LEUPHANA
UNIVERSITÄT LÜNEBURG

PISA, Bildungsstandards und die National Educational Panel Study (NEPS)

Vergleich der Rahmenkonzepte und Validierung der NEPS-Testinstrumente in den Naturwissenschaften und in der Mathematik

Abschlussbericht

Förderkennzeichen LSA009

Hauptantragsteller: Prof. Dr. habil. Timo Ehmke,
Leuphana Universität Lüneburg

Mitantragstellerinnen: Dr. Silke Rönnebeck und
Dr. Inga Hahn,
Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik (IPN), Kiel

wissenschaftliche Mitarbeit: Ann-Katrin van den Ham,
Leuphana Universität Lüneburg und
Helene Wagner,
IPN Kiel

GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

Inhaltsverzeichnis

1 Kurze Darstellung

1.1 Aufgabenstellung

Ziel des Projektes war es, die Naturwissenschafts- und Mathematiktests, die im Rahmen der Studie National Educational Panel Study (NEPS) – Bildungsverläufe in Deutschland für die 9. Klasse erstellt wurden, mit den Studien Programme for International Student Assessment (PISA) 2012 und dem IQB-Ländervergleich(LV) für den Mittleren Schulabschluss zu vergleichen und gegebenenfalls zu verankern. Auf diese Weise sollten die NEPS-Tests validiert und in einem internationalen Referenzmaßstab beziehungsweise in den nationalen Standards verortet werden. Der Forderung nach einer wechselseitigen Verbindung von nationalen und internationalen Testinstrumenten aus Large-Scale-Assessments (LSA) in Deutschland wurde damit Rechnung getragen.

1.2 Voraussetzungen, unter denen das Vorhaben durchgeführt wurde

Das vom Bundesministerium für Bildung und Forschung geförderte Forschungsvorhaben wurde unter der Leitung von Herrn Prof. Dr. habil. Ehmke im Rahmen des Projektes „PISA, Bildungsstandards und die National Educational Panel Study (NEPS). Vergleich der Rahmenkonzepte und Validierung der NEPS-Testinstrumente in den Naturwissenschaften und in der Mathematik“ durchgeführt. An der Projektarbeit beteiligt waren ebenfalls die Mitantragstellerinnen Dr. Silke Rönnebeck, Dr. Katrin Schöps und Dr. Inga Hahn vom Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik (IPN) in Kiel. Während der Projektlaufzeit wurden außerdem die zwei wissenschaftliche

Mitarbeiterinnen Helene Wagner am IPN Kiel und Ann-Katrin van den Ham an der Leuphana Universität Lüneburg eingestellt.

Für die Durchführung des Vorhabens konnte auf drei Expertise-Bereiche zurückgegriffen werden. Diese beziehen sich auf die Entwicklung der NEPS-Rahmenkonzeptionen für Mathematik und Naturwissenschaften, auf die Testentwicklung und -auswertung und auf die Durchführung von Experten-Reviews. Sowohl in den Naturwissenschaften als auch in der Mathematik haben die Antragstellerinnen und Antragsteller maßgeblich an der Entwicklung der NEPS-Rahmenkonzeptionen mitgewirkt. Bei der Entwicklung von Mathematik- und Naturwissenschaftstests kann das IPN auf reichhaltige Erfahrungen aus PISA zurückgreifen. Sowohl in der Mathematik (im Rahmen von PISA 2012) als auch in den Naturwissenschaften (im Rahmen von PISA 2006) ist bzw. war das IPN als eines von mehreren internationalen Zentren an der Aufgabenentwicklung beteiligt. Die dabei gewonnenen Erfahrungen der Antragstellerinnen und Antragsteller flossen in die Entwicklung der NEPS-Testaufgaben ein. Sowohl für Mathematik als auch für die Naturwissenschaften sind unter Mitarbeit der Antragstellerinnen und Antragsteller, basierend auf der jeweiligen NEPS-Rahmenkonzeption, Tests für Schülerinnen und Schüler der 9. Jahrgangsstufe entwickelt worden. Im Bereich der Datenanalyse weisen die Antragstellerinnen und Antragsteller ebenfalls weitreichende Expertise auf, die insbesondere in der Auswertung der Studien PISA-I 2003, PISA-E 2003, PISA-MWH 2004, PISA-I 2006, PISA-E 2006 und PISA-I 2009, aber auch der im Rahmen von PISA durchgeführten Entwicklung eines Testverfahrens zur Überprüfung der Bildungsstandards für den Mittleren Schulabschluss in Mathematik (Prenzel, Artelt et al., 2007) begründet sind. Die entsprechenden Daten wurden sowohl in PISA 2003 als auch in PISA 2006 am IPN aufbereitet, ausgewertet und in verschiedenen Publikationen veröffentlicht. Diese Expertise wurde bei der Auswertung der NEPS-Daten genutzt. Auch im Bereich des Itemreviews konnten die Antragstellerinnen und Antragsteller ihre Expertise aus PISA einbringen. Sowohl im Rahmen von PISA 2003 (Blum et al., 2004) als auch von PISA 2006 (Prenzel, Artelt et al., 2007) wurden am IPN Experten-Reviews zur curricularen Validität bzw. zur Validität der Testaufgaben in Hinblick auf die länderübergreifenden Bildungsstandards des Sekretariats der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (KMK) organisiert, durchgeführt und ausgewertet.

1.3 Planung und Ablauf des Vorhabens

Das Vorhaben war in Arbeitspakete unterteilt, die gemäß der Planung bearbeitet wurden. Anpassungen ergaben sich zum einen auf inhaltlicher Ebene und zum anderen im Hinblick auf die zeitliche Abfolge einzelner Arbeitsschritte (AS). Die für die Planung wichtigen Arbeitsschritte sind in Abbildung 1 (vgl. Vorhabenbeschreibung) dargestellt.

Arbeitsschritte (AS) \ Zeitpunkt (Quartale)	1/2012	2/2012	3/2012	4/2012	1/2013	2/2013	3/2013	4/2013	1/2014	2/2014	3/2014	4/2014
AS 1: Haupterhebung												
Identifikation und Gewinnung von langfristig aktiven SINUS-Schulen	X											
Testdesign entwickeln	X											
Testhefte zusammenstellen	X											
Datenerhebung (PISA, Bildungsstandards, NEPS)		X										
AS 2: Experten-Reviews												
Vorbereitung der Expertenratings	X											
Durchführung der Expertenratings		X										
Kodierung und Auswertung der Expertenratings			X									
Publikation der Ergebnisse				X								
AS 3: Auswertung von Bildungsstandards- und NEPS-Daten												
Bildungsstandards- und NEPS-Datensatz (2. Testtag) verfügbar					X							
Datenaufbereitung (Cleaning, Scoring)						X						
Gemeinsame IRT-Skalierung							X					
Analysen zur dimensionalen und skalenbezogenen Äquivalenz								X				
Publikation der Ergebnisse									X			
AS 4: Auswertung der PISA- und NEPS-Daten												
PISA-Datensatz (1. Testtag) verfügbar							X					
Datenaufbereitung (Cleaning, Scoring)								X				
Gemeinsame IRT-Skalierung									X			
Analysen zur dimensionalen und skalenbezogenen Äquivalenz										X		
Publikation der Ergebnisse											X	X
AS 5: Analyse der Effekte des Modellversuchsprogramms SINUS												
Entwicklung des SINUS-Schulleiterfragebogens	X											
Datenaufbereitung von Prozessindikatoren aus der SINUS-Evaluation					X	X	X	X				
Datenaufbereitung des SINUS-Schulleiterfragebogens					X	X	X	X				
Analysen zu Programmeffekten										X		
Publikation der Ergebnisse											X	X
Berichte verfassen				X				X				X

Abbildung 1: Zeitplan

1.3.1 AS1: Haupterhebung

Zur Vorbereitung der Haupterhebung wurde das bestehende Testdesign der PISA- und des LV in 2012 so erweitert, dass die Testbooklets der NEPS-Studie in Mathematik und Naturwissenschaften (NaWi) integriert werden konnten. Dabei wurde sichergestellt, dass die Kompetenzskalen der NEPS-Tests für Mathematik und Naturwissenschaften mit den Kompetenzskalen der PISA- und des LV-Tests sowohl innerhalb der Tests als auch zwischen den Tests verankert sind.

Ferner wurde aus der Grundgesamtheit aller SINUS-Schulen eine Auswahl von etwa 200 aktiven SINUS-Schulen identifiziert und angeschrieben, um sie für die Teilnahme an der Studie zu gewinnen. Die konkrete Stichprobenziehung erfolgte in Absprache mit dem IEA Data Processing Center (DPC), das auch für die Stichprobenziehung im Ländervergleich verantwortlich ist. Insgesamt wurden 80 Schulen mit 1965 Schülerinnen und Schülern für die Teilnahme an der Studie ausgewählt. Für 1718 Schülerinnen und Schüler lag eine Einverständniserklärung für die Teilnahme an der Studie vor. Eine Beschreibung der Stichprobe ist in Tabelle 1 dargestellt.

Tabelle 1: Schüleranzahl, Geschlechterverteilung und Vorliegen der Einverständniserklärung in Abhängigkeit der Schulform für die Validierungsstudie 2012

Schulform	Schülerzahl	Anzahl an Schulen	Anteil der Jungen (%)	Einverständniserklärung liegt vor (%)
Hauptschule	56	31	48.2	91.1
Schule mit mehreren Bildungsgängen	226	13	50.9	79.2
Realschule	399	16	54.6	96.5
Integrierte Gesamtschule	432	17	48.6	71.5
Gymnasium	852	3	50.5	93.2
Gesamt	1965	80		

Im letzten Vorbereitungsschritt wurden die Testhefte und Codebooks entsprechend des Multi-Matrix-Design erstellt und an das DPC weitergeleitet. Mit der organisatorischen Durchführung der Studie wurde das IEA Data Processing Center beauftragt, das auch

die PISA- und Bildungsstandard-Erhebung in 2012 in den Schulen organisiert und durchführt. Die Haupterhebung in den Schulen fand im zweiten Quartal in 2012 statt. Das Testdesign der Studie wird in Abbildung 2 gezeigt. Die Testung der Schülerinnen und Schüler wurde an zwei Tagen durchgeführt. Am ersten Testtag bearbeiteten die Schülerinnen und Schüler Mathematik- und Naturwissenschaften (NaWi) -Testhefte aus PISA 2012. Die reine Bearbeitungszeit betrug 120 Minuten. Am zweiten Testtag beantworteten die Testpersonen entweder erst Aufgaben aus dem IQB-Ländervergleich (LV) 2012 Mathematik- und NaWi-Test und danach den Mathematik- und NaWi -Test des NEPS oder sie bearbeiteten erst die NEPS-Tests und nachfolgend die Aufgaben aus den LV -Tests. Für die Bearbeitung der LV-Aufgaben und der NEPS-Tests standen den Schülerinnen und Schülern jeweils 60 Minuten zur Verfügung. Im Anschluss an die Testbearbeitung beantworteten die Schülerinnen und Schüler 20 Minuten lang Aufgaben aus dem Subtest des Berliner Tests zur Erfassung fluider und kristalliner Intelligenz für die 8. bis 10. Jahrgangsstufe (BEFKI) zur Erfassung figuraler Aspekte. Für die Schülerinnen und Schüler aus Thüringen ergab sich aufgrund der Besonderheit im Stichprobendesign ein abweichendes Testdesign. Diese Testpersonen haben bereits an der originalen Erhebung des LV 2012 teilgenommen. Im Rahmen dieser Studie bearbeiten die Jugendlichen daher an einem Testtag Aufgaben aus dem PISA 2012 Mathematik- und NaWi-Test und die NEPS Mathematik- und NaWi-Tests. Die reine Bearbeitungszeit betrug auch hier 120 Minuten. Das Einscannen und die Kodierung der Testhefte erfolgten anschließend durch das DPC in Hamburg.

