgang mit Fachwissen im LV (die Anzahl der Personen wurde auf den Balken abgetragen)

Individuelle Zuordnung zu den Kompetenzstufen im Fachwissen Biologie

Nachdem im vorhergehenden Abschnitt die Frage der Äquivalenz in der Kompetenzstufenverteilung der Studien NEPS und LV geklärt wurde, wird nun im nächsten Schritt die (prozentuale) Übereinstimmung (*PÜ*) der beiden Studien hinsichtlich der individuellen Zuordnung der Schülerinnen und Schüler auf die Kompetenzstufen des LV berechnet. Tabelle 4 zeigt, dass diese im Umgang mit dem Fachwissen Biologie zwischen 47% (Kompetenzstufe 1) und 59% (Kompetenzstufe 5) variiert. Im Mittel liegt die *PÜ* bei 53%. Der über fünf PVs erfasste Cohens-Kappa-Wert beträgt k = 0,38 und kann als ausreichend bewertet werden.

Tab. 4 Prozentuale Übereinstimmung zwischen den Studien NEPS und LV hinsichtlich der Kompetenzstufenzuordnung im Umgang mit Fachwissen Biologie

		Kompetenzstufen des LV				
		1	2	3	4	5
NEPS in der LV-Metrik	1	**47**	9	1		
	2	30	**47**	14	2	
	3	23	40	**55**	26	4
	4		4	27	**57**	37
	5			3	15	**59**
	Total	100	100	100	100	100

Individuelle Zuordnung zu den Kompetenzstufen im Fachwissen Chemie

In Tabelle 5 ist die prozentuale Übereinstimmung der Studien NEPS und LV hinsichtlich der individuellen Zuordnungen von Schülerinnen und Schülern zu den Kompetenzstufen im Umgang mit dem Fachwissen Chemie zusammengefasst. Insgesamt werden 45% (Kompetenzstufe 2) bis 67% (Kompetenzstufe 5) der Schülerinnen und Schüler aufgrund des NEPS-Tests und des Umgangs mit dem Fachwissen Chemie im LV derselben Kompetenzstufe zugeordnet. Die durchschnittliche PÜ liegt bei knapp 58% und das durchschnittliche Cohens-Kappa erreicht den Wert von k = 0,41. Somit liegt im Kompetenzbereich *Umgang mit Fachwissen Chemie* zwischen den Studien NEPS und LV eine moderate Übereinstimmung vor.

Tab. 5 Prozentuale Übereinstimmung zwischen den Studien NEPS und LV hinsichtlich der Kompetenzstufenzuordnung im Umgang mit Fachwissen Chemie

		Kompetenzstufen des LV				
		1	2	3	4	5
NEPS in der LV-Metrik	1	**63**	10			
	2	31	**45**	12	2	
	3	6	44	**62**	28	4
	4		1	24	**51**	29
	5			2	19	**67**
	Total	100	100	100	100	100

Individuelle Zuordnung zu den Kompetenzstufen im Fachwissen Physik

Wie aus Tabelle 6 ersichtlich ist, variiert die prozentuale Übereinstimmung der beiden Studien hinsichtlich der individuellen Kompetenzstufenzuordnung im Umgang mit dem Fachwissen Physik zwischen 37% (Kompetenzstufe 2) und 68% (Kompetenzstufe 5) und erreicht im Durchschnitt $PÜ$ = 52%. Cohens Kappa liegt bei k = 0,4 und zeigt somit eine moderate Übereinstimmung zwischen den Studien.

Tab. 6 Prozentuale Übereinstimmung zwischen den Studien NEPS und LV hinsichtlich der Kompetenzstufenzuordnung im Umgang mit Fachwissen Physik

		Kompetenzstufen des LV				
		1	2	3	4	5
NEPS in der LV-Metrik	1	**44**	9			
	2	49	**37**	13	1	
	3	7	49	**56**	24	3
	4		5	28	**54**	29
	5			3	21	**68**
	Total	100	100	100	100	100

4.5.3. Skalenäquivalenz und Linking zwischen NEPS und dem Bereich der Erkenntnisgewinnung des LV

Nachdem in den vorangehenden Abschnitten die Ergebnisse des Linkings zwischen NEPS und dem Bereich Fachwissen des LV vorgestellt wurden, steht das Linking des NEPS-Tests mit den Tests zur Erkenntnisgewinnung im Fokus der nächsten Abschnitte.

Skalenäquivalenz im 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-Skala und den LV-Skalen, im Bereich des Fachwissens signifikante Mittelwertsunterschiede ($p < 0{,}05$) der beiden Skalen bei ansonsten gleich verteilter Schiefe und gleich verteiltem Exzess (Tab. 7). Durch die Verlinkung wird die NEPS-Verteilung an die Verteilung des LV angeglichen, wodurch die Unterschiede in den Mittelwerten (44-65 Punkte) verschwinden.

Tab. 7 Verteilung der Personenfähigkeit im NEPS, den Tests zur Erkenntnisgewinnung im LV und im NEPS in der LV-Metrik

	M (*SE*)	*SD*	Schiefe	*SD*	Exzess	*SD*
NEPS	540 (3.13)	81	-0.01	0.09	-0.39	0.19
Biologie-Erkenntnisgewinnung (BE)	584 (3.22)	84	-0.19	0.09	-0.15	0.19
NEPS in der BE-Metrik	584 (3.39)	88	-0.17	0.09	-0.19	0.19
Chemie-Erkenntnisgewinnung (CE)	605 (4.11)	107	- .14	0.09	- .23	0.19
NEPS in der CE-Metrik	605 (4.28)	111	- .17	0.09	- .24	0.19
Physik-Erkenntnisgewinnung (PE)	589 (3.84)	100	- .13	0.09	- .17	0.19
NEPS in der PE-Metrik	589 (3.98)	104	- .17	0.09	- .26	0.19

Legende: *M*: Mittelwert; *SE*: Standardfehler; *SD*: Standardabweichung.

Die Einordnung der verlinkten NEPS- und LV-Werte auf die Kompetenzstufen der Erkenntnisgewinnung in Abbildung 5 zeigt, dass beide Skalen auf äquivalente Verteilungen kommen. Der Chi-Quadrat-Wert liegt im Bereich zwischen $\chi^2 = 0{,}23$ in der Biologie-Erkenntnisgewinnung und $\chi^2 = 0{,}45$ in der Physik-Erkenntnisgewinnung ($df = 4$, *n.s.*) und widerspricht somit der Annahme von unterschiedlichen Verteilungen der Schülerinnen und

Schüler auf die Kompetenzstufen des LV auf Grundlage ihrer Testwerte im NEPS und dem LV.

Der Anteil der Schülerinnen und Schüler, die den festgelegten Mindeststandard (Kompetenzstufe 2) im Bereich der Erkenntnisgewinnung verfehlen, liegt zwischen 3% in Biologie und 6% in Chemie und ist somit etwas höher als der entsprechende Anteil im Bereich des Fachwissens. Mindestens 80% der untersuchten Stichprobe, und damit 5% weniger als im Bereich des Fachwissens, erreichen den Regelstandard (Kompetenzstufe 3) in der Erkenntnisgewinnung. Der Anteil der Schülerinnen und Schüler auf der höchsten Kompetenzstufe (Optimalstandard) variiert allerdings sehr stark zwischen den Fächern (Biologie: 6%; Physik: 32%; Chemie 40%).