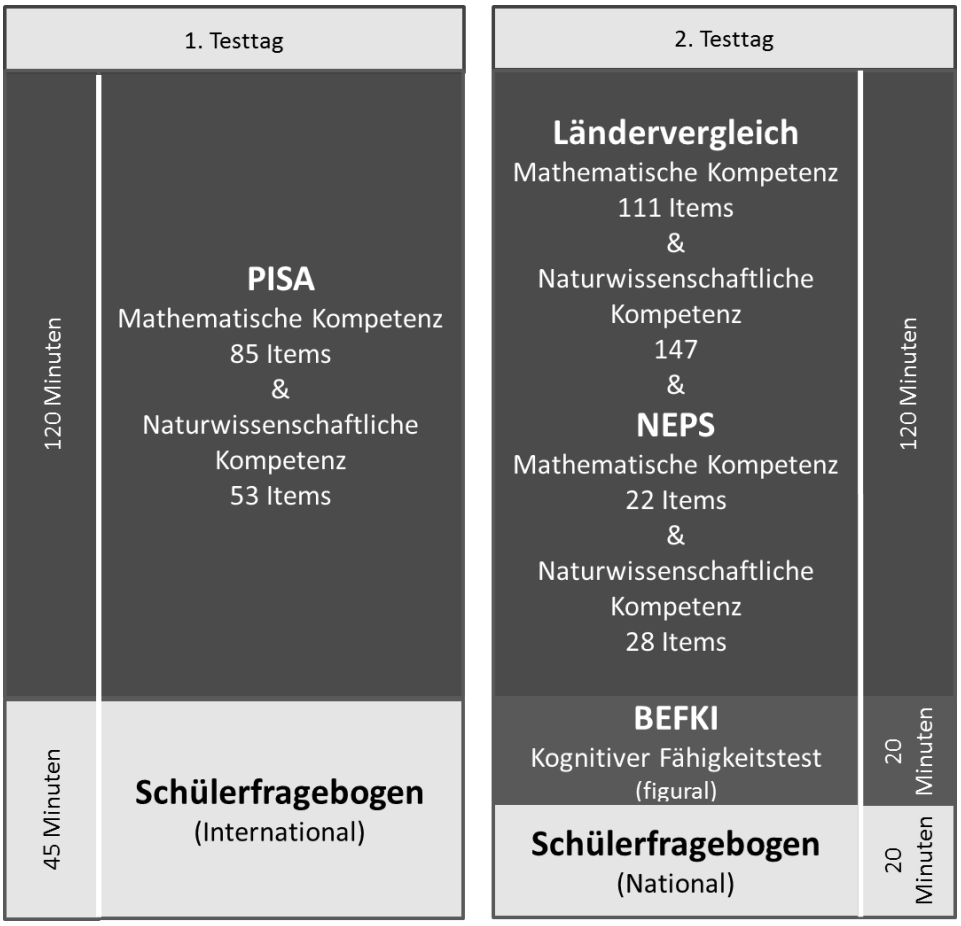

Abbildung 2: Testdesign

1.3.2 AS2: Expertenreviews

Die Durchführung eines Experten-Reviews bezieht sich auf die erste Forschungsfrage der Studie, bei der die konzeptionelle Äquivalenz der Testkonzeptionen für die Domänen Mathematik und Naturwissenschaften in PISA 2012, im LV und in der NEPS-Studie untersucht wird. Um die konzeptionelle Äquivalenz empirisch zu überprüfen, wurden die einzelnen Items des NEPS-Mathematik- und NaWi-Tests von in der Konstruktion von Aufgaben erfahrenen Experten hinsichtlich der curricularen Validität, der Passung zu den Testdimensionen der PISA- und LV-Rahmenkonzeptionen sowie hinsichtlich der Kompetenzstufenmodelle aus PISA und den LV eingestuft. Im Vorfeld der Expertenbefragung wurde dafür ein Fragebogen mit Ratingskalen entwickelt, welches die curriculare Validität und die Passung der NEPS-Testaufgaben zu den genannten Merkmalen erfasst. Das Expertenreview wurde anschließend ausgewertet. Für die Überprüfung der Interraterreliabilität wurden Generalisierbarbeit, die prozentualen Übereinstimmung ($P\ddot{U}$) und das zufallskorrigierte Übereinstimmungsmaß Cohens Kappa verwendet. Die Aufgaben wurden anschließend basierend auf den Experteneinschätzungen in die jeweiligen Kategorien der Inhaltsbereiche, Prozesse, Anforderungsbereiche und Kontexte der Rahmenkonzeptionen von PISA und Ländervergleich eingeordnet. Dabei wurde jeweils die Kategorie bzw. Abstufung für eine Aufgabe gewählt, die von mindestens zwei Experten für diese Aufgabe eingeschätzt wurde. Auf diese Weise war es möglich, alle Aufgaben in die Kategorien bzw. Abstufungen einzuordnen (für eine präzise Darstellung der der Ergebnisse siehe Kapitel 2.1).

Zusätzlich wurde ein zweites Expertenreview durchgeführt. Dafür wurden die Mathematikaufgaben aus NEPS, LV 2012 und PISA 2012 hinsichtlich formaler, linguistischer und sprachlicher Kriterien untersucht. Im Vorfeld wurde ein Fragebogen entwickelt, welches die sprachlichen, linguistischen und formalen Merkmale der Aufgaben erfasst. Die Analyse wurde von drei Ratern durchgeführt, die in einem Training gezielt für die Aufgabenanalyse anhand eines detaillierten Kodiermanuals geschult wurden. Für die Berechnung der Raterübereinstimmung wurden 36 Items von allen Ratern eingestuft. Als Übereinstimmungsmaße wurden der Median der prozentualen Übereinstimmung und das Cohens Kappa zwischen den drei Raterpaaren über die dreifach eingestuften Aufgaben berechnet. Anschließend wurden die Merkmale der Aufgaben aus den verschiedenen Tests miteinander verglichen (für eine präzise Darstellung der Ergebnisse siehe Kapitel 2.1).

Die Ergebnisse aus den Expertenreviews wurden dem Arbeitsplan entsprechend veröffentlicht (siehe Kapitel 2.6).

1.3.3 AS3: Auswertung von den Bildungsstandards- und NEPS-Daten

Die Daten zu den Testaufgaben aus dem NEPS wurden aufbereitet und in Testscores transformiert. Die Aufbereitung und das Scoring der Testaufgaben für den LV erfolgten in Absprache mit dem IQB in Berlin. Die aufbereiteten Daten aus beiden Tests wurden anschließend sowohl getrennt als auch zusammen nach Modellen der Item-Response-Theorie skaliert. Anhand dieser Kompetenzskalen ließ sich dann die dimensionale Äquivalenz für die mathematische und naturwissenschaftliche Kompetenz des Bildungsstandardtests und des NEPS-Tests analysieren. Entsprechend der Forschungsfrage zur dimensionalen Testäquivalenz wurden die messfehlerkorrigierten Korrelationen zwischen den Teildimensionen (Inhaltsbereiche) innerhalb der Tests und zwischen den Tests bestimmt. In einem nächsten Schritt wurden die Skalen der Mathematiktests aus NEPS und dem Ländervergleich verlinkt. Dafür wurden die Methoden des IRT-Linking und des Equipercentile Linking verwendet.

Die Ergebnisse dieses Arbeitsschrittes werden in Kürze zur Publikation eingereicht (vgl. auch Kapitel 2.6.3):

Ehmke, T., van den Ham, A.-K., Nissen, A., Sälzer, C., & Heine, J.-H. (in Vorbereitung). Measuring Mathematics in international and national Large Scale Assessment: Linking PISA with NEPS.

1.3.4 AS4: Auswertung der PISA- und NEPS-Daten

Im nächsten Arbeitsschritt wurden die PISA-Daten des ersten Testtags der Haupterhebung in die Auswertungen miteinbezogen. Die Aufbereitung und das Scoring der Testaufgaben des PISA-Tests erfolgten in Absprache mit dem Zentrum für internationale Bildungsvergleichsstudien (ZIB) an der Technischen Universität München (TUM). Die aufbereiteten Daten aus beiden Tests wurden analog zum vorigen Schritt sowohl ge-

trennt als auch zusammen nach Modellen der Item-Response-Theorie skaliert. Auf diese Weise ließ sich dimensionale Äquivalenz für die mathematische und naturwissenschaftliche Kompetenz des PISA-Tests und des NEPS-Tests analysieren. Auch hier wurden die messfehlerkorrigierten Korrelationen zwischen den Teildimensionen (Inhaltsbereiche) innerhalb der Tests und zwischen den Tests bestimmt. In einem nächsten Schritt wurden die Skalen der Mathematiktests aus NEPS und PISA verlinkt. Dafür wurden die Methoden des IRT-Linking und des Equipercentile Linking verwendet.

Auch die Ergebnisse dieses Arbeitsschrittes werden in Kürze zur Publikation eingereicht (vgl. auch Kapitel 2.6.3):

van den Ham, A.-K., Ehmke, T., Roppelt, A., & Nissen, A. (in Vorbereitung). Assessments verbinden, Interpretationen erweitern? Lässt sich die Mathematikskala des Ländervergleichs 2012 auf die Mathematikergebnisse aus dem Nationalen Bildungspanel übertragen?

1.3.5 AS5: Analyse der Effekte des Modellversuchsprogramms SINUS

Vor der Durchführung der Erhebung wurde ein Fragebogen für die Schulleiter der SINUS-Schulen entwickelt. Nach der Datenlieferung durch das DPC wurden die Daten aus dem SINUS-Schulleiterfragebogen aufbereitet. Die Programmeffekte wurden bisher in einer Masterarbeit analysiert (vgl. auch Kapitel 2.4.1 und 2.6.2):

Segnitz, J. (2013). Die Wirkung des Schulentwicklungsprogramms SINUS auf den Mathematikunterricht. Unveröffentlichte Masterarbeit, Leuphana Universität Lüneburg.

1.4 Wissenschaftlicher Stand, an den angeknüpft wurde

Im Folgenden wird zunächst auf internationale und im Anschluss auf nationale Studien eingegangen, die den Versuch unternommen haben, nationale und internationale

Testinstrumente aus Large-Scale-Assessments zu verknüpfen. Eine solche Verknüpfung erlaubt es, nationale Ergebnisse an einem internationalen Referenzmaßstab zu messen und einzuordnen. Die Ergebnisse nationaler oder internationaler Schulleistungsvergleiche können dabei in großem Maße bildungspolitische Entscheidungen beeinflussen. Aus bildungspolitischer, aber auch aus wissenschaftlicher Perspektive ist es daher von hohem Interesse zu untersuchen, inwieweit unterschiedliche Testinstrumente zu vergleichbaren Aussagen hinsichtlich der Kompetenzen der untersuchten Schülerinnen und Schüler - und damit zu vergleichbaren bildungspolitischen Empfehlungen - gelangen. Abgeschlossen wird der Abschnitt durch eine Verbindung der methodischen Fragestellungen mit der Untersuchung langfristiger Effekte des Modellversuchsprogramms SINUS der Bund-Länder-Kommission (BLK) beziehungsweise des Nachfolgeprojektes SINUS-Transfer (im Folgenden zusammenfassend als SINUS bezeichnet) auf die mathematische und naturwissenschaftliche Kompetenz von Schülerinnen und Schülern.

1.4.1 Internationale Studien

Ein Versuch, die Ergebnisse nationaler Tests in einen internationalen Rahmen einzuordnen, wurde beispielsweise von Cartwright, Lalancette, Mussio und Xing (2003) unternommen. Die Autoren verknüpften die Ergebnisse, die 15-Jährige im nationalen Test zum Leseverständnis FSA 2000 (Foundation Skills Assessment) erreichten, mit den Ergebnissen von PISA 2000 (Cartwright et al., 2003). Die Stichprobe bestand aus 2 800 Schülerinnen und Schülern der 10. Klasse, die beide Tests bearbeiteten. Die Testaufgaben konnten von Leseexperten in der jeweils anderen Rahmenkonzeption zugeordnet werden, woraus geschlossen wurde, dass beide Tests vergleichbare Teildimensionen messen. Mit Hilfe der Equipercentile Linking Methode wurde dann jeder Schülerin und jedem Schüler mit einem gültigen FSA-Ergebnis ein PISA-Wert zugeordnet (Cartwright et al., 2003). Auf ähnliche Weise wurde der FSA 2006 mit PIRLS 2006 (Progress in International Reading Literacy Study) verlinkt. Beide Tests erheben das Leseverständnis von Schülerinnen und Schüler am Ende der 4. Klasse. Auch in diesem Fall wurden die Aufgaben aus beiden Tests im Rahmen von Experten-Ratings als vergleichbar eingestuft. Über 3 500 Schülerinnen und Schüler bearbeiteten beide Tests. Die Skalen beider Tests wurden wiederum mit der Equipercentile Linking Methode verbunden. Aufgrund eines sich überschneidenden Itempools konnten im zweiten Schritt die mit Verfahren der Item-Response-Theory (IRT) skalierten Ergebnisse von FSA 2008 und FSA 2006

miteinander verknüpft werden, so dass auch die Werte von FSA 2008 auf der PIRLS Skala von 2006 verortet werden konnten (B.C. Ministry of Education, o. J.). Das National Center for Education Statistics (NCES) führte 2011 eine Linkingstudie durch, mit welcher die Ergebnisse des National Assessment of Educational Progress (NAEP) 2011 auf die TIMSS-Metrik von 2011 für die achte Klassenstufe übertragen wurden (NCES, 2014). Das Design der Linkingstudie beinhaltete Daten aus den offiziellen NEAP- und TIMSS-Erhebungen aus 2011 und einer Validierungsstichprobe bestehend aus ca. 19 600 Achtklässlerinnen und Achtklässlern, welche zusätzlich zur TIMSS-Erhebung aus neun Staaten repräsentativ gezogen wurde. Für das Linking wurde u.a. die Methode der statistischen Moderation genutzt, mit welcher der Mittelwert und die Standardabweichung der NAEP-Ergebnisse aus der Haupterhebung 2011 an den Mittelwert und die Standardabweichung der TIMSS-Ergebnisse aus der Haupterhebung 2011 angeglichen wurde. Die so entstehenden TIMSS-Scoreäquivalente wurden für Unterschiede in den Ausschlussraten zwischen den beiden Studien korrigiert. Anschließend wurden diese korrigierten Scoreäquivalente genutzt um die TIMSS-Ergebnisse für die Staaten und den Prozentsatz auf bzw. über den internationalen TIMSS-Benchmarks zu schätzen. Auf diese Weise konnte gezeigt werden, dass die mittleren Ergebnisse in Mathematik für öffentliche Schulen in 36 Staaten und die mittleren Ergebnisse für Naturwissenschaften in 47 Staaten höher liegen als der Mittelwert von 500 in TIMSS. Für die Evaluation des Linking wurde die Linkingfunktion auf die Validierungsstichprobe übertragen und die TIMSS- Scoreäquivalente mit den tatsächlich erreichten TIMSS-Ergebnissen in dieser Stichprobe verglichen. Die Konfidenzintervalle wiesen auf statistisch signifikante Unterschiede zwischen den empirischen und den geschätzten TIMSS-Ergebnissen hin. Diese Unterschiede wurden durch die Korrektur für Unterschiede zwischen den Ausschlussraten der Studien jedoch verringert. Weiterhin haben einige US-Bundesstaaten begonnen (z. B. Delaware, Oregon und Hawaii), in der Sekundarstufe I ihre State Assessments im Lesen, in der Mathematik und in den Naturwissenschaften an die Metrik der PISA-Tests anzubinden. Dabei zeigte sich anhand repräsentativer Stichproben, dass die stark curriculumsbasierten State Assessments so hoch mit den PISA-Maßen korrelieren, dass sie sich gemeinsam skalieren ließen (Mean-Mean Equating; vgl. Phillips, 2009).

1.4.2 Nationale Studien – Verknüpfung von PISA (national und international) mit anderen nationalen und internationalen Large-Scale-Assessments

In Deutschland hat es seit Beginn der PISA-Untersuchungen im Jahr 2000 verschiedene Ansätze gegeben, die im internationalen Rahmen erhobenen Daten mit anderen nationalen beziehungsweise internationalen Daten zu verknüpfen. Sowohl in der PISA-Erhebung im Jahr 2000 als auch im Jahr 2003 wurden die internationalen Tests in Mathematik und in den Naturwissenschaften durch nationale Tests ergänzt. Im Bereich der Mathematik stellte der nationale Zusatztest bei PISA 2000 eine konzeptionelle Erweiterung des internationalen Tests dar (Klieme, Neubrand & Lüdtke, 2001). Er war insgesamt stärker curricular orientiert und bezog auch dekontextualisierte Aufgaben mit ein sowie solche, die rein technische Fähigkeiten erforderten. Eine Analyse der um Messfehler korrigierten Korrelation konnte zeigen, dass beide Tests sich analytisch kaum trennen lassen ($r = .91$). Ebenfalls wurden im Rahmen von PISA 2000 die Korrelationen zwischen dem internationalen Mathematiktest sowie Aufgaben, die aus TIMSS beziehungsweise landesweiten Vergleichstest in Brandenburg, Baden-Württemberg und Bayern stammten, untersucht. Anforderungsanalysen durch Experten aus der Fachdidaktik konnten zeigen, dass der internationale PISA-Mathematiktest ein Anforderungsprofil aufweist, das sich sowohl von TIMSS als auch von in Deutschland gängigen unterrichtlichen Anforderungen abhebt (Klieme et al., 2001). Trotzdem zeigten die empirischen Analysen, dass sowohl PISA und TIMSS als auch die Ländertests zwar nicht identische Kompetenzen erfassen, jedoch hohe korrelative Zusammenhänge aufweisen (mit Korrelationen zwischen $r = .89$ und $r = .91$). Ähnliche Ergebnisse lieferte die gemeinsame Skalierung des internationalen Mathematiktests und des nationalen Zusatztests in PISA 2003 (Blum et al., 2004). Auch hier bezog sich der nationale Zusatztest auf die internationale Rahmenkonzeption, differenzierte und erweiterte diese jedoch, indem unterschiedliche Typen mathematischen Arbeitens (technische, rechnerische und begriffliche Aufgaben) sowie die in der Schule behandelten Stoffgebiete Arithmetik, Algebra, Geometrie und Stochastik (als Subskalen) unterschieden wurden. Die messfehlerbereinigten Korrelationen dieser nationalen Subskalen mit den internationalen Subskalen Quantität, Veränderung und Beziehungen, Raum und Form sowie Unsicherheit betrugen zwischen $r = .80$ und $r = .90$; die beiden Gesamttests korrelierten zu $r = .92$. Einen weiteren Versuch der Verknüpfung nationaler und internationaler Large-Scale-Assessments im

Bereich Mathematik unternahmen Hartig und Frey (2012), indem sie die Konstruktvalidität des Tests zur Messung der Bildungsstandards in Mathematik für den Mittleren Schulabschluss untersuchten. In dieser Studie bearbeiteten 9 577 Schülerinnen und Schüler der für den internationalen Vergleich in Deutschland gezogenen Stichprobe am ersten Testtag die PISA-Aufgaben und am zweiten Testtag die neu entwickelten Aufgaben zum LV in Mathematik. In die IRT-Skalierung gingen die Daten der LV Mathematik, des PISA-Mathematik-, des Naturwissenschafts- und des Lesetests ein. Die Ergebnisse konfirmatorischer Faktorenanalysen zeigten eine sehr hohe Korrelation der Dimension LV-Mathematik“ mit der Dimension „PISA-Mathematik“ ($r = .95$). Diese Korrelation war signifikant höher als die Korrelation zwischen „Bildungsstandards-Mathematik“ und „PISA-Naturwissenschaften“ ($r = .87$) und ebenfalls signifikant höher als die Korrelation zwischen „Bildungsstandards-Mathematik“ und „PISA-Lesen“ ($r = .77$). Die Befunde der Studie zeigen, dass zwischen dem LV-Test in Mathematik und dem PISA-Mathematiktest ein sehr hoher Zusammenhang besteht. Ebenso wie der nationale Zusatztest in der Mathematik war der nationale Zusatztest in den Naturwissenschaften bei PISA 2000 stärker an den Lehrplänen orientiert (und unterschied daher die klassischen Schulfächer Biologie, Chemie und Physik) als der internationale Test und sollte damit zum einen die Akzeptanz der internationalen Befunde in Deutschland sichern, zum anderen eine Möglichkeit bieten, den internationalen Test unter bestimmten Gesichtspunkten zu validieren (Prenzel, Häußler, Rost & Senkbeil, 2002). Er stellte weiterhin höhere Anforderungen an das vorhandene Faktenwissen und differenzierte zwischen fünf kognitiven Kompetenzen, die für naturwissenschaftliches Denken und Verstehen typisch sind (Prenzel, Rost, Senkbeil, Häußler & Klopp, 2001). Die um Messfehler korrigierte Korrelation zwischen dem internationalen und nationalen Naturwissenschaftstest lag in PISA 2000 bei $r = .84$ und damit in der gleichen Größenordnung wie die Korrelation zwischen dem internationalen Mathematik- und Naturwissenschaftstest. In PISA 2003 wurde das Design des nationalen Naturwissenschaftstests erweitert und ausdifferenziert. Im Unterschied zu PISA 2000 waren die nationalen Aufgaben nun sieben kognitiven Teilkompetenzen zugeordnet, die mit den naturwissenschaftlichen Fächern und Inhaltsbereichen ein vollständiges Facettendesign bilden. Die Korrelationen zwischen dem internationalen Gesamttest und den kognitiven Teilkompetenzen des nationalen Tests lagen zwischen $r = .63$ und $r = .81$; für fachspezifische Teilkompetenzen fand sich in den Daten keine empirische Evidenz. Ein Vergleich zwischen PISA und anderen internationalen Large-Scale-Assessments hat in Deutschland bislang nur für den Bereich der Lesekompetenz stattgefunden. Artelt, Drechsel, Bos und Stubbe (2008) un-