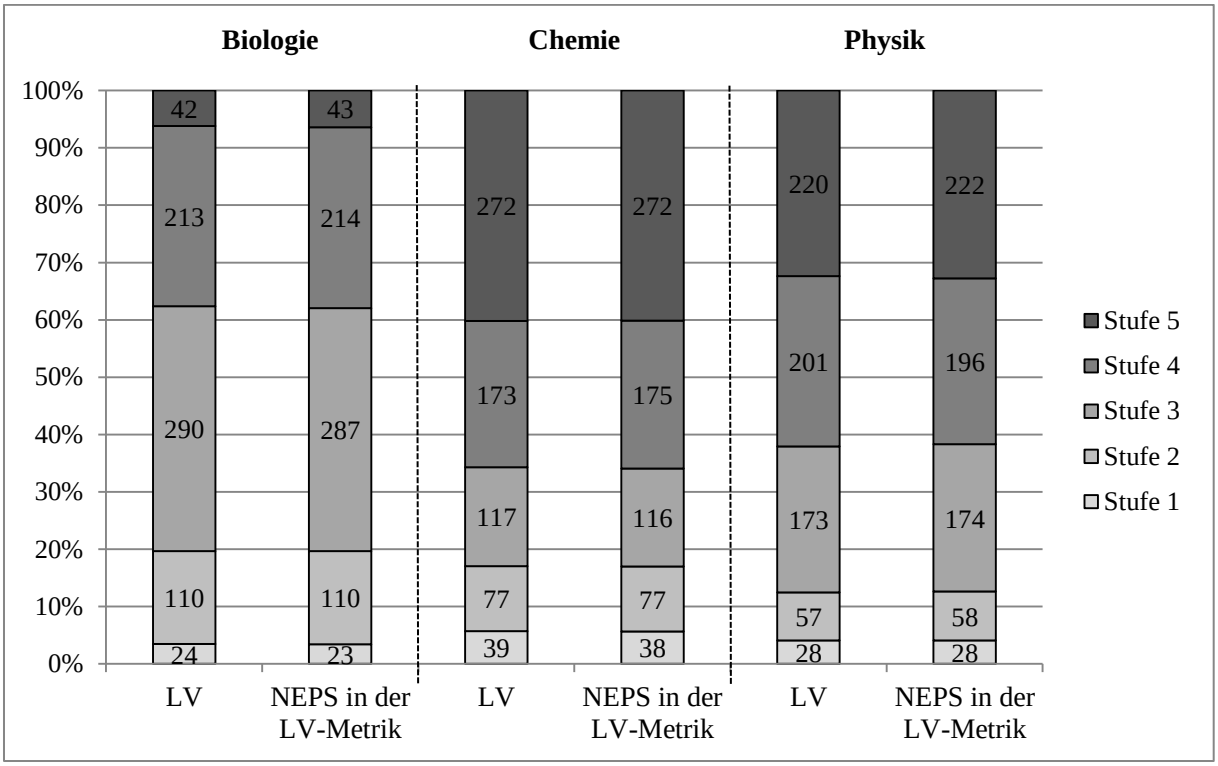

Abb. 5 Einordnung der Schülerinnen und Schülern auf die Kompetenzstufen der Tests zur Erkenntnisgewinnung im LV (die Anzahl der Personen wurde auf den Balken abgetragen)

Individuelle Zuordnung zu den Kompetenzstufen im Bereich Biologie-Erkenntnisgewinnung

Im Unterschied zu den Zellbesetzungen der Kompetenzstufen, die zwischen den Studien praktisch identisch sind, weisen die individuellen Zuordnungen von Befragten zu den Kompetenzstufen im Bereich Biologie-Erkenntnisgewinnung sichtbare Unterschiede auf (Tab. 8). Die Studien NEPS und LV ordnen mindestens 42% (Kompetenzstufe 1) und maximal 61% (Kompetenzstufe 3) der Schülerinnen und Schüler derselben Kompetenzstufe

zu. Die durchschnittliche prozentuale Übereinstimmung zwischen den beiden Studien liegt bei *PÜ* = 52% und der durchschnittliche Cohens-Kappa-Wert beträgt *k* = 0,37.

Tab. 8 Prozentuale Übereinstimmung zwischen den Studien NEPS und LV hinsichtlich der Kompetenzstufenzuordnung in Biologie-Erkenntnisgewinnung

		Kompetenzstufen des LV				
		1	2	3	4	5
NEPS in der LV-Metrik	1	**42**	11			
	2	50	**48**	14	2	
	3	8	41	**61**	29	8
	4			24	**59**	44
	5			1	10	**48**
	Total	100	100	100	100	100

Individuelle Zuordnung zu den Kompetenzstufen im Bereich Chemie-Erkenntnisgewinnung

In Tabelle 9 ist die prozentuale Übereinstimmung zwischen den Studien NEPS und LV hinsichtlich der Kompetenzstufenzuordnung im Bereich Chemie-Erkenntnisgewinnung zusammengefasst. Die Höhe der Übereinstimmung variiert sehr stark zwischen den Kompetenzstufen (36% auf den Kompetenzstufen 2 und 3, und 76 % auf der Kompetenzstufe 5) und liegt im Durchschnitt bei 47%. Der durchschnittliche Cohens-Kappa-Wert ist ähnlich hoch wie im Bereich Biologie-Erkenntnisgewinnung (*k* = 0,37).

Tab. 9 Prozentuale Übereinstimmung zwischen den Studien NEPS und LV hinsichtlich der Kompetenzstufenzuordnung in Chemie-Erkenntnisgewinnung

		Kompetenzstufen des LV				
		1	2	3	4	5
NEPS in der LV-Metrik	1	**47**	16	5	1	
	2	22	**36**	19	8	1
	3	19	31	**36**	19	4
	4	12	16	31	**41**	19
	5		1	9	31	**76**
	Total	100	100	100	100	100

Individuelle Zuordnung zu den Kompetenzstufen im Bereich Physik-Erkenntnisgewinnung

Die in Tabelle 10 abgebildete prozentuale Übereinstimmung zwischen NEPS und LV im Bereich Physik-Erkenntnisgewinnung lässt eine sehr hohe Varianz zwischen den Kompetenzstufen erkennen, die sich zwischen 22% auf der Kompetenzstufe 2 und 72% auf der Kompetenzstufe 5 bewegt. Im Mittel werden 46% der Befragten unabhängig vom verwendeten Testinstrument derselben Kompetenzstufe zugeordnet. Das durchschnittliche Cohens Kappa liegt bei *k* = 0,34 und kann als ausreichend bewertet werden.

Tab. 10 Prozentuale Übereinstimmung zwischen den Studien NEPS und LV hinsichtlich der Kompetenzstufenzuordnung in Physik-Erkenntnisgewinnung

		Kompetenzstufen des LV				
		1	2	3	4	5
NEPS in der LV-Metrik	1	**47**	17	2		
	2	31	**22**	18	3	
	3	18	50	**44**	27	5
	4	4	10	28	**45**	23
	5		1	8	25	**72**
	Total	100	100	100	100	100

4.6. Zusammenfassung und Diskussion

Das Ziel dieser Studie war es, den Naturwissenschaftstest des NEPS für die neunte Klassenstufe mit den Naturwissenschaftsskalen des LV 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.

4.6.1. Dimensionale Äquivalenz

Die erste Fragestellung zielte auf die Untersuchung der latenten Korrelationen zwischen den Naturwissenschaftsskalen der Studien NEPS und LV ab. Die Ergebnisse sprechen für einen hohen Zusammenhang. Dieser Zusammenhang fällt zwar kleiner als die latente Korrelation der Mathematik-Werte im NEPS und dem LV für die neunte Klasse aus (van den Ham et al., 2016), seine Höhe liegt allerdings sehr nah an der Höhe der in der vorliegenden Studie ermittelten Zusammenhänge der einzelnen LV-Tests (Biologie, Chemie und Physik) untereinander. Wenn man davon ausgehen kann, dass die Naturwissenschaftstests des LV verwandte Konstrukte messen, die neben der Kernkompetenz der Naturwissenschaft jeweils noch fachspezifische (Teil-)Kompetenzen beinhalten, dann kann der in dieser Studie vorgefundene latente Zusammenhang als Evidenz für die Ähnlichkeit der Konstrukte im NEPS und dem LV interpretiert werden. Das bedeutet, dass die Testinstrumente beider Studien in Bezug auf ihre Kernkompetenzen äquivalent sind. Darüber hinaus messen sie allerdings spezifische, der jeweiligen Studie eigene Inhalte.

Die Ergebnisse der konfirmatorischen Faktorenanalyse, die zur Analyse der zweiten Fragestellung herangezogen wurde, untermauern diese Beobachtung, indem sie zeigen, dass die NEPS- und LV-Testwerte nicht durch eine gemeinsame Dimension abgebildet werden

können. Die untersuchten informationstheoretischen Indizes sprechen mit einer starken Evidenz für die Abgrenzung der Dimensionen *Fachwissen* und *Erkenntnisgewinnung* im LV sowohl voneinander als auch von der NEPS-Dimension. Eine Erklärung für die Annahme getrennter Dimensionen für NEPS und LV könnte - wie bereits bei der Untersuchung ihrer konzeptionellen Äquivalenz festgestellt (Wagner et al., 2014) - darin liegen, dass die zu vergleichenden Testinstrumente das Konstrukt *Naturwissenschaftliche Kompetenz* unterschiedlich breit messen, und die LV-Items aus diesem Grund ihre eigenen Dimensionen benötigen.

Ein weiterer Grund für die Annahme getrennter Dimensionen für die NEPS- und LV-Tests könnte in den Unterschieden der Studien hinsichtlich des Umgangs mit fehlenden Werten liegen. Insbesondere bei den halboffenen und offenen Items kann die Bewertung der fehlenden Werte als falsch zur Überschätzung der Itemschwierigkeiten und folglich zur Unterschätzung der Personenfähigkeiten führen, da den Befragten zur Beantwortung der Items eventuell nicht nur die fachliche Kompetenz, sondern auch die Bereitschaft zur Verfassung eines (wenn auch nur kurzen) Textes fehlt. Insbesondere unter der Berücksichtigung des hohen Anteils der Items mit halboffenem und offenem Antwortformat am Gesamttest des LV kann der Effekt des unterschiedlichen Umgangs mit fehlenden Werten auf die Vergleichbarkeit der Kompetenzwerte im NEPS und dem LV nennenswert sein. Da die Itemschwierigkeiten in den Hauptstudien NEPS 2010 und LV 2012 unter der oben beschriebenen Bedingung des Umgangs mit fehlenden Daten geschätzt wurden und in allen folgenden Erhebungen (inklusive der vorliegenden Studie) fixiert werden müssen, um die neu erhobenen Testwerte auf die Metrik der Hauptstudien zu bringen, entfällt an dieser Stelle die Möglichkeit der Überprüfung, welchen Einfluss die Art des Umgangs mit fehlenden Daten auf die Vergleichbarkeit der Kompetenzwerte im NEPS und LV hat.