tersuchten die Zusammenhänge zwischen PISA und PIRLS/IGLU mit dem Ziel, die Vergleichsmöglichkeiten der Studien und ihre bildungspolitischen Konsequenzen zu analysieren. Auch wenn beide Studien am Literacy-Konzept ausgerichtet sind, weisen sie doch konzeptionelle Unterschiede auf. Während PIRLS/IGLU den Erfolg des „Lesenlernens“ abbildet, ist PISA stärker funktional ausgerichtet und bildet den Erfolg des „Lesens, um zu lernen“ ab. Die Studien unterscheiden sich jedoch nicht nur inhaltlich, sondern auch methodisch, so dass die erhaltenen Kennwerte nicht direkt vergleichbar sind. Die einzige Möglichkeit, die beiden Studien miteinander zu verknüpfen, bot eine Untersuchung der jeweiligen Veränderungen in der Lesekompetenz über die Zeit. Die Autoren gelangten zu der Schlussfolgerung, dass ein empirischer Vergleich, der sich auf die gemeinsame Vorgabe von PISA- und PIRLS/IGLU-Aufgaben stützt, sowohl aus wissenschaftlicher als auch aus bildungspolitischer Sicht unbedingt wünschenswert wäre. Einen Vergleich zwischen den länderübergreifenden Bildungsstandards im Kompetenzbereich Lesen und IGLU 2006 nahmen Pietsch, Böhme, Robitzsch und Stubbe (2009) vor. Sie stellten die in beiden Untersuchungen verwendeten Kompetenzstufenmodelle einander gegenüber und fanden aus inhaltlicher Perspektive eine große Übereinstimmung. Aus empirischer Sicht zeigte sich, dass die zentralen Tendenzen beider Studien weitgehend miteinander vergleichbar waren. Die Verteilung der Schülerinnen und Schüler auf die jeweiligen Kompetenzstufen erfolgte mit einer mittleren Klassifikationskorrektheit von 33 Prozent hingegen eher unscharf – ein Befund, der allerdings Simulationsstudien folgend erwartungskonform war.

1.5 Zusammenarbeit mit anderen Stellen

Für die Durchführung der Studie wurden dem Zuwendungsempfänger von der Leuphana Universität Lüneburg Arbeitsräume und notwendiges Computerequipment zur Verfügung gestellt. Das Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik (IPN) ist Nebenantragssteller und Zweitempfänger. Das Projekt wurde in Zusammenarbeit mit dem IPN durchgeführt. Zudem fand eine Kooperation mit dem IQB-Berlin statt. Das IQB stellte dem Zuwendungsempfänger die Testaufgaben sowie weitere Informationen bezüglich der Testskala des vom IQB durchgeführten Ländervergleichs 2012 Mathematik und Naturwissenschaften zur Verfügung. Zudem überließ das IQB dem Zuwendungsempfänger die Ländervergleichsdaten von 2012 für Mathematik

und Naturwissenschaften in der neunten Jahrgangsstufe der Schülerinnen und Schüler aus Thüringen, welche auch an der Studie „PISA, Bildungsstandards und die National Educational Panel Study (NEPS) teilnahmen. Vergleich der Rahmenkonzepte und Validierung der NEPS-Testinstrumente in den Naturwissenschaften und in der Mathematik“ teilgenommen haben. Die Testaufgaben, sowie weitere Informationen bezüglich der Testskala der Mathematiktest und Naturwissenschaftstests aus PISA 2012 wurden vom ZIB an der TUM zur Verfügung gestellt. Die IEA erbrachte mit dem „Oversampling von SINUS-Schulen in PISA 2012 zum Thema PISA, Bildungsstandards und die National Educational Panel Study (NEPS)“ die Feldarbeit als Teilleistung des Foschungsvorhabens. Die Feldarbeit umfasste alle zugehörigen Aufgaben wie Stichprobenziehung (90 Schulen, eine Klasse innerhalb der Schulen), Schulrekrutierung, Druck, Testdurchführung (zwei Testtage), Kodierung, Datenerfassung und Datenaufbereitung sowie Schulrückmeldungen.

2 Eingehende Darstellung

2.1 Ergebnisse des Vorhabens

Mit dem Projekt wurde zwei übergeordneten Zielsetzungen nachgegangen: einer diagnostisch-testtheoretischen und einer inhaltlichen Fragestellung. Im Rahmen der diagnostisch-testtheoretischen Fragestellung zielte das Projekt darauf ab, die Testkonzeptionen und Kompetenzmodelle der Erhebungen des Programme for International Student Assessment (PISA 2012), der länderübergreifenden Bildungsstandards für den Mittleren Abschluss und des NEPS in den Domänen Mathematik und Naturwissenschaften vergleichend zu analysieren. Die inhaltliche Fragestellung bezog sich auf die Effekte des BLK-Modellversuchsprogramms SINUS beziehungsweise dem Nachfolgeprogramm SINUS-Transfer auf das schulspezifische Kompetenzniveau von Neuntklässlerinnen und Neuntklässlern in Mathematik und Naturwissenschaften. Hier sollte analysiert werden, inwieweit sich an Schulen, die seit mehreren Jahren an SINUS beziehungsweise SINUS-Transfer aktiv beteiligt waren, Effekte in der mathematischen und naturwissenschaftlichen Kompetenz auf Schulebene nachweisen lassen.

Zielsetzung 1: Analyse der Äquivalenz von Testkonzeptionen und Kompetenzmodellen von PISA, länderübergreifenden Bildungsstandards und NEPS

Ein Vergleich von Testinstrumenten und empirischen Befunden aus Erhebungen sollte sich nach Kolen und Brennan (2004) auf wenigstens vier Aspekte beziehen (vgl. Pietsch et al., 2009):

a) Merkmale und Umstände der Messung: Wie unterscheiden sich die verwendeten Aufgabenformate, die Durchführungsbedingungen oder die verfügbaren Testzeiten?
b) Zielpopulationen: In welchen Zielpopulationen werden die Testinstrumente einge-

setzt?

c) Operationalisierte Konstrukte: Welche Konstrukte werden durch die Testinstrumente erhoben? Inwieweit erfassen diese dieselben inhaltlichen Teilbereiche und kognitiven Prozesse?
d) Schlussfolgerungen: Welche Erkenntnisse können aus den empirischen Erhebungen gezogen werden? Inwieweit lassen sich die Ergebnisse verschiedener Tests aufeinander beziehen und inwieweit führen unterschiedliche Test zu ähnlichen Personenklassifikationen?

Die ersten beiden Aspekte sind deskriptiv und teilweise bereits eingangs skizziert worden. Für die Studie sind daher vor allem der dritte und vierte Aspekt bedeutsam. Nach Pietsch et al. (2009, vgl. auch van de Vijver, 1998) ist es beim Vergleichen von Testergebnissen wichtig, mögliche Ursachen für einen Konstrukt-, Methoden- und/oder Itembias zu erkennen und aufzuzeigen. Ein Gegenüberstellen der Ergebnisse aus zwei oder mehreren Studien hinsichtlich eines zu messenden Konstrukts setzt demnach voraus, dass zwischen latentem, also nicht direkt beobachtbarem Konstrukt, den verwendeten Testmodellen und den eingesetzten Testitems gleiche Beziehungen bestehen. Dies wird erreicht, wenn neben der konzeptuellen Äquivalenz auch eine dimensionale Äquivalenz und eine Skalenäquivalenz gegeben sind. Inwieweit dies für die mathematik- und naturwissenschaftsbezogenen Kompetenzmessungen in PISA 2012, in LV und NEPS zutrifft, wurde in dieser Studie untersucht. Konkret wurde den folgenden Fragestellungen nachgegangen:

(I) Inwieweit sind die theoretischen Testkonzeptionen für die Domänen Mathematik und Naturwissenschaften in PISA 2012, in den länderübergreifenden Bildungsstandards und im NEPS vergleichbar?
(II) Inwieweit ist die faktorielle Struktur für die latenten Konstrukte „mathematische Kompetenz“ und „naturwissenschaftliche Kompetenz“ der Tests in PISA 2012, in den länderübergreifenden Bildungsstandards und im NEPS vergleichbar?
(III) Inwieweit lassen sich durch ein Linking der mathematischen und naturwissenschaftlichen Skalen des NEPS mit den Studien PISA und LV vergleichbare Verteilungen produzieren und die Kompetenzstufenmodelle übertragen?

Als Ergebnis dieser Analysen soll erreicht werden, dass die Kompetenzskalen für Mathematik und Naturwissenschaften in den Studien PISA 2012, dem LV und im NEPS verankert werden können. Somit lassen sich künftig beispielsweise die Ergebnisse der Er-

hebungen im NEPS auch an einem internationalen Referenzmaßstab oder im Vergleich mit den Ergebnissen einzelner Länder verorten. Der Forderung nach einer wechselseitigen Verbindung von nationalen und internationalen Testinstrumenten aus Large-Scale-Assessments in Deutschland wird demnach Rechnung getragen (Blossfeld, Schneider & Doll, 2009).

Zielsetzung 2: Effekte des BLK-Modellversuchsprogramms SINUS auf das Kompetenzniveau von Neuntklässlerinnen und Neuntklässlern

Entsprechend den Zielsetzungen des SINUS-Projekts sollten sich spezifische Effekte in der mathematischen und naturwissenschaftlichen Kompetenz auf Schulebene nachweisen lassen. Wie hoch diese langfristigen Effekte ausfallen, ist in Deutschland weitgehend unerforscht. Bisherige Evaluation von Schulentwicklungsprogrammen fokussieren überwiegend formative Programmmerkmale (Akzeptanz, Kooperationen) oder decken wesentlich kürzere Zeiträume ab (vgl. Prenzel, Carstensen, Senkbeil, Ostermeier & Seidel, 2005). Konkret sollen zwei Fragestellungen in diesem Bereich analysiert werden:

(I) Inwieweit unterscheiden sich die mathematischen und naturwissenschaftlichen Kompetenzen von Neuntklässlerinnen und Neuntklässlern an langjährig aktiven SINUS-Schulen von den Kompetenzen der Jugendlichen an Schulen, die nicht an SINUS teilgenommen haben?

(II) Welche Bedeutung haben Prozessmerkmale der Programmbeteiligung der Schulen für die mathematischen und naturwissenschaftlichen Kompetenzen von Neuntklässlerinnen und Neuntklässlern an langjährig aktiven SINUS-Schulen?

Als Ergebnisse der Analysen werden empirische Aussagen darüber erwartet, welche langfristigen Effekte mit der Beteiligung von Schulen und Lehrkräften an dem Modellversuchsprogramm SINUS verbunden sind und inwieweit weitere Programme zur Professionalisierung und Unterrichtsentwicklung aussichtsreich sein könnten.

2.1.1 Ergebnisse zur konzeptionellen Äquivalenz der Testkonzeptionen für die Domänen Mathematik und Naturwissenschaften in PISA 2012, Ländervergleich und NEPS

Im Rahmen dieses Ziels wurde analysiert, inwieweit die Kompetenzmessungen in Mathematik und in den Naturwissenschaften in NEPS, PISA 2012 und dem LV 2012 auf der konzeptionellen Ebene vergleichbar sind. Dafür wurde zunächst 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 untersucht.

Naturwissenschaften

Dazu wurden drei Aspekte der konzeptionellen Äquivalenz betrachtet: Die Passung der NEPS-Items zu den Inhalten der PISA- und LV-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 die Generalisierbarkeitstheorie herangezogen. Die Ergebnisse zeigen, dass der Generalisierbarkeitskoeffizient Φ eine zufriedenstellende Höhe erreicht und die systematische Fehlervarianz einen geringen Anteil an der Gesamtvarianz einnimmt. Somit kann die Interrater-Reliabiltät 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 geringere 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 LV ist hoch. Insgesamt konnte hier für vier von fünf NEPS-Konzepten eine Übereinstimmung mit den Wissensbereichen von PISA und den Kompetenzbereichen des LV gefunden werden. Nur die Items des Konzepts „Systeme“ des NEPS konnten den Inhalten von LV und PISA nicht einheitlich zugeordnet werden. Diese Schwierigkeit in der Zuordnung kann bei Betrachtung der Teilbereiche des Kompetenzbereichs „Umgang mit Fachwissen“ dadurch begründet werden, dass der Begriff „Systeme“ in der Rahmenkonzeption des LV 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 Studien Bildungsstandards und PISA. Nur die Kompetenzbereiche der Bildungsstandards und die Teilkompetenzen der naturwissenschaftlichen Grundbildung in 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, dass die Berücksichtigung des beruflichen und fachlichen Hintergrundes der Expertinnen und Experten auf den Dimensionen kognitive Prozesse und Komplexität 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 Komplexitäts-

stufen sein, die eine einheitliche Zuordnung der NEPS-Items unmöglich macht und zu einer unsystematischen Variation der Raterurteile 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 aus 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.

Mathematik

In einem weiteren Schritt wurde die Vergleichbarkeit der Kompetenzmessungen in Mathematik in NEPS, PISA 2012 und dem LV 2012 auf der konzeptionellen Ebene untersucht. Dazu wurden die Testkonzeptionen analysiert: (a) hinsichtlich der Einordbarkeit der Aufgaben aus dem NEPS-Mathematiktest in die entsprechenden Teildimensionen des PISA- bzw. LV-Mathematiktests und (b) hinsichtlich der Gewichtung der einzelnen Teildimensionen im Vergleich zwischen dem NEPS einerseits und den Mathematiktests aus PISA 2012 bzw. aus dem LV 2012 andererseits. Das zweite Ziel war der Vergleich auf Aufgabenebene hinsichtlich (a) formaler und linguistischer Kriterien sowie (b) der sprachlichen Komplexität zwischen den drei Tests. Der Vergleich der mathematischen Rahmenkonzeptionen und Teildimensionen ließ keine deutliche Abgrenzung des NEPS-Tests gegenüber den Tests aus PISA 2012 und dem LV 2012 zu. Kleine

Abweichungen in der konzeptionellen Definition der Teildimensionen zeigen sich unter anderem in Zuordnungsunterschieden zwischen den NEPS-Teilbereichen und den PISA-2012- bzw. LV-2012-Inhaltsbereichen. So werden beispielsweise fast alle NEPS-Aufgaben aus dem Inhaltsbereich „Raum und Form“ in den LV-Inhaltsbereich „Messen“ eingeordnet. Bei diesen Aufgaben müssen z. B. Flächen und Winkel berechnet werden. Das Bestimmen von Maßen fällt in der LV-Rahmenkonzeption unter den Inhaltsbereich „Messen“ und wird in der NEPS-Rahmenkonzeption nicht explizit genannt. Hier wird jedoch das Analysieren von und das gedankliche Operieren mit geometrischen Strukturen unter den Inhaltsbereich „Raum und Form“ gefasst. Den Experten war es möglich, jede NEPS-Aufgabe den Teildimensionen aus PISA 2012 bzw. dem LV 2012 zuzuordnen. Die hieraus entstehenden Verteilungen auf die Teildimensionen zeigten keine bedeutsamen Unterschiede zu den Verteilungen in den originalen Tests aus NEPS, PISA 2012 und dem LV 2012. Lediglich in den Aufgabenkontexten zeigen sich signifikante Unterschiede. Während sich die PISA-Aufgaben relativ gleichmäßig über die persönlichen, ausbildungs-/berufsbezogenen, gesellschaftsbezogenen und wissenschaftlichen Kontexte verteilen, beziehen sich die meisten NEPS-Aufgaben auf einen persönlichen Kontext. Ein Vergleich auf Basis der Rahmenkonzeptionen mit ihren ausführlichen, aber keinesfalls erschöpfenden Beschreibungen der mathematischen Konstrukte kann das Vorhandensein von „jingle fallacy“ und „jangle fallacy“ nicht vollständig aufklären. Tiefgehende Analysen der Aufgabeneigenschaften können evtl. bestehende Unterschiede u. a. bezüglich des Umfangs, den Anforderungen und der Komplexität der Aufgaben zwischen den Studien aufdecken. Diese Befunde decken sich mit den Befunden aus anderen Vergleichsuntersuchungen (Neidorf, Binkley, Gattis & Nohara, 2006). Ein detaillierter Vergleich der formalen und linguistischen Aufgabengestaltung offenbart wichtige Unterschiede, die in einem Vergleich der Rahmenkonzeptionen und Teildimensionen unentdeckt blieben. Es konnte gezeigt werden, dass sich die Aufgaben aus den Mathematiktests in NEPS, PISA 2012 und dem LV 2012 signifikant bezüglich einiger formaler und linguistischer Merkmale und bezüglich ihrer sprachlichen Komplexität unterscheiden. Der PISA-Mathematiktest ist hinsichtlich der Wortschwierigkeiten und der Komplexität der Satzstrukturen in den Mathematikaufgaben anspruchsvoller als der NEPS-Mathematiktest. Außerdem haben die Aufgaben aus dem PISA-2012-Mathematiktest signifikant mehr mathematische Begriffe, Tabellen, Sätze, Wörter, offene und halboffene Aufgabenformate als die Aufgaben des NEPS-Mathematiktests. Dass der NEPS-Test vor allem geschlossene und nur eine halboffene Aufgabe verwendet, ist pragmatischen Gründen zuzuschreiben (kurze verfügbare Testzeit, Kosten für die Kodierung). Im Antworttext der PISA-Aufgaben werden

ebenfalls mehr Wörter gefunden als in den NEPS-Aufgaben. Insgesamt sind also die PISA-Aufgaben sprachlich deutlich schwieriger als die NEPS-Aufgaben. Weniger Unterschiede zeigen sich zwischen den formalen und sprachlichen Merkmalen der Aufgaben aus dem LV-2012- und NEPS-Test. Die NEPS-Aufgaben haben prozentual mehr geschlossene Aufgabenstellungen und weniger mathematische Begriffe als die LV-Aufgaben. Damit kann der NEPS-Test als sprachlich etwas leichter als der LV 2012-Test eingestuft werden. Die gefundenen Unterschiede in der formalen und sprachlichen Aufgabengestaltung, vor allem zwischen den Mathematiktests aus NEPS und PISA 2012, können die Testleistungen von Schülerinnen und Schülern beeinflussen. So können sprachlich komplexe Formulierungen bestimmte Subgruppen (z. B. niedriger sozialökonomischer Status, vom Test abweichende Muttersprache, niedrige mathematische Kompetenz) bei der Lösung der Aufgaben benachteiligen (Abedi & Lord, 2001; Barbu & Beal, 2010; Haag, Heppt, Stanat, Kuhl & Pant, 2013; Wolf & Leon, 2009). Insgesamt lässt sich festhalten, dass die Tests oberflächlich betrachtet große Überschneidungen in den Testkonzeptionen aufweisen. Bei einer detaillierten Untersuchung zeigen sich jedoch Unterschiede in der Verteilung der Aufgaben auf die Kontexte zwischen PISA 2012 und NEPS und deutliche Unterschiede auf der Aufgabenebene zwischen NEPS und PISA 2012 bzw. LV 2012. Der NEPS-Mathematiktest für die neunte Klasse ist dementsprechend nicht als deckungsgleiches, austauschbares Instrument für den PISA-2012- und den LV-2012-Mathematiktest zu betrachten.

2.1.2 Ergebnisse zur dimensionalen Äquivalenz der latenten Konstrukte „mathematische Kompetenz“ und „naturwissenschaftliche Kompetenz“ der Tests in PISA 2012, Ländervergleich und NEPS

Eine dimensionale Äquivalenz setzt voraus, dass zwischen latentem, also nicht direkt beobachtbarem Konstrukt, den verwendeten Testmodellen und den eingesetzten Testitems gleiche Beziehungen bestehen. Dementsprechend wird erwartet, dass Zusammenhänge des Tests mit anderen Messungen des gleichen Konstruktes dem Testkonstrukt entsprechen (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 2014). Wird also von einer Äquivalenz der Mathematischen und naturwissenschaftlichen Konstrukte von NEPS und PISA

bzw. LV ausgegangen, so sollten die Korrelationen der Teildimensionen innerhalb der Tests und die Korrelationen der Teildimensionen zwischen den Tests vergleichbar sein. Außerdem sollten die Korrelationen zwischen den Tests einen hohen Zusammenhang aufweisen und die Informationskriterien für eine eindimensionale Skalierung sprechen. Um die dimensionale Äquivalenz zu überprüfen, werden die drei Tests zuerst getrennt voneinander skaliert. Dabei werden die Inhaltsbereiche der Tests auf jeweils einer eigenen Dimension modelliert. Die hierbei entstehenden Korrelationsmuster der jeweiligen Tests werden einander gegenübergestellt. Diese Berechnungen werden für die Prozesse allerdings nicht durchgeführt, da die Einordnungen der Aufgaben in die Prozesse für den NEPS-Test nicht bekannt sind. In einem zweiten Schritt werden die Korrelationen der Inhaltsbereiche zwischen den Tests berechnet. Für die Naturwissenschaftstests werden dafür zwei vierdimensionale Modelle geschätzt, wobei sowohl die zwei Inhaltsbereiche aus dem PISA-Test als auch die zwei Inhaltsbereiche aus dem NEPS-Test jeweils auf einer eigenen Dimension modelliert werden. Für Korrelationen der Teildimensionen zwischen dem NEPS- und LV-Mathematiktest werden vier sechsdimensionale Modelle berechnet, wobei die Aufgaben des LV-Test auf einer Dimension und die Inhaltsbereiche des NEPS-Tests auf jeweils einer eigenen Dimension modelliert werden. Die Korrelationen der Teildimensionen zwischen dem NEPS- und dem PISA-Mathematiktest wurden auf ähnliche Weise ermittelt. Es wurden vier fünfdimensionale Modelle berechnet, wobei der PISA-Mathematiktest auf einer Dimension und die NEPS-Inhaltsbereiche auf jeweils einer eigenen modelliert werden. Dieses Vorgehen wurde gewählt da eine acht- bzw. neundimensionale Modellierung des NEPS-Mathematiktests mit dem PISA- bzw. LV-Mathematiktest, mit einer Modellierung von sowohl den Inhaltsbereichen des NEPS-Tests als auch des jeweils anderen Tests auf einer eigenen Dimension technisch nicht möglich war. Abschließend wurden die NEPS-Mathematik und -Naturwissenschaftstests mit sowohl den NEPS-Mathematik und -Naturwissenschaftstests aus PISA und LV jeweils eindimensional und zweidimensional modelliert. Dabei wurde bei der zweidimensionalen Modellierung die Korrelationen zwischen den Tests ermittelt. Für beide Modellierungen wurden die Modellgütekriteien Akaike's Information Criterion (AIC), Bayesian Information Criterion (BIC) und das Consistent AIC (CAIC) berechnet und gegenübergestellt. Dabei berücksichtigt der BIC-Kennwert die Anzahl der geschätzten Parameter und verhindert so eine Überparametrisierung (Pohl & Carstensen, 2012) und der CAIC-Wert stellt eine Korrektur des AIC dar un soll auch bei größeren Stichprobenumfängen konsitent sein (Rost, 1996). Sowohl für den AIC, als auch für den BIC und CAIC gilt, dass ein kleinerer Index auf eine bessere Modellpassung weist (Rost, 1996). Im Folgenden

werden die Ergebnisse zur dimensionalen Äquivalenz erst für die Naturwissenschaftstests und anschließend für die Mathematiktests dargelegt. Tabelle 2 und 3 zeigen die Korrelationen der Inhaltsbereiche Knowledge of Science (KOS) und Knowledge about Science (KAS) innerhalb des NEPS bzw. PISA Tests. Mit Korrelationen von $r = .92$ und .96 zeigen beide Tests einen sehr hohen Zusammenhang der Inhaltsbereiche KOS und KAS untereinander.