Ein weiteres Ergebnis der konfirmatorischen Faktorenanalyse war das bessere Abschneiden des oben genannten dreidimensionalen Modells im Vergleich zum siebendimensionalen Modell, das neben der NEPS-Dimension sechs Dimensionen des LV unterscheidet. Die erhobenen Daten widersprechen somit der inhaltlichen Unterscheidung der Fächer und Kompetenzbereiche im LV. Da dieses Ergebnis auf Daten von (nur) 678 Schülerinnen und Schüler basiert, sollte es mit Vorsicht interpretiert werden und an einer größeren Stichprobe repliziert werden.

Die Untersuchung der konzeptionellen Äquivalenz zeigte bereits eine große inhaltliche Überschneidung zwischen dem NEPS-Test und den Tests des LV in den Bereichen des Fachwissens und der Erkenntnisgewinnung. Auch die Analyse der latenten Zusammenhänge

legt nahe, dass die Messungen im NEPS und LV nicht unabhängig voneinander sind. Gleichzeitig machen die Befunde der Äquivalenzprüfung deutlich, dass die Definition der naturwissenschaftlichen Kompetenz in den Bildungsstandards im Vergleich zu NEPS breiter ist.

Dieser Unterschied entsteht aus unserer Sicht aus der unterschiedlichen Zielsetzung der Studien. Während das NEPS die Definition der naturwissenschaftlichen Kompetenz am Literacy-Begriff ausrichtet, die in einen alltäglichen Kontext eingebettet ist, liegt das Ziel der Kompetenzmessung im Ländervergleich in der Überprüfung der fachbezogenen Standards in der Schule. Gleichzeitig ist die NEPS-Studie mehr als *nur* die längsschnittliche Messung der Kompetenz. Im Rahmen des NEPS werden über die Kompetenzmessung hinaus unter anderem der Einfluss von Kompetenzentwicklung auf die Entscheidungen an kritischen Übergängen der Bildungskarriere sowie Einflüsse des sozioökonomischen und kulturellen Hintergrundes auf die Kompetenzentwicklung untersucht (Blossfeld et al., 2009). Die Menge der in diesem Zuge erhobenen Konstrukte führt unweigerlich zu einer Einschränkung der Testzeit der einzelnen Konstrukte und als Folge davon zu einer Schwerpunktsetzung der Testinhalte. Abschließend kann festgehalten werden, dass die Struktur der NEPS- und LV-Daten die Unterschiede in den Zielsetzungen der Studien widerspiegeln. Trotz dieses Unterschiedes messen die zu vergleichenden Testinstrumente äquivalente Kernkompetenzen und werden daher als gegenseitig anschlussfähig betrachtet.

4.6.2. Skalenäquivalenz und Linking

Bei der Untersuchung der Skalenäquivalenz und damit der dritten Fragestellung wurde ein Mittelwertsunterschied zwischen den Originalskalen des NEPS und des LV festgestellt: Die Testpersonen erreichten im LV-Test im Durchschnitt einen höheren Fähigkeitswert als im NEPS-Test. Im nächsten Schritt erfolgte die Verlinkung der NEPS-Skala mit den Skalen des LV, in deren Folge der NEPS-Mittelwert an die entsprechenden Mittelwerte des LV angeglichen wurde. Die auf diese Weise verlinkten Fähigkeitsschätzungen wurden anschließend in die Kompetenzstufen des LV eingeordnet. Die Untersuchung der Robustheit des Linkings ergab, dass die Skalen des NEPS und des LV in der Population der Schülerinnen und Schüler der neunten Klasse äquivalente Verteilungen auf die Kompetenzstufen des LV aufweisen. In dieser Hinsicht sind die Ergebnisse dieser Tests untereinander austauschbar.

Allerdings betrifft die festgestellte Äquivalenz der Verteilungen von NEPS und dem LV lediglich die Zellbesetzungen der Kompetenzstufen. Die individuellen Zuordnungen zu den Kompetenzstufen unterscheiden sich jedoch zwischen den Studien. Die Überschneidung der

Studien hängt nicht nur von der Vergleichbarkeit ihrer Testwerte, sondern auch von der Reliabilität ihrer Messungen ab (Huynh, 1990). Doch selbst bei den Testinstrumenten mit einer Reliabilität von 1 und einer Korrelation von 0,95 liegt die Wahrscheinlichkeit für die Zuordnungsäquivalenz laut der Simulationsstudie von Pietsch et al. (2009) bei nur 67%. Je niedriger die Reliabilität der Tests oder die Korrelation zwischen den Tests ausfällt, desto geringer fällt auch die Übereinstimmung in der Kompetenzstufenzuordnung aus. Die für die Naturwissenschaftsskalen von NEPS und dem LV aufgrund ihrer Reliabilitäts- und Korrelationskoeffizienten zu erwartende PÜ beträgt 37%. Die beobachtete PÜ liegt zwischen 46% und 58% und ist somit als zufriedenstellend einzuschätzen. Auch der durchschnittliche Cohens-Kappa-Wert variiert zwischen k = 0,34 und k = 0,41 und zeigt eine ausreichende Übereinstimmung zwischen den Studien an. Beide Koeffizienten übersteigen außerdem in ihrer Höhe die entsprechenden Indizes in den Studien von van den Ham et al. (2016) und Nissen et al. (2015).

Insgesamt kann also die in der vorliegenden Studie ermittelte Linking-Funktion als ausreichend robust angesehen werden. Mit Ausnahme des Bereichs Biologie-Erkenntnisgewinnung besteht die höchste Übereinstimmung zwischen NEPS und dem LV auf der Kompetenzstufe 5. Die niedrigste Übereinstimmung findet sich kompetenzbereichs- und fachübergreifend auf den Kompetenzstufen 1 und 2. Diese vergleichsweise niedrige Übereinstimmung könnte ihre Ursache in der niedrigen Zellbesetzung auf den unteren Kompetenzstufen haben, was den Linking-Fehler in diesem Kompetenzbereich groß werden lässt. Folglich wird davon abgeraten, die Ergebnisse des Linkings für die Untersuchung der Kompetenzentwicklung von Schülerinnen und Schülern der NEPS-Studie, deren Leistung nicht dem Mindeststandard im LV entspricht, zu nutzen. Stattdessen wird vorgeschlagen, das Linking zur Untersuchung der Bedingungsfaktoren des Kompetenzzuwachses auf den höheren Stufen des LV zu verwenden.

Weiterhin offenbart der Vergleich der Zuordnungen zu den Kompetenzstufen des LV zwischen NEPS und dem LV die Tendenz von NEPS, einen hohen Anteil der Schülerinnen und Schüler, die vom LV den Kompetenzstufen 1 bis 3 zugeordnet wurden, in die nächsthöhere Stufe zu klassifizieren. Auf den Stufen 4 und 5 in der Verteilung der LV wird dagegen mit Ausnahme des Bereichs Chemie-Erkenntnisgewinnung ein hoher Anteil der Befragten durch den NEPS-Test in die nächstniedrigere Stufe eingeordnet. Möglicherweise kann dieser Befund damit erklärt werden, dass halboffene und offene Items besonders gern von Schülerinnen und Schülern mit niedriger Kompetenz gemieden werden und ihre Leistung im LV deswegen niedriger bewertet wird als im NEPS. Bei den Schülerinnen und Schülern

mit einer höheren Kompetenz wird die Leistung im LV aufgrund ihres Antwortverhaltens in den halboffenen und offenen Items dagegen überbewertet. Dieser Interaktionseffekt könnte dazu führen, dass die Leistungsbewertung der Studien je nach dem Kompetenzstand unterschiedlich ausfällt. Diese Hypothese bedarf einer gesonderten Prüfung und wird an die zukünftige Forschung adressiert.

Die Verbindung der Naturwissenschaftsskalen 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. Zusammenfassend kann hier also geschlossen 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.