Tabelle 2: Korrelationen der Inhaltsbereiche innerhalb des NEPS-Naturwissenschaftstests

		NEPS	
		KOS	KAS
NEPS	KOS	1	
	KAS	0.96	1

Tabelle 3: Korrelationen der Inhaltsbereiche innerhalb des PISA-Naturwissenschaftstests

		PISA	
		KOS	KAS
PISA	KOS	1	
	KAS	0.92	1

Tabelle 4 zeigt die Korrelationen der Inhaltsbereiche zwischen den Naturwissenschaftstest aus NEPS und PISA. Die jeweils gleichnamigen Inhaltsbereiche in NEPS und PISA hängen mit Korrelationen von $r = .90$ (KOS) und $r = .87$ (KAS) deutlich miteinander zusammen. Auch der Inhaltsbereich KOS aus NEPS korreliert hoch mit dem Inhaltsbereich KAS aus PISA ($r = .90$). Insgesamt fallen die Korrelationen der Inhaltsbereiche zwischen den Tests jedoch schwächer aus als innerhalb der Tests.

In Tabelle 5 werden die Modellgütekriterien der getrennten und gemeinsamen Skalierung der Naturwissenschaftstests aus NEPS und PISA dargestellt. Sowohl der AIC, der BIC als auch der CAIC weisen geringfügig niedrigere Werte für die zweidimensionale Skalierung auf.

Tabelle 4: Korrelationen der Inhaltsbereiche zwischen den Naturwissenschaftstests aus NEPS und PISA

		PISA	
		KOS	KAS
NEPS	KOS	0.90	0.90
	KAS	0.77	0.87

Insgesamt kann festgehalten werden, dass die Naturwissenschaftstest aus PISA und NEPS hohe dimensionale Zusammenhänge aufweisen. Beide Tests bilden jedoch nicht eine einzige Dimension naturwissenschaftlicher Kompetenz ab, sonder weisen leichte Unterschiede auf. Diese Befunde decken sich mit den Ergebnissen aus der Untersuchung der konzeptionellen Äquivalenz.

Tabelle 5: Gütekriterien für getrennte und gemeinsame Skalierung von den Naturwissenschftstest aus NEPS und PISA

	N	Parameter	Devianz	AIC	BIC	CAIC
NEPS-PISA Eindimensional	1720	108	86371	86587	87176	87284
NEPS-PISA Zweidimensional	1720	110	86307	86527	87127	87237

Im Folgenden werden die Ergebnisse zur dimensionalen Äquivalenz des NEPS-Mathematiktests mit den Mathematiktests aus PISA und LV dargelegt. Tabellen 6, 7 und 8 zeigen die Korrelationen der Inhaltsbereiche innerhalb der Mathematiktests aus NEPS, PISA und LV. Die Korrelationen fallen in allen drei Tests hoch aus. Im LV fallen sie jedoch mit $r = .78$ bis .89 im Gegensatz zu NEPS ($r = .82$ bis .93) und LV ($r = .82$ bis .91) etwas niedriger aus.

Die Tabellen 9 und 10 zeigen die Korrelationen der Inhaltsbereiche zwischen dem NEPS- und dem PISA- respektive dem LV-Mathematiktest. Die Inhaltsbereiche „Raum und Form“ sowie „Veränderung und Beziehungen“ des NEPS-Mathematiktests hängen stark mit den gleichnamigen Inhaltsbereichen des PISA-Mathematiktests zusammen ($r = .85$ und $r = .84$). Die Bereiche „Quantität“ und „Daten und Zufall“ hängen mit Korrelationen von $r = .75$ und .74 nicht höher mit den gleichnamigen Inhaltsbereichen des PISA-Mathematiktests zusammen als mit den übrigen Inhaltsbereichen ($r = .73$ bis

Tabelle 6: Korrelationen der Inhaltsbereiche innerhalb des NEPS-Mathematiktests

Sub-Dimensionen	Quantität	Raum und Form	Veränderung und Beziehungen	Daten und Zufall
Quantität	1			
Raum und Form	0.86	1		
Veränderung und Beziehungen	0.93	0.89	1	
Daten und Zufall	0.92	0.82	0.91	1

Tabelle 7: Korrelationen der Inhaltsbereiche innerhalb des PISA-Mathematiktests

Sub-Dimensionen	Quantität	Raum und Form	Veränderung und Beziehungen	Daten und Zufall
Quantität	1			
Raum und Form	0.86	1		
Veränderung und Beziehungen	0.91	0.85	1	
Daten und Zufall	0.85	0.82	0.90	1

Tabelle 8: Korrelationen der Inhaltsbereiche innerhalb des LV-Mathematiktests

Sub-Dimensionen	Zahl	Messen	Raum und Form	Funktionaler Zusammenhang	Daten und Zufall
Zahl	1				
Messen	0.84	1			
Raum und Form	0.79	0.89	1		
Funktionaler Zusammenhang	0.78	0.82	0.86	1	
Daten und Zufall	0.80	0.87	0.88	0.87	1

$r = .79$)

Tabelle 9: Korrelation der Inhaltsbereiche zwischen dem NEPS- und dem PISA-Mathematiktest

		NEPS			
	Sub-Dimensionen	Quantität	Raum und Form	Veränderung und Beziehungen	Daten und Zufall
PISA	Quantität	0.75	0.77	0.79	0.75
	Raum und Form	0.78	0.85	0.76	0.73
	Veränderung und Beziehungen	0.79	0.79	0.84	0.75
	Daten und Zufall	0.76	0.75	0.77	0.74

Bei der Betrachtung der Zusammenhänge der Inhaltsbereiche zwischen dem NEPS- und dem LV-Mathematiktest lassen sich hohe Zusammenhänge zwischen den konzeptionell ähnlichen Bereichen „Quantität“ und „Zahl“ ($r = .86$), „Raum und Form“ und „Messen“ ($r = .87$), „Raum und Form“ und „Raum und Form“ ($r = .85$) sowie „Veränderung und Beziehungen“ und „funktionaler Zusammenhang“ ($r = .81$) finden. Lediglich der Zusammenhang der gleichnamigen Bereiche „Daten und Zufall“ hängen mit einer Korrelation von $r = .78$ nicht höher miteinander zusammen als mit den übrigen Bereichen.

Sowohl für den PISA- als auch für den LV-Mathematiktest werden keine Inhaltsbereiche gefunden, die sich klar von den Bereichen des NEPS-Mathematiktests abgrenzen lassen. Tendenziell hängen die konzeptionell ähnlichen Inhaltsbereiche stärker miteinander, als mit den übrigen Inhaltsbereichen zusammen. Niedrigere Zusammenhänge können unter anderem durch eine geringe Anzahl an Aufgaben in den NEPS-Inhaltsbereichen erklärt werden oder auch durch kleinere konzeptionelle Unterschiede, welche auch in Kapitel 2.1 dargelegt werden.

In den Tabellen 11 und 12 werden die Modellgütekriterien für die getrennte und gemeinsame Skalierung des NEPS-Mathematiktests mit respektive dem PISA- und dem LV-Mathematiktests gezeigt. Die getrennte Skalierung weist sowohl mit dem PISA- als auch mit dem LV-Mathematiktest niedrigere AIC-, BIC- und CAIC-Werte aus. Dieses Ergebnis weist darauf hin, dass der NEPS-Mathematiktest nicht dieselbe Dimension

Tabelle 10: Korrelation der Inhaltsbereiche zwischen dem NEPS- und dem LV-Mathematiktest

		NEPS			
	Sub-Dimensionen	Quantität	Raum und Form	Veränderung und Beziehungen	Daten und Zufall
LV	Zahl	0.86	0.79	0.77	0.81
	Messen	0.80	0.87	0.77	0.68
	Raum und Form	0.77	0.85	0.76	0.73
	Funktionaler Zusammenhang	0.80	0.77	0.81	0.74
	Daten und Zufall	0.82	0.84	0.78	0.78

mathematischer Kompetenz abbildet, wie der PISA- oder der LV-Mathematiktest.

Tabelle 11: Modellgütekriterien für die gemeinsame und getrennte Skalierung des NEPS- und PISA-Mathematiktests

	N	Parameter	Devianz	AIC	BIC	CAIC
NEPS-PISA Eindimensional	1297	118	67644	67880	68490	68608
NEPS-PISA Zweidimensional	1297	120	67405	67645	68265	68386

Insgesamt kann festgehalten werden, dass der NEPS-Naturwissenschaftstest konzeptionell und dimensional dem PISA-Naturwissenschaftstest sehr ähnlich ist, diese jedoch nicht äquivalent sind. Auch für den NEPS-Mathematiktest kann geschlossen werden, dass dieser dem PISA- und LV-Mathematiktest konzeptionell und dimensional sehr ähnlich ist.

2.1.3 Ergebnisse zur Skalenäquivalenz in den drei Studien PISA, Ländervergleich und NEPS

In einem letzten Schritt wurde die Möglichkeit einer Übertragung der Kompetenzstufen des PISA-Mathematiktests und der Kompetenzstufen sowie Mindest-, Regel- und Opti-

Tabelle 12: Modellgütekriterien für die gemeinsame und getrennte Skalierung des NEPS- und LV-Mathematiktests

	N	Parameter	Devianz	AIC	BIC	CAIC
NEPS-LV Eindimensional	649	136	32419	32691	33300	33436
NEPS-LV Zweidimensional	649	138	32367	32643	33260	33398

malstandards des LV-Mathematiktests auf die Testergebnisse des NEPS-Mathematiktests überprüft. Dafür werden zuerst die Ähnlichkeit der Testwerte im NEPS-Mathematiktest und des PISA- bzw. LV-Mathematiktest statistisch geprüft. In einem zweiten Schritt werden die Skalen des NEPS-Mathematiktests und des Mathematiktests aus PISA-2012 bzw. LV-2012 verlinkt. Die Verteilungen der auf der PISA-Skala bzw. LV-Skala verlinkten NEPS-Testwerte ($NEPS_{PISA}$-Testwerte/ $NEPS_{LV}$-Testwerte) werden sowohl für die Gesamtgruppe als auch für Subgruppen mit der Verteilung der PISA- bzw. LV-Testwerte verglichen. Zusätzlich werden Abweichungen der Linkingfunktion für die Gesamtgruppe von den Linkingfunktionen der Subgruppen näher betrachtet. Abschließend wird die Klassifikationskorrektheit des Linking für die Kompetenzstufen aus PISA und die Kompetenzstufen sowie die Mindest-, Regel- und Optimalstandards aus LV analysiert. Dafür werden die aus dem Linking entstehende Verteilung auf die Kompetenzstufen und Standards auf Basis der $NEPS_{PISA}$- bzw. $NEPS_{LV}$-Testwerte mit der Verteilung der Kompetenzstufen und Standards auf Basis der PISA- bzw. LV-Testwerte verglichen. Anhand dieser Ergebnisse werden Rückschlüsse auf die Robustheit des Linking über Subgruppen sowie auf mögliche kriteriale Interpretationen auf Basis der aus dem Linking entstehenden Verteilung gezogen. Um die Möglichkeit einer Übertragung der Kompetenzstufen aus PISA und der Kompetenzstufen sowie den Standards des LV für Mathematik auf die Testergebnisse des NEPS-Mathematiktests zu überprüfen, geht der vorliegende Artikel der Frage nach, inwiefern das Linking zu äquivalenten Testergebnissen führt. Folgende Forschungsfragen werden in diesem Zusammenhang untersucht:

a) Sind die Verteilungen der Testwerte der Mathematiktests aus NEPS und PISA statistisch vergleichbar?
b) Inwieweit können durch ein Skalenlinking mit dem NEPS-Mathematiktest äquivalente Testergebnisse zum PISA-Mathematiktest erzeugt werden?
c) Inwieweit lassen sich durch ein Skalenlinking mit dem NEPS-Mathematiktest gleiche Verteilungen auf die Kompetenzstufen produzieren wie mit dem PISA-Mathematiktest?

d) Sind die Verteilungen der Testwerte der Mathematiktests aus NEPS und dem LV statistisch vergleichbar?
e) Inwieweit können durch ein Skalenlinking mit dem NEPS-Mathematiktest äquivalente Testergebnisse zum LV-Mathematiktest erzeugt werden?
f) Inwieweit lassen sich durch ein Skalenlinking mit dem NEPS-Mathematiktest gleiche Verteilungen auf die Kompetenzstufen und Standards produzieren wie mit dem LV-Mathematiktest?

Insgesamt nahmen 80 Klassen aus 80 Schulen an der Studie teil (siehe Kapitel 1.3.1). Es bearbeiteten N = 1270 Neuntklässlerinnen und Neuntklässler (50% männlich, 50% weiblich) sowohl den NEPS-Mathematiktest als auch Aufgaben aus dem PISA-Mathematiktest und N = 636 Neuntklässlerinnen und Neuntklässler (52% männlich, 48% weiblich) sowohl den NEPS-Mathematiktest als auch Aufgaben aus dem LV. Für die Analysen des Linking werden jeweils die Testergebnisse der Schülerinnen und Schüler verwendet, die sowohl den NEPS-Test als auch den PISA- bzw. LV-Test bearbeitet haben. Vor der Durchführung des Linking wird in einem ersten Schritt wird die Vergleichbarkeit der Verteilungen der Testwerte aus dem LV-Test und dem NEPS-Test analysiert. Dafür werden zunächst die Streuungen der Testwerte auf Normalverteilung geprüft und die Eigenschaften der Verteilungen einander gegenübergestellt. Anschließend werden beide Verteilungen durch ein Streudiagramm mit der NEPS-Skala auf der X-Achse und der PISA- bzw. LV-Skala auf der Y-Achse miteinander in Beziehung gesetzt. Des Weiteren wird die latente Korrelation zwischen den Tests berechnet. Hierfür werden beide Tests mit der Software ConQuest (Wu, Adams, Wilson & Haldane, 2007) zweidimensional skaliert.

In sowohl der Studie von Cartwright (2012) und Nissen, Ehmke, Köller und Duchhardt (eingereicht) wurde die Linking-Methode des Equipercentile Equating einer IRT-basierten Verlinkung gegenübergestellt. Beide Studien kommen zu dem Schluss, dass eine IRT-basierte Verlinkung die Verteilungseigenschaften der originalen Skala weniger gut repräsentiert. Aus diesem Grund wird für die Verlinkung NEPS-Testergebnisse und den Berichtsskalen aus PISA und LV die Methode des Equipercentile Equating gewählt.

Die Equipercentile Linkingfunktion entsteht durch das Identifizieren von Testwerten im Test X, welche den gleichen Perzentilrang wie Testwerte im Tests Y haben. Dabei wird in dieser Studie die Definition des Equipercentile Linking von Braun und Holland (1982) verwendet (Kolen & Brennan, 2010, vgl. auch).

Die folgenden Funktionen sind nach Braun und Holland (1982) Equipercentile Linking-funktionen, wenn X und Y zufällige, kontinuierliche Variablen sind:

$ey(x) = G^{-1}[F(x)]$

und

$ex(y) = F^{-1}[G(y)]$

ey ist eine symmetrische Equating-Funktion, welche verwendet wird, um Testwerte des Tests X auf die Skala des Tests Y zu konvertieren.

ex ist eine symmetrische Equating-Funktion, welche verwendet wird, um Testwerte des Tests Y auf die Skala des Tests X zu konvertieren.

X ist eine zufällige Variable, welche einen Testwert des Tests X repräsentiert und x ist ein bestimmter Wert von X.

Y ist eine zufällige Variable, welche einen Testwert des Tests Y repräsentiert und y ist ein bestimmter Wert von Y.

F ist die kumulative Verteilungsfunktion von X in der Population.

F^{-1} ist die Umkehrfunktion der kumulativen Verteilungsfunktion F.

G ist die kumulative Verteilungsfunktion von Y in der Population.

G^{-1} ist die Umkehrfunktion der kumulativen Verteilungsfunktion G.

Entsprechend der Funktion des Equipercentile Linking können äquivalenten Ergebniswerte wie folgt gebildet werden. In einem ersten Schritt werden für jeden Rohwert des NEPS-Tests und des PISA- bzw. LV-Tests die kumulativen Prozente der Schülerinnen und Schüler, die diesen Wert oder einen niedrigeren Werte erreicht haben, sowie die dazugehörigen Perzentilränge berechnet. Anschließend wird jedem NEPS-Wert ein Testwert aus dem PISA- bzw. LV-Test mit dem gleichen kumulativen Prozentwert und Perzentilrang ($NEPS_{LV}$-Wert) zugeordnet (Kolen & Brennan, 2010). Haben zum Beispiel 1.2% der Schülerinnen und Schüler im NEPS-Test einen WLE -2.75 oder niedriger erreicht

und im LV-Test haben 1.2% der Schülerinnen und Schüler einen Wert von 370 oder niedriger erreicht, so wird angenommen, dass ein WLE von -2.75 im NEPS-Test die gleiche Kompetenz wie ein Wert von 370 im LV-Test repräsentiert. Das Equipercentile Equating wurde in dieser Studie mit der Computersoftware LEGS (Kolen & Brennan, 2004) durchgeführt. Die Computer Software LEGS nutzt lediglich positive ganzzahlige Werte. Aus diesem Grund wurde die NEPS Werte mit folgender Funktion transformiert: $x_t = rnd(\Theta * 100) + 500$.

Im Folgenden werden zunächst die Ergebnisse für das Linking des NEPS-Mathematiktests mit dem PISA-Mathematiktest und anschließend die Ergebnisse für das Linking des NEPS-Mathematiktests mit dem LV-Mathematiktest dargelegt.

Linking der Mathematiktests aus NEPS und PISA

In Tabelle 13 werden die Verteilungsmerkmale der Ergebnisse aus dem NEPS- und PISA-Test in der Linkingstudie gegenübergestellt. Die Schiefe und Kurtosis des PISA-Tests unterscheiden sich nicht signifikant von 0 bzw. 3 und deuten auf eine symmetrische Gaußsche Normalverteilung der Testergebnisse (siehe auch Abbildung 7). Die Testwerte aus dem NEPS-Test haben im Vergleich zum PISA-Test eine höhere positive Schiefe und eine niedrigere Kurtosis. Die Verteilung der Testwerte unterscheidet sich signifikant von einer gaußförmige Verteilung (siehe auch Abbildung 4).

Tabelle 13: Verteilungsmerkmale des PISA- und NEPS-Testwerte

	Min	Max	MW	SD	Skewness	Kurtosis
PISA	280	809	541	78	-0.02	3.09
NEPS	-3.47	3.75	0.53	1.27	0.15	2.57

In Abbildung 5 wird der Zusammenhang der mathematischen Testergebnisse aus NEPS und PISA abgebildet. Mit einer beobachteten Korrelation von $r = .68$ und einer latenten Korrelation von $r = .90$ zeigen die Tests einen starken Zusammenhang. Trotz des starken konzeptionellen und korrelativen Zusammenhanges der Studien wird in Abbildung 5 deutlich, dass die Testwerte nicht untereinander austauschbar sind. Beispielsweise erreichen Schülerinnen und Schüler mit einem WLE von Θ = -1 bei NEPS im LV Testwerte von ca. 380 bis 614.

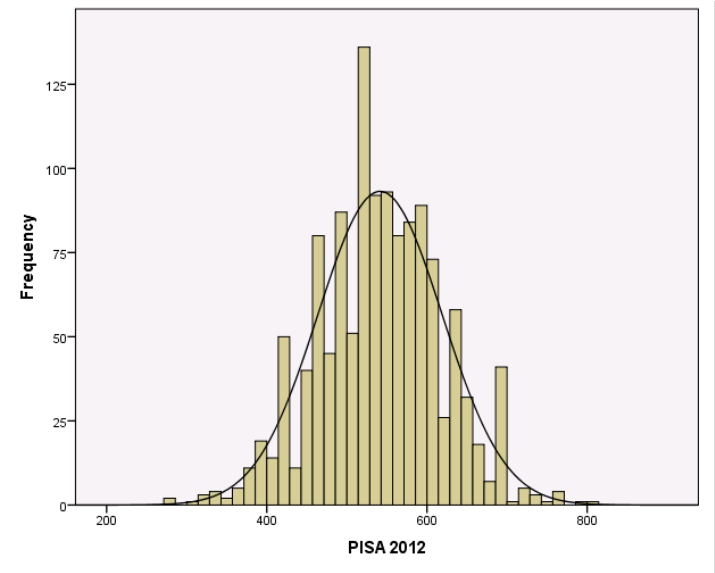

Abbildung 3: Verteilung der Mathematikergebnisse des PISA-Tests

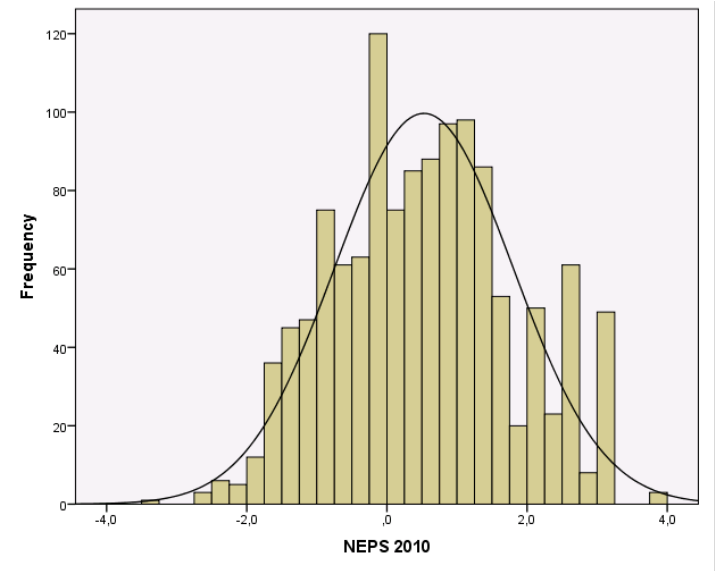

Abbildung 4: Verteilung der Mathematikergebnisse des NEPS-Tests

Für das Linking wurden in einem ersten Schritt die Perzentilränge der Testwerte im PISA-Mathematiktest festgestellt. In Tabelle 14 ist eine Auswahl an niedrigen, mittleren und hohen Testergebnissen im PISA-Test, die Häufigkeit, mit der diese von den Schülerinnen und Schülern erreicht wurden, die kumulative Häufigkeit, die prozentuale und kumulative prozentuale Häufigkeit sowie die dazugehörigen Perzentilränge angegeben.