4.6.3. 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 Equipercentile Equating durchzuführen. Jedoch würde bei einer größeren Stichprobe vermutlich sowohl die Robustheit des Linkings hinsichtlich der Randverteilung als auch die Aussagekraft der Ergebnisse der Faktorenanalyse höher sein. Die untersuchte Stichprobe könnte im Hinblick auf die Höhe ihres Kompetenzstandes selektiv sein, da die Schülerinnen und Schüler von den Lehrkräften unterrichtet wurden, die am SINUS-Programm teilgenommen haben. Dies könnte dazu geführt haben, dass Schülerinnen und Schüler dieser Stichprobe über eine höhere Leistung verfügen als die Population der Schülerinnen und Schüler in der neunten Klassenstufe. Zuletzt sollte angemerkt werden, dass die eingesetzten LV-Aufgaben nur eine Auswahl der Aufgaben der Hauptstudie darstellen, und aus diesem Grund die Ergebnisse der vorliegenden Studie nur mit Vorsicht auf den Gesamttest übertragen werden sollten.

Danksagung

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

4.7. 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, Hungary: Akademai Kiado.

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.

Blossfeld, H.-P., Schneider T., & Doll, J. (2009). Die Längsschnittstudie Nationales Bildungspanel: Notwendigkeit, Grundzüge und Analysepotential. *Pädagogische Rundschau*, *63*, 249–259.

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.

Bos, W., Wendt, H., Köller, O., & Selter C. (2012). *TIMSS 2011. Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich*. Münster: Waxmann.

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.* Kiel, 37–68.

Cartwright, F. (2012). *Linking the British Columbia English examination to the OECD combined reading scale*. Prepared for the British Columbia Ministry of Education.

Cartwright, F., Lalancette, D., Mussio, J., & Xing, D. (2003). Linking provincial student assessments with national and international assessments. *Education, skills and learning, research papers*, Bd. 005.Ottawa: Statistics Canada.

De Ayala, R. J., Plake, B. S., & Impara, J. C. (2001). The impact of omitted responses on the accuracy of ability estimation in item response theory. *Journal of Educational Measurement, 38*, 213–234.

Dziak, J. J., Coffman, D. L., Lanza, S. T., & Li, R. (2012). *Sensitivity and specificity of information criteria.* Technical report series, #12-119. The Methodology Center, the Pennsylvania State University.

Finch, H. (2008). Estimation of item response theory parameters in the presence of missing data. *Journal of Educational Measurement, 45*, 225–245.

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.

Hambleton, R. K., Sireci, S. G., & Smith, Z. R (2009). How Do Other Countries Measure Up to the Mathematics Achievement Levels on the National Assessment of Educational Progress? *Applied Measurement in Education, 22*(4), 376-393.

Hanushek, E. A., & Wößmann, L. (2015). *The knowledge capital of nations: education and the economics of growth.* Cambridge, MA: MIT Press.

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.

Huynh, H. (1990). Computation and statistical inference for decision consistency indexes based on the Rasch Model. *Journal of Educational and Statistical Statistics, 15* (4), 353-368.

Kauertz, 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.

Kauertz, 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.

Kauertz, 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 (2005a) = Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland. (2005a). *Beschlüsse der Kultusministerkonferenz: Bildungsstandards im Fach Biologie für den Mittleren Schulabschluss. Beschluss vom 16.12.2004.* München: Luchterhand.

KMK (2005b) = Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland. (2005b). *Beschlüsse der Kultusministerkonferenz: Bildungsstandards im Fach Chemie für den Mittleren Schulabschluss. Beschluss vom 16.12.2004.* München: Luchterhand.

KMK (2005c) = Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland. (2005c). *Beschlüsse der Kultusministerkonferenz: Bildungsstandards im Fach Physik für den Mittleren Schulabschluss. Beschluss vom 16.12.2004.* München: Luchterhand.

KMK (2006) = Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland. (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.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159–174.

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, NJ: ETS Educational Testing Service.

Lord F. M. (1974). Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika, 39*, 247–264.

Ludlow, L. H., & O'leary, M. (1999). Scoring omitted and not-reached items: Practical data analysis implications. *Educational and Psychological Measuremant, 59*, 615–630.

Maurice, J. von, 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 das Fach Biologie. 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. 74-83). Münster: Waxmann.

Mislevy, R. J. (1992). *Linking educational assessments: concepts, issues, methods, and prospects.* Princeton, NJ: ETS Policy Information Center.

Muthén, L. K., & Muthén, B. O. (2012). *Mplus User's Guide.* Seventh Edition. Los Angeles, CA: Muthén and Muthén.

National Center for Education Statistics (2013). *U.S. States in a Global Context: Results from the 2011 NAEP-TIMSS Linking Study.* Washington, DC: Institute of Education Sciences.

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 equipercentile and IRT methods. *Studies in Educational Evaluation, 47*, 58–67. DOI: 10.1016/j.stueduc.2015.07.003.

OECD (2006). *Assessing scientific, reading and mathematical literacy: A framework for PISA 2006.* Paris: OECD.

Pant, H. A., Stanat, P., Schröders, U., Roppelt, A., Siegle, T., & Pöhlmann, C. (Hrsg.) (2013). *LV 2012. Mathematische und Naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I.* Münster: Waxmann.

Pietsch, M. Böhme, K., Robitzsch, A., & T. C. Stubbe (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 und Basel: Belz Verlag.

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

Pohl, S., Gräfe, L., & Rose, N. (2014). Dealing with omitted and not reached items in competence tests: Evaluating approaches accounting for missing responses in IRT models. *Educational and Psychological Measurement, 74*, 423–452.

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.

Prenzel, M., Stadler, M. Friedrich, A., Knickmeier, K., & Ostermeier, C. (2009). *Increasing the efficiency of mathematics and science instruction (SINUS) – A large scale teacher professional development programme in Germany.* Kiel: Leibniz-Institute for Science and Mathematics Education. https://www.ntnu.no/wiki/download/attachments/8324749/SINUS_en_fin.pdf?version=1andmodificationDate=1251384255000

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology, 25*, 111-163.

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

Rose, N., von Davier, M., & Xu, X. (2010). *Modeling nonignorable missing data with item response theory (IRT).* ETS Research Rep. no. RR-10-11. Educational Testing Service. Princeton, NJ.

Schöps, K., & Saß, S. (2013). *NEPS Technical Report for Science. Scaling Results of Starting Cohort 4 in Ninth Grade. NEPS (Working Paper No. 23).* Otto-Friedrich-Universität. Bamberg.

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 den Ham, A.-K., Ehmke, T., Nissen, A., & Roppelt, A. (2016): Assessments verbinden, Interpretationen erweitern? *Z Erziehungswiss. DOI:* 10.1007/s11618-016-0686-2.

van de Vijver, F.J.R. (1998). Towards a Theory of Bias and Equivalence. In J. Harkness (Hrsg.), *ZUMA-Nachrichten Spezial*, *3*, 41-65. Mannheim: ZUMA.

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., 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.

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.

5

Gesamtdiskussion

5.1. Zusammenfassung

Das Ziel dieser Forschungsarbeit lag in der Untersuchung der Anschlussfähigkeit des Naturwissenschaftstests von NEPS für die neunte Klassenstufe an die naturwissenschaftlichen Messungen in PISA 2012 und im Ländervergleich (LV) 2012. Dazu wurde der Ansatz von Kolen und Brennan (2004) genutzt, dessen Fokus beim Vergleich der Testinstrumente auf den Schlussfolgerungen, Zielpopulationen, Messintentionen und Konstrukten liegt. Der letzte Aspekt kann im Rahmen des kulturvergleichenden Ansatzes von van de Vijver (1998) in Bezug auf die konzeptionelle Äquivalenz, dimensionale Äquivalenz und die Skalenäquivalenz der Testinstrumente beurteilt werden.

Im Folgenden werden zunächst die Ergebnisse zu den jeweiligen Fragestellungen dieser Arbeit einschließlich der Ergebnisse des Linkings zusammengefasst und diskutiert. Zur besseren Lesbarkeit folgt die Ergebnisdarstellung der inhaltlichen Logik des Vergleichs, d.h. die Befunde zur Untersuchung der einzelnen Schritte der Konstrukt-Äquivalenz werden nacheinander zusammengefasst und diskutiert. Anschließend werden die Ergebnisse der durchgeführten Vergleiche studienübergreifend diskutiert. Abgeschlossen wird die vorliegende Arbeit mit den Limitationen der empirischen Befunde und einem Ausblick auf den Nutzen der Linking-Ergebnisse für die zukünftige Forschung.

5.1.1. Der konzeptionelle Vergleich des NEPS-Naturwissenschaftstests mit den PISA- und LV-Testinstrumenten

Bei der Untersuchung der konzeptionellen Äquivalenz der Testinstrumente wird die Vergleichbarkeit der Rahmenkonzeptionen verschiedener Studien geprüft. Ziel dieser Untersuchung ist es, auf der inhaltlichen Ebene Gemeinsamkeiten und Unterschiede der Testinstrumente festzustellen, um eine Aussage darüber zu treffen, inwiefern die zu vergleichenden Testinstrumente ihre Konstrukte auf dieselbe Art und Weise definieren.

Die erste Studie dieser Dissertation konnte zeigen, dass der NEPS-Naturwissenschaftstest mit den PISA- und LV-Rahmenkonzeptionen auf der Ebene der Testinhalte kompatibel ist. Der Grad der Passung ist jedoch davon abhängig, in welche Dimension des PISA-Tests bzw. der LV-Tests die NEPS-Items eingeordnet werden. Während die Zuordnungsrate zu den Wissensbereichen im LV (93%) höher als in PISA (79%) ausfällt, können den Teilkompetenzen der naturwissenschaftlichen Grundbildung in PISA (96%) ein höherer Anteil von NEPS-Items zugeordnet werden als den kognitiven Prozessen (75%) und der Komplexität im LV (64%). Dabei scheint der Fokus des NEPS-Tests eher im mittleren

Anforderungsbereich der Rahmenkonzeptionen von PISA und dem LV zu liegen, da jeweils mehr als die Hälfte der NEPS-Items der Kategorie *naturwissenschaftliche Phänomene erklären* in PISA und der Komplexitätsstufe *Ein Zusammenhang* im LV zugewiesen wurde.

Die Zuordnung der NEPS-Items zu den Wissensbereichen von PISA und dem LV lässt erkennen, dass der NEPS-Test nicht alle Inhalte dieser Rahmenkonzeptionen abdecken kann und somit aus theoretischer Sicht nicht mit den PISA- und LV-Tests aus 2012 austauschbar ist. Verglichen mit der PISA-Rahmenkonzeption enthält der NEPS-Test keine Aufgaben im Bereich der *Erd- und Weltraumsysteme.* Weiterhin wurden im Expertenreview keine NEPS-Items den *Technologischen Systemen* in PISA zugeordnet. Dieser Befund ist umso erstaunlicher, als dass die Rahmenkonzeption des NEPS explizit einen technologischen Kontext vorsieht. Allerdings ordneten die Experten diese Items den physikalischen Systemen in der PISA-Rahmenkonzeption zu, da bei ihnen die physikalischen Inhalte augenscheinlich mehr im Vordergrund stehen.

Der konzeptionelle Vergleich des NEPS-Tests mit der LV-Rahmenkonzeption zeigte außerdem, dass der NEPS-Test keine Aufgaben enthält, die die Komplexitätsstufe *Zwei Fakten* abbilden bzw. das Reproduzieren des naturwissenschaftlichen Wissens (*Kognitive Prozesse)* erfordern. Ein Grund dafür könnte in den unterschiedlichen Zielsetzungen der Studien liegen: Während sich die Studien NEPS und PISA in der Definition der naturwissenschaftlichen Kompetenz am Literacy-Begriff orientieren, welcher sich auf alltägliche Kontexte und naturwissenschaftliche Problemstellungen bezieht, ist die Messung der naturwissenschaftlichen Kompetenz im LV an das Curriculum angelehnt und dient der Überprüfung der fachbezogenen Standards in der Schule. Weiterhin beinhaltet die Literacy-Orientierung für die Studien NEPS und PISA, dass ihre Aufgaben nicht auf das Reproduzieren von Wissen abzielen, sondern stattdessen die Anwendung naturwissenschaftlichen Wissens in alltagsnahen Situationen untersuchen (OECD, 2006).

Ein weiterer Grund für die Unterschiede in den Rahmenkonzeptionen könnte in den verschiedenen Testdesigns der Studien liegen. Während die Aufgaben im LV und in PISA in einem Multi-Matrix-Design dargeboten werden, bearbeiten im NEPS alle Testpersonen einer bestimmten Altersstufe dieselben Aufgaben. Weiterhin bestehen zwischen den Studien große Unterschiede in der vorgegebenen Testzeit. Während die Bearbeitung der Aufgaben in der NEPS-Untersuchung 2010 und in der PISA-Untersuchung 2012 in knapp 30 Minuten erfolgte, lag die Testzeit in der LV-Studie 2012 bei 120 Minuten. Die Kombination des Testdesigns mit der Testzeit führt unweigerlich dazu, dass die naturwissenschaftlichen Tests der Studien unterschiedlich lang sind. Von allen drei Naturwissenschaftstests, die in dieser Arbeit

untersucht wurden, hat der NEPS-Test die geringste Itemanzahl (28). Im Vergleich dazu hat PISA fast die doppelte Itemanzahl (53), die auf drei Testblöcke aufgeteilt wird, während die LV-Tests 386 Aufgaben und somit das Dreizehnfache der Itemanzahl des NEPS-Tests enthalten. Die LV-Items sind ebenfalls auf mehrere Aufgabenblöcke aufgeteilt. Die beschriebene Einschränkung des Testdesigns aufgrund der Testzeit führt im NEPS zu einer Schwerpunktsetzung der Testinhalte, die einen Unterschied in der Breite der NEPS-Rahmenkonzeption im Vergleich zu den PISA- und LV-Rahmenkonzeptionen bedingen könnte.

Abschließend kann festgehalten werden, dass die NEPS-Rahmenkonzeption auf Ebene der Testinhalte eine hohe Überschneidung mit den PISA- und LV-Rahmenkonzeptionen zeigt. Gleichzeitig erweist sich der NEPS-Naturwissenschaftstest nicht als vollständig deckungsgleich mit den PISA- und LV-Tests. Das bedeutet, dass der NEPS-Naturwissenschaftstest zwar nicht exakt dasselbe Konstrukt misst wie die PISA- und LV-Tests, aber aufgrund der hohen Überschneidung im Kern ihrer Rahmenkonzeptionen mit ihnen vergleichbar ist.

5.1.2. Der dimensionale Vergleich des NEPS-Naturwissenschaftstests mit den PISA- und LV-Testinstrumenten

Bei der Untersuchung der dimensionalen Äquivalenz stehen die Analyse der korrelativen Zusammenhänge sowie die Analyse der Datenstruktur im Vordergrund. Diese wurden für den Vergleich von NEPS und PISA in der zweiten Studie sowie für den Vergleich von NEPS und LV in der dritten Studie dargestellt.

Die Analyse der NEPS- und PISA-Daten zeigt einen hohen Zusammenhang zwischen den beiden Testdimensionen sowie eine hohe Tendenz der Daten zur Eindimensionalität. Beide Ergebnisse werden durch den Unterschied im Umgang mit fehlenden Daten beeinflusst und fallen am höchsten aus, wenn fehlende Werte in beiden Studien ignoriert werden. Auch die Korrelationen der Itemresiduen (nach Herauspartialisierung der *naturwissenschaftlichen Kompetenz*) von NEPS und PISA bieten keinen Anhaltspunkt für eine Diskrepanz der Konstrukte, die mit den jeweiligen Testinstrumenten gemessen werden.

Die Analyse der NEPS- und LV-Daten zeigt ebenfalls hohe korrelative Zusammenhänge zwischen den Tests beider Studien, deren Höhe sehr nah an der Höhe der Zusammenhänge der einzelnen LV-Tests untereinander liegt und somit als Beleg für die Verwandtheit der Konstrukte interpretiert werden kann. Die Ergebnisse der konfirmatorischen

Faktorenanalyse deuten ebenfalls darauf hin, dass die beiden Tests in ihrem Kern vergleichbare aber nicht identische Konstrukte messen.

Ein weiterer Unterschied zwischen NEPS und LV liegt in den eingesetzten Antwortformaten und dem unterschiedlichen Umgang mit fehlenden Daten. Die Analysen machen deutlich, dass sich die NEPS- und LV-Daten hinsichtlich des Anteils fehlender Werte am Gesamttest signifikant unterscheiden. Dieser Anteil ist in den LV-Daten in den Aufgaben mit halboffenem und offenem Antwortformat besonders hoch. Da diese Formate im NEPS-Test nicht vorkommen, könnte dieser Umstand unter Berücksichtigung des unterschiedlichen Umgangs mit fehlenden Werten im NEPS und im LV eine Auswirkung auf die Vergleichbarkeit ihrer Testwerte haben. Sollte dieser Effekt für das bessere Abschneiden des dreidimensionalen Modells, das die Dimensionen *Umgang mit Fachwissen* und *Erkenntnisgewinnung* im LV sowie eine zusätzliche NEPS-Dimension unterscheidet, mitverantwortlich sein, würde es bedeuten, dass die dimensionale Trennung der NEPS- und LV-Testwerte nicht nur inhaltlicher Art ist, sondern auch auf die Aufgabenformate und auf den Umgang mit fehlenden Werten zurückzuführen ist.

Abschließend kann festgehalten werden, dass die naturwissenschaftliche Kompetenz im NEPS-Test auf ähnliche Art und Weise gemessen wird, wie in den PISA- und LV-Tests. Darüber hinaus zeigen die Ergebnisse der zweiten Studie, dass der NEPS- und der PISA-Test vergleichbare Konstrukte messen. Dieses kann für den NEPS-Test und die LV-Tests nur eingeschränkt festgestellt werden. Neben den Kernbereichen, in denen sich LV und NEPS überschneiden, gibt es im LV studienspezifische Anteile, die möglicherweise für die dimensionale Trennung der NEPS- und LV-Werte mitverantwortlich sind. Gleichwohl wird die festgestellte Überschneidung der Tests im Bereich der Kernkompetenzen aufgrund der hohen Zusammenhänge zwischen ihren Testwerten als ausreichend bewertet, um eine Verlinkung ihrer Skalen durchführen zu können.

5.1.3. Der skalenbezogene Vergleich des NEPS-Naturwissenschaftstests mit den PISA- und LV Testinstrumenten

Die Untersuchung der Skalenäquivalenz zielt auf die Bewertung von Testleistungen ab und beinhaltet zumeist den Vergleich der Kompetenzstufenklassifikationen von Studien. Da die NEPS-Studie bisher keinen Gebrauch von Kompetenzstufenmodellen macht, wurde die Analyse der Skalenäquivalenz des NEPS-Tests mit den PISA- und LV-Tests unter der Heranziehung des Mittelwertes, der Schiefe und des Exzesses der Testwertverteilungen beider Studien durchgeführt.

Die Ergebnisse der zweiten und der dritten Studie dieser Dissertation belegen, dass die Verteilungen der untersuchten Tests sowohl untereinander als auch in Bezug auf die Normalverteilung vergleichbar sind. Lediglich in den Mittelwerten bestehen zwischen dem NEPS-Test und den PISA- und LV-Tests signifikante Unterschiede: Die Testpersonen erreichen im Durchschnitt in den PISA- und LV-Tests einen höheren Fähigkeitswert als im NEPS-Test, wenn die Items auf die Schwierigkeiten der jeweiligen Hauptstudie fixiert sind. Dieser Unterschied stellt jedoch keine ernsthafte Verletzung der Äquivalenz-Bedingung dar: Der Unterschied der Skalen kann durch eine lineare Transformation der Testwerte ausgeglichen werden.

Es kann festgehalten werden, dass die Untersuchung der Skalenäquivalenz für die NEPS-Testwerte mit den PISA- und LV-Skalen aufgrund der fehlenden Kompetenzstufenklassifikation im NEPS nur bedingt möglich war. Der Vergleich der untersuchten Kennwerte lässt jedoch den Schluss zu, dass die Verteilung der NEPS-Werte mit den Testwert-Verteilungen der beiden anderen Studien vergleichbar ist.

Das Vorliegen der Konstrukt-Äquivalenz dient als eine der Voraussetzungen für die Anschlussfähigkeit der Testinstrumente aus verschiedenen Studien. Zusammenfassend lässt sich feststellen, dass die Ergebnisse der durchgeführten Analysen einerseits eine enge (nicht zuletzt entwicklungsbedingte) Verknüpfung der NEPS-Rahmenkonzeption mit den PISA- und LV-Rahmenkonzeptionen zeigen, die sich sowohl in der hohen konzeptionellen Überschneidung als auch im hohen Zusammenhang ihrer Testwerte niederschlägt. Andererseits lassen sich zwischen NEPS und den zu vergleichenden Studien (insbesondere bezüglich des LV) deutliche Unterschiede mit Blick auf ihre Zielsetzungen und das in den Haupterhebungen angewendete Testdesign konstatieren, die teilweise zur dimensionalen Trennung der Testwerte in der vorliegenden Forschungsarbeit geführt haben. Trotz dieser Unterschiede lässt der inhaltliche, der dimensionale und der skalenbezogene Vergleich des NEPS-Tests mit den PISA- und LV-Tests die übergreifende Schlussfolgerung zu, dass diese Testinstrumente sich in einem Kernbereich in der Messung der naturwissenschaftlichen Kompetenz überschneiden und aus diesem Grund aneinander anschlussfähig sind. Dieses Ergebnis kann weiterhin als Hinweis auf die Validität des NEPS-Tests genutzt werden. Die hohen Zusammenhänge der NEPS-Testwerte mit den Kompetenzwerten in PISA und dem LV können als Beleg für die konvergente Validität des NEPS-Tests interpretiert werden.

5.1.4. Verlinkung des NEPS-Naturwissenschaftstests mit den PISA- und LV-Testinstrumenten

Die Verlinkung der Kompetenzwerte zweier Tests mithilfe des Equipercentile Equating ermöglicht eine Übertragung der Kompetenzskala des einen Tests auf den anderen. Durch die Anwendung dieser Linking-Methode in der vorliegenden Forschungsarbeit konnte die Verteilung des NEPS-Tests an die Verteilung der PISA- und LV-Tests angeglichen werden. Weiterhin ermöglichte die Verlinkung der Naturwissenschaftstests die Übertragung der PISA- und LV-Kompetenzstufenmodelle auf die Kompetenzwerte im NEPS.

Bei der Untersuchung der Robustheit des Linkings wurden keine signifikanten Unterschiede zwischen den Studien in der Klassifikation der Testpersonen zu den PISA- und LV-Kompetenzstufen festgestellt. Die verlinkten Tests zeigen außerdem auf der individuellen Ebene eine hohe prozentuale Übereinstimmung (NEPS-PISA: 60%, NEPS-LV: 46%- 58%) hinsichtlich der Kompetenzstufenzuordnung. Lediglich auf den unteren Kompetenzstufen kommt NEPS im Vergleich zu PISA und dem LV teilweise zu anderen Einstufungen. Dabei ordnet der NEPS-Test, verglichen mit den Tests von PISA und dem LV, einen hohen Anteil der Schülerinnen und Schüler in die jeweils nächsthöhere Kompetenzstufe ein.

Insgesamt kann die Linking-Funktion zwischen dem NEPS-Naturwissenschaftstest und den PISA- und LV-Tests als ausreichend robust bezeichnet werden. Besonders das Linking im mittleren und höheren Kompetenzbereich zeichnet sich durch eine hohe Klassifikationskonsistenz zwischen dem NEPS-Test und den PISA- und LV-Testinstrumenten aus. Die hohe Robustheit der Linking-Funktion in diesen Bereichen macht sie besonders geeignet zur Untersuchung von Fragestellungen, die auf Testpersonen abzielen, deren Leistung im LV zwischen dem Regel- und Optimalstandard bzw. in PISA zwischen der Kompetenzstufe 2 (baseline proficiency level) und der Kompetenzstufe 5 liegt.

5.2. Studienübergreifende Diskussion der Befunde zur Vergleichbarkeit der naturwissenschaftlichen Kompetenz am Ende der Sekundarstufe I

Die vorliegende Arbeit konnte in ihrer Gegenüberstellung der NEPS-, PISA- und LV-Rahmenkonzeptionen zeigen, dass alle drei Studien trotz der (zum Teil) unterschiedlichen Zielsetzungen ähnliche Schlussfolgerungen aus ihren Messungen ableiten. Ähnlich bis identisch sind auch die Zielpopulationen der Studien: lediglich bei PISA gibt es eine Abweichung vom NEPS und dem LV, da sie die Kompetenz der Fünfzehnjährigen erfasst, die

in Deutschland sowohl neunte als auch zehnte Klasse besuchen dürfen. Ein wesentlicher Unterschied zwischen den Studien besteht dagegen in den Messbedingungen: insbesondere die Testlänge, der Umgang mit fehlenden Werten und die eingesetzten Antwortformate können die Vergleichbarkeit der Testinstrumente beeinflussen. Beispielsweise zeigte die Analyse der konzeptionellen Äquivalenz der Testkonstrukte in der ersten Studie, dass die naturwissenschaftliche Kompetenz in den drei untersuchten Tests unterschiedlich breit gemessen wird. Der Grund für diese Diskrepanz könnte neben den unterschiedlichen Zielsetzungen der Studien in der Anzahl der Items liegen, die zur Messung der Kompetenz in diesen Studien eingesetzt werden. Auch der Unterschied der Studien in den eingesetzten Formaten und im Umgang mit fehlenden Werten kann die Vergleichbarkeit der Testinstrumente beeinflussen, wie der dimensionale Vergleich in der zweiten und der dritten Studie dieser Dissertation gezeigt hat. Die zwischen den Studien festgestellten Unterschiede bedeuten jedoch nicht, dass die Vergleichbarkeit der Testinstrumente nicht gegeben ist. Die Argumente für eine Vergleichbarkeit werden im Folgenden diskutiert.

Ein Beleg für die Vergleichbarkeit ist ein Ergebnis des konzeptionellen Vergleichs der Testinstrumente: hier zeigte sich eine hohe Überschneidung der NEPS-Testinhalte mit den Rahmenkonzeptionen von PISA und dem LV. Des Weiteren belegen die Ergebnisse des dimensionalen Vergleichs einen hohen Zusammenhang der NEPS-Testdimension mit den PISA und LV-Dimensionen. Dieser Zusammenhang ist ähnlich stark wie der Zusammenhang der LV-Tests untereinander, was als Hinweis auf die Vergleichbarkeit des NEPS-Tests mit den LV-Tests im Bereich der Kernkompetenzen interpretiert werden kann. Zum anderen entspricht dieser Zusammenhang der Höhe der Korrelationen, die üblicherweise zwischen zwei Testinstrumenten, die das gleiche Merkmal erfassen, vorgefunden werden (vgl. Cartwright, 2012; Cartwright et al., 2003; Hambleton, Sireci & Smith, 2009; National Center for Educational Statistics, 2013; Nissen et al., 2015; Pietsch et al., 2009; van den Ham et al., 2016).

Auch die Ergebnisse der Skalenäquivalenz-Überprüfung liefern keinen Anhaltspunkt für die Annahme unterschiedlicher Verteilungen im NEPS, in PISA und im LV. Der einzig festgestellte Unterschied zwischen den Studien betrifft ihre Mittelwerte und kann lediglich als Hinweis auf eine höhere Schwierigkeit des NEPS-Tests im Vergleich zu den PISA- und LV-Tests interpretiert werden. Insgesamt sprechen die Ergebnisse dieser Arbeit für die Anschlussfähigkeit des NEPS-Testinstruments an die PISA- und LV-Tests. Diese Anschlussfähigkeit stellt die Basis für die Übertragung der naturwissenschaftlichen Skalen von PISA und dem LV auf die NEPS-Naturwissenschaftswerte anhand des Equipercentile

Equating dar, welches nachfolgend diskutiert und hinsichtlich seiner Qualität bzw. Robustheit beurteilt wird.

Die Verlinkung der NEPS-Testwerte mit den PISA- und LV-Skalen mithilfe des Equipercentile Equating macht die Testwerte der Studien untereinander austauschbar. Das bedeutet, dass für jeden in dieser Studie untersuchten NEPS-Testwert ein entsprechendes Äquivalent in der PISA- und LV-Skala existiert und umgekehrt. Die Verlinkung der Skalen kann dazu genutzt werden, die Testwertinterpretationen der drei Studien zu erweitern. Dieses hängt jedoch entscheidend davon ab, inwiefern das durchgeführte Linking ausreichend robust ist.

Die vorliegenden Ergebnisse konnten in Bezug auf die Robustheit der Linking-Funktion zeigen, dass die Verteilungen der Personen auf die Kompetenzstufen der jeweiligen Studien zwischen den Tests nahezu identisch sind. Auch die Kompetenzstufen-Zuordnung der Schülerinnen und Schüler auf der individuellen Ebene kann als ausreichend hoch bezeichnet werden. Lediglich auf den unteren Kompetenzstufen kommt NEPS im Vergleich zu PISA und dem LV zum Teil zu abweichenden Einstufungen. Dies kann einerseits mit einer im Vergleich zu den anderen Kompetenzstufen niedrigen Zellbesetzung erklärt werden, was den Linking-Fehler in diesem Kompetenzbereich groß werden lässt. Zum anderen könnte die Ursache für die beobachtete Diskrepanz darin liegen, dass Aufgaben mit halboffenem und offenem Antwortformat besonders gern von Schülerinnen und Schülern mit niedriger Kompetenz gemieden werden. In Verbindung mit dem unterschiedlichen Umgang der Studien mit fehlenden Werten könnte die beschriebene Interaktion des Antwortformats mit der Leistung der Schülerinnen und Schüler in diesem Kompetenzbereich zu einer unterschiedlichen Kompetenzbewertung geführt haben.

Insgesamt implizieren die Ergebnisse der Verlinkung der NEPS- und PISA-Skalen eine Abnahme in der Robustheit der Linking-Funktion, wenn die fehlenden Werte in PISA studienspezifisch, also als falsch kodiert werden. Dieser Befund ist neu. Bisher wurde der Einfluss des Umgangs mit fehlenden Werten auf das Linking nicht untersucht. Besonders relevant könnte dieser Befund für die Verlinkung der Testwerte von Personen mit Zuwanderungshintergrund sein, die aufgrund der häufig mangelnden deutschen Sprachkenntnisse möglicherweise mehr Zeit für die Bearbeitung der Aufgaben benötigen. Werden die am Ende des Tests nicht bearbeiteten Aufgaben als falsch kodiert, könnte dieses Vorgehen in der Gruppe von Personen mit Zuwanderungshintergrund die Vergleichbarkeit der Ergebnisse aus verschiedenen Studien sowie die Robustheit des Linkings dieser Ergebnisse beeinträchtigen.

Insgesamt hängt die Robustheit einer Linking-Funktion maßgeblich davon ab, wie reliabel die Messungen in den jeweiligen Studien sind und wie hoch sie miteinander zusammenhängen (Huynh, 1990). Laut der Simulationsstudie von Pietsch et al. (2009) kann eine hundertprozentige Klassifikationskorrektheit der Studien in der Zuordnung von Schülerkompetenzen nur bei einer Reliabilität von 1 für beide Tests und einer Korrelation von 1 zwischen den Tests erwartet werden. Fällt die Korrelation zwischen den Tests um 0.05 Punkte auf 0.95 ab, beträgt die maximal zu erwartende Klassifikationskorrektheit 67%. Die PV-Reliabilität der in dieser Forschungsarbeit analysierten Tests liegt im LV zwischen 0.75 und 0.79, in PISA bei 0.80 und im NEPS bei 0.85. Folglich liegt die maximale Klassifikationskorrektheit des NEPS-Tests unter Berücksichtigung der Korrelation mit den LV-Tests ($0.8 \leq r \leq 0.91$) und dem PISA-Test ($r = 0.9$) bei ca. 42%. Die in dieser Arbeit errechnete prozentuale Übereinstimmung des NEPS-Tests mit den PISA- und LV-Tests variiert zwischen 46% und 60% und kann somit als Beleg für eine hohe Robustheit des Linkings in dieser Arbeit interpretiert werden. Beide Linking-Funktionen liefern eine zuverlässige Basis für die Untersuchung der inhaltlichen Fragestellungen, die im Abschnitt zu den Implikationen der Studie dargestellt werden.

Es kann zusammengefasst werden, dass die in dieser Arbeit vorgestellte Untersuchung die Vergleichbarkeit des NEPS-Tests mit den PISA- und LV-Tests im Bereich der naturwissenschaftlichen Kompetenz am Ende der Sekundarstufe I belegt. Die im Rahmen dieser Arbeit durchgeführten Analysen der konzeptionellen, der dimensionalen und der skalenbezogenen Äquivalenz zeigen, dass die Tests der drei Studien im Kern vergleichbare Konstrukte messen und folglich aneinander anschlussfähig sind. Dabei scheint die Anschlussfähigkeit des NEPS-Naturwissenschaftstests an den entsprechenden PISA-Test verglichen mit den LV-Tests höher zu sein. Dieser Befund kann möglicherweise mit der Äquivalenz der Zielsetzungen in diesen Studien sowie einer daraus folgenden stärkeren Anlehnung der Rahmenkonzeption von NEPS an die PISA-Rahmenkonzeption erklärt werden.

Die Verlinkung der NEPS-Skala mit den PISA- und LV-Skalen in der vorliegenden Dissertation kann als ausreichend robust angesehen werden. Sie liefert eine Grundlage für die Untersuchung inhaltlicher Fragestellungen, die die Testwertinterpretationen der drei Studien erweitern könnten.

5.3. Limitationen

Die Limitationen dieser Arbeit betreffen zum einen die Wahl der Stichprobe, die zur Beantwortung der Fragestellungen herangezogen wurde, und zum anderen die in dieser Studie verwendeten Testinstrumente.

In Bezug auf die Stichprobe nennen Kolen und Brennan (2004) zwar eine Stichprobengröße von N=250 als ausreichend, um eine Verlinkung mittels des Equipercentile Equating durchzuführen, jedoch würde bei einer größeren Stichprobe vermutlich sowohl die Robustheit des Linkings auf den unteren Kompetenzstufen als auch die Aussagekraft der Ergebnisse der Faktorenanalyse stärker ausfallen. Dieser Punkt bezieht sich insbesondere auf die dritte Studie zum Vergleich der NEPS-Testwerte mit den Werten des LV, da dort lediglich 678 Schülerinnen und Schüler untersucht werden konnten.

Weiterhin könnte die zur Beantwortung der Fragestellungen herangezogene Stichprobe hinsichtlich ihres Kompetenzstandes aufgrund der Teilnahme der Lehrkräfte in den untersuchten Schulen am Programm *Steigerung der Effizienz des mathematisch-naturwissenschaftlichen Unterrichts* (SINUS; Prenzel & Ostermeier, 2006; Prenzel, Stadler, Friedrich, Knickmeier, & Ostermeier, 2009) selektiv sein. Es kann vermutet werden, dass die untersuchten Testpersonen im Vergleich zur Population der Neuntklässlerinnen und Neuntklässler über eine höhere naturwissenschaftliche Kompetenz verfügen. Diese Vermutung wird durch die in der zweiten und der dritten Studien berichteten Mittelwerte unterstützt, die höher als die jeweiligen Mittelwerte der PISA- und LV-Haupterhebungen ausfallen.

Hinsichtlich der Testinstrumente, die in dieser Arbeit analysiert wurden, ist anzumerken, dass die eingesetzten LV-Aufgaben nur eine Auswahl der Aufgaben der Hauptstudie darstellen, und die Ergebnisse der vorliegenden Studie aus diesem Grund nur mit Vorsicht auf den Gesamttest übertragen werden sollten.

Einen weiteren limitierenden Faktor können die Weiterentwicklungen der Testinstrumente von NEPS und PISA darstellen. Dieser Punkt schränkt zwar die berichteten Ergebnisse nicht ein, ist jedoch für die Generalisierbarkeit der Ergebnisse auf die aktuellen Erhebungen von Bedeutung. In der letzten PISA-Erhebung in 2015 wurden die naturwissenschaftliche Rahmenkonzeption des Tests sowie die Merkmale und Umstände der Messung verändert. Beispielsweise werden vorgelegte, aber nicht zu Ende bearbeitete Aufgaben nun auch bei der Schätzung der Personenparameter ignoriert und fließen in die Schätzung der Plausible Values als Teil des Hintergrundmodells ein (OECD, 2016).

Weiterhin verzichtet PISA ab 2015 auf die Untersuchung der naturwissenschaftlichen Kompetenz im Bereich der technologischen Systeme. Zusätzlich dazu hat PISA 2015 das Erhebungsmodus vom Papier- und Bleistifttest auf computerbasierte Testungen sowie das Skalierungsmodell vom 1PL-Modell auf das 2-PL-Modell umgestellt. Auch der hier analysierte NEPS-Test von 2010 wurde durch neue Items erweitert und wird ab 2014 in drei Stufen (leicht, mittel und schwer) dargeboten. Diese Neuerung ermöglicht dem NEPS-Test eine genauere Schätzung der Personenfähigkeit. Die Weiterentwicklung der NEPS- und PISA-Testinstrumente könnte eventuell eine erneute Verlinkung ihrer Testwerte erfordern, um die Vergleichbarkeit der Kompetenzwerte in den neuen Tests sicher zu stellen. Dieses wird jedoch an die zukünftige Forschung adressiert.

5.4. Implikationen und Ausblick

Die im Rahmen dieser Arbeit untersuchte Vergleichbarkeit der naturwissenschaftlichen Messung im NEPS mit den entsprechenden Messungen in PISA und dem LV impliziert, dass diese Tests im Kernbereich ihrer Kompetenzen miteinander vergleichbar sind. Das bedeutet, dass die Ergebnisse der naturwissenschaftlichen Kompetenzmessung im NEPS mit der Messung in PISA und dem LV in Beziehung gesetzt werden können, was die Testwertinterpretationen dieser Studien erweitern kann.

Die im Abschnitt zu den Limitationen berichtete Selektivität der Stichprobe könnte gleichzeitig einen Gewinn für die Untersuchung weiterführender inhaltlicher Fragestellungen auf Basis der in dieser Arbeit ermittelten Linking-Funktionen darstellen. Wie die Verlinkung der NEPS-Testwerte mit den Skalen von PISA und LV bereits gezeigt hat, zeichnen sich die Linking-Funktionen im mittleren und hohen Kompetenzbereich durch eine besonders hohe Robustheit aus. Folglich kann die ermittelte Linking-Funktion dafür genutzt werden, die Bedingungsfaktoren des Kompetenzzuwachses von Schülerinnen und Schülern im NEPS, deren Leistung in den mittleren oder hohen PISA- oder LV-Kompetenzbereich eingestuft wurde, zu untersuchen.

Gleichzeitig kann die auf die PISA- und LV-Kompetenzstufen verlinkte Leistung der NEPS-Teilnehmerinnen und -Teilnehmer am Ende der Sekundarstufe I selbst als Prädiktor für die Vorhersage des Erfolgs in den Naturwissenschaften in der Sekundarstufe II und anschließend des Studiums eines naturwissenschaftlichen Faches an einer Universität fungieren.

Darüber hinaus könnten die verlinkten Ergebnisse dafür genutzt werden, die Kompetenzmodelle der Studien PISA und LV auf die NEPS-Skala zu übertragen. Da die

NEPS-Studie bisher keinen Gebrauch von Kompetenzstufenmodellen macht, war es bis jetzt nicht möglich, die NEPS-Werte kriterial zu interpretieren. Haschke, Kampa, Hahn und Köller (2017) haben mithilfe der Item Descriptor Matching Methode die Kompetenzstandards für den NEPS-Erwachsenen-Test entwickelt und sie anschließend mit den Daten der Studie validiert. Dieser Prozess ist allerdings relativ aufwendig und mit hohen Kosten verbunden. Die in dieser Forschungsarbeit ermittelte Linking-Funktion kann eine Basis für die Modellierung der Kompetenzstandards im NEPS am Ende der Sekundarstufe I schaffen, die einen wichtigen Übergang in die Oberstufe bzw. in die berufliche Ausbildung darstellt.

5.5. Literatur

Cartwright, F. (2012). *Linking the British Columbia English examination to the OECD combined reading scale*. Prepared for the British Columbia Ministry of Education.

Cartwright, F., Lalancette, D., Mussio, J. & Xing, D. (2003). *Linking provincial student assessments with national and international assessments.* Education, skills and learning, research papers, Bd. 005.Ottawa: Statistics Canada.

Hambleton, R. K., Sireci, S. G. & Smith, Z. R (2009). How Do Other Countries Measure Up to the Mathematics Achievement Levels on the National Assessment of Educational Progress? *Applied Measurement in Education, 22*(4), 376-393.

Haschke, L., Kampa, N., Hahn, I. & Köller, O. (2017). Setting standards to a scientific literacy test for adults using the item-descriptor (ID) matching method. In S. Blömeke & J.-E. Gustafsson (Eds.), *Standard Setting in Education. The Nordic countries in an international perspective* (pp. 319 –339). Cham: Springer.

Huynh, H. (1990). Computation and statistical inference for decision consistency indexes based on the Rasch Model. *Journal of Educational and Statistical Statistics, 15* (4), 353-368.

Kolen, M. J. & Brennan, R. L. (2004). *Test Equating, Scaling, and Linking: Methods and Practices.* New York: Springer.

National Center for Education Statistics (2013). *U.S. States in a Global Context: Results from the 2011 NAEP-TIMSS Linking Study.* Washington, DC: Institute of Education Sciences.

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

equipercentile and IRT methods. *Studies in Educational Evaluation, 47,* 58–67. DOI: 10.1016/j.stueduc.2015.07.003.

OECD (2006). *Assessing scientific, reading and mathematical literacy: A framework for PISA 2006.* Paris: OECD.

OECD (2016). *PISA 2015 Results (Volume I): Excellence and Equity in Education.* Paris: OECD Publishing.

Pietsch, M. Böhme, K., Robitzsch, A. & T. C. Stubbe (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 und Basel: Belz Verlag.

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.

Prenzel, M., Stadler, M. Friedrich, A., Knickmeier, K. & Ostermeier, C. (2009). *Increasing the efficiency of mathematics and science instruction (SINUS) – A large scale teacher professional development programme in Germany.* Kiel: Leibniz-Institute for Science and Mathematics Education. https://www.ntnu.no/wiki/download/attachments/8324749/SINUS_en_fin.pdf?version=1andmodificationDate=1251384255000

van den Ham, A.-K., Ehmke, T., Nissen, A. & Roppelt, A. (2016): Assessments verbinden, Interpretationen erweitern? *Zeitschrift für Erziehungswissenschaft. DOI:* 10.1007/s11618-016-0686-2.

van de Vijver, F.J.R. (1998). Towards a Theory of Bias and Equivalence. In J. Harkness (Hrsg.), *ZUMA-Nachrichten Spezial, 3,* 41-65. Mannheim: ZUMA.

CURRICULUM VITAE

Persönliche Angaben

Name	Helene Wagner
Geburtsdaten	29.04.1982 in Jalutorowsk (Russische Föderation)
Nationalität	deutsch

Schulbildung und Studium

September 1999 bis März 2002	Studium der Linguistik und interkultureller Kommunikation an Tjumener Staatlichen Universität (Russ. Föderation)
Oktober 2003 bis März 2012	Studium der Psychologie an der Freien Universität Berlin
Juni 2012 bis Mai 2018	Promotion im Fach Psychologie an der Christian-Albrechts-Universität zu Kiel

Berufliche Tätigkeiten

Seit Juni 2012	Wissenschaftliche Angestellte am Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik (IPN), Kiel

Kiel, im Februar 2018