Tabelle 14: Häufigkeiten der Testwerte und Perzentilwerte für den PISA-Test

PISA-Werte	Häufigkeit	Kumulative Häufigkeit	Prozentuale Häufigkeit	Prozentuale, kumulative Häufigkeit	Perzentilrang
280	1	1	0.00079	0.00079	0.039
284	1	2	0.00079	0.00157	0.118
313	1	3	0.00079	0.00236	0.197
326	3	6	0.00236	0.00472	0.354
341	2	8	0.00157	0.0063	0.551
342	2	10	0.00157	0.00787	0.709
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
529	14	575	0.01102	0.45276	44.724
533	14	589	0.01102	0.46378	45.827
535	12	601	0.00945	0.47323	46.850
537	16	617	0.01260	0.48583	47.953
539	13	630	0.01024	0.49606	49.094
543	23	653	0.01811	0.51417	50.512
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
752	1	1264	0.00079	0.99528	99.488
761	2	1266	0.00157	0.99685	99.606
767	1	1267	0.00079	0.99764	99.724
771	1	1268	0.00079	0.99843	99.803
793	1	1269	0.00079	0.99921	99.882
809	1	1270	0.00079	1	99.961

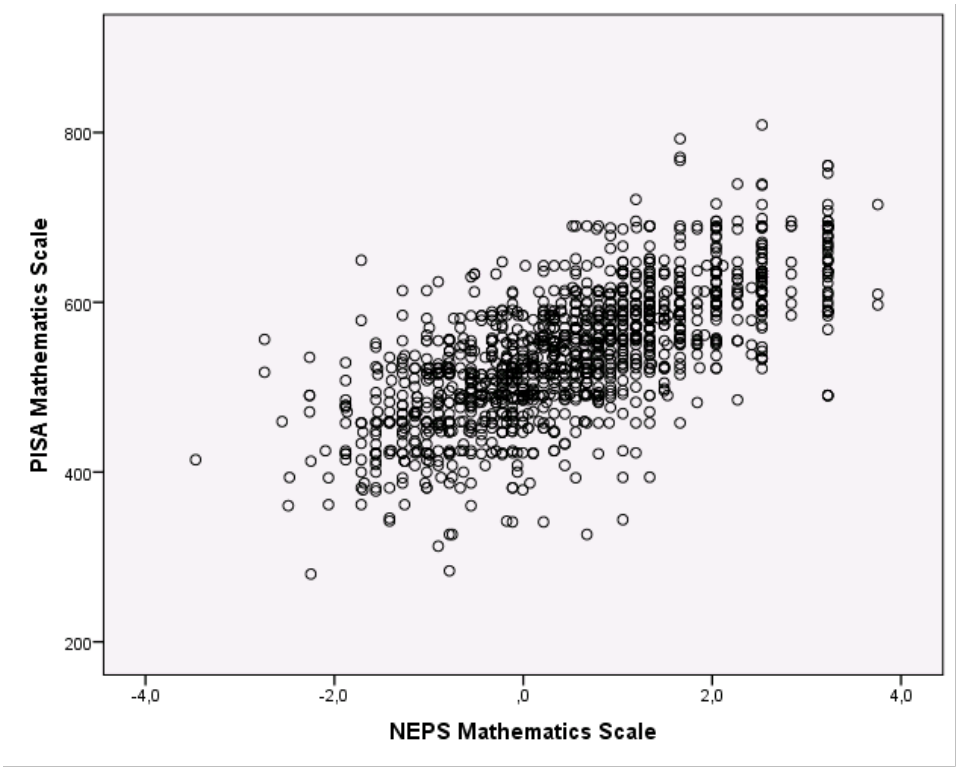

Abbildung 5: Streudiagramm der Testergebnisse für PISA und NEPS

Tabelle 15 zeigt die Ergebnisse des Linking für einige niedrige, mittlere und hohe Kompetenzwerte. Den NEPS-WLEs werden die positiv transformierten Werte, die Häufigkeit und kumulative Häufigkeit mit der X erreicht wurde sowie die prozentuale und kumulative prozentuale Häufigkeit mit der X erreicht wurde, zugeordnet. In den letzten beiden Spalten werden die Perzentilränge der NEPS-WLEs und der $NEPS_{PISA}$-Werte aus dem Equipercentile Equating dargestellt.

Tabelle 15: Testwert-Häufigkeiten und Perzentilwerte für NEPS und $NEPS_{PISA}$

NEPS-WLE	NEPS-Werte	Häufigkeit	Kumulative Häufigkeit	Prozentuale Häufigkeit	Prozentuale, kumulative Häufigkeit	Perzentilrang	$NEPS_{PISA}$ Wert
-3.47	153	1	1	0.00079	0.00079	0.039	280
-2.74	226	2	3	0.00157	0.00236	0.157	313
-2.55	245	1	4	0.00079	0.00315	0.276	326
-2.49	251	1	5	0.00079	0.00394	0.354	326
-2.48	252	1	6	0.00079	0.00472	0.433	326
-2.26	274	4	10	0.00315	0.00787	0.630	342
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
-0.01	499	28	474	0.02205	0.37323	36.220	516
0.02	502	1	475	0.00079	0.37402	37.362	518
0.03	503	1	476	0.00079	0.37480	37.441	518
0.05	505	8	484	0.00630	0.38110	37.795	518
0.07	507	1	485	0.00079	0.38189	38.150	522
0.08	508	1	486	0.00079	0.38268	38.228	522
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
2.27	727	19	1145	0.01496	0.90157	89.409	636
2.42	742	4	1149	0.00315	0.90472	90.315	637
2.53	753	61	1210	0.04803	0.95276	92.874	650
2.84	784	8	1218	0.00630	0.95906	95.591	686
3.22	822	49	1267	0.03858	0.99764	97.835	690
3.75	875	3	1270	0.00236	1	99.882	793

Für die Kontrolle der Robustheit des Linking wurde die Linkingfuntkion der Gesamtgruppe mit den Linkingfunktionen für die Subgruppe Geschlecht verglichen. Für den Vergleich wurden die Linkingfunktionen zunächst graphisch abgebildet (vgl. Abbildung 6). Anschließend wurden die Verteilungsmerkmale der PISA-Testwerte für die Gesamtgruppe und die Subgruppen den Verteilungsmerkmalen der $NEPS_{PISA}$-Werte gegenübergestellt (Tabelle 16). Das Linking zeigt große Übereinstimmungen zwischen der Gesamtgruppe, den Jungen und den Mädchen. Kleinere Abweichungen werden im oberen Kompetenzbereich (über einem NEPS-WLE von $\Theta = 2.4$) deutlich. Für diese Werte erhalten die Mädchen etwas niedrigere $NEPS_{PISA}$-Ergebniswerte als die Gesamtgruppe oder die Jungen. Im unteren Kompetenzbereich (unter einem NEPS-WLE Θ = -1.9) lassen sich etwas stärkere Abweichungen erkennen. In diesem Bereich erhalten die Mädchen höhere und die Jungen niedrigere $NEPS_{PISA}$-Ergebniswerte als die Gesamtgruppe. Die Abweichungen gelten allerdings nur für die zwei Prozent der Schülerinnen und Schüler mit den niedrigsten und die acht Prozent der Schülerinnen und Schüler mit den höchsten Kompetenzwerten.

Für die Gesamtgruppe unterscheiden sich der Mittelwert und die Standardabweichung kaum zwischen den Testwerten des PISA-Tests und den $NEPS_{PISA}$-Ergebniswerten. Abweichungen zeigen sich jedoch in der Schiefe und der Kurtosis der beiden Verteilungen. Jedoch unterscheiden sich beide Verteilungen nicht signifikant von einer Normalverteilung. Für die Subgruppe der Mädchen weichen die Mittelwerte und Standardabweichungen der PISA und der $NEPS_{PISA}$ Ergebniswerte ebenfalls nur geringfügig voneinander ab. Auch diese Verteilungen unterscheiden sich nicht signifikant von einer Normalverteilung. Die Subgruppe der Jungen zeigt bei annähernd gleichen Mittelwerten eine leicht höhere Standardabweichung, eine etwas niedrigere negative Schiefe und eine höhere Kurtosis für die PISA-Ergebnisse. Für die Subgruppen der Jungen unterscheidet sich Schiefe der $NEPS_{PISA}$-Verteilung signifikant von einer Normalverteilung.

Die kontinuierliche Kompetenzskala aus PISA ist in voneinander abgrenzende Abschnitte eingeteilt, welche als Kompetenzstufen bezeichnet werden. Diese Kompetenzstufen beschreiben die kognitiven Fähigkeiten der Schülerinnen und Schüler, die das jeweilige Kompetenzniveau erreicht haben. Die Schülerinnen und Schüler können anhand ihrer Testwerte in dem PISA-Test und anhand der $NEPS_{PISA}$ Ergebniswerte in die Kompetenzstufen aus PISA eingeordnet werden. Für einen Vergleich der Zuordnungen in die Kompetenzstufen mit dem PISA-Test und den $NEPS_{PISA}$ Ergebniswerten werden die mit beiden Tests entstehenden prozentualen Kompetenzstufenverteilungen gegenüberge-

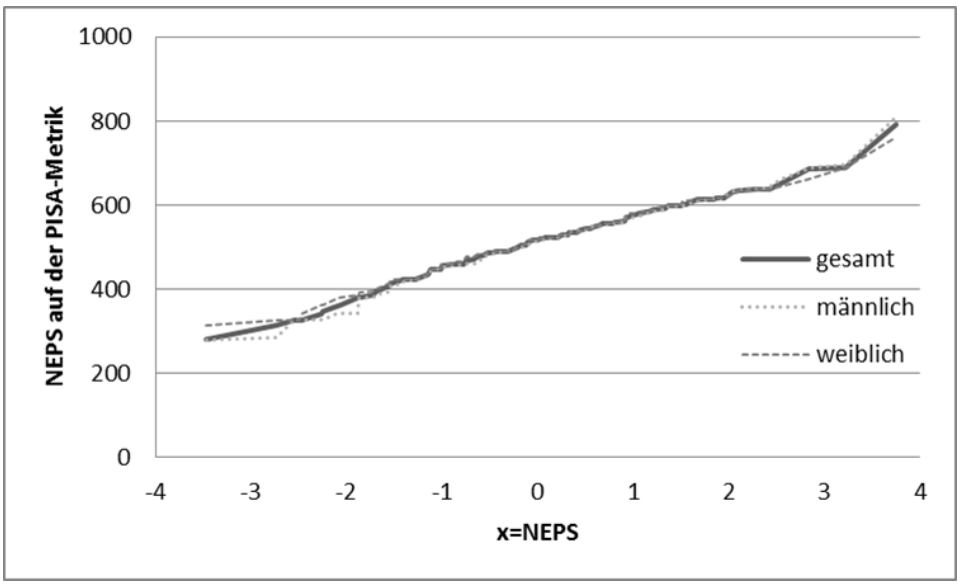

Abbildung 6: Zuordnung der NEPS-WLEs zu den äquivalenten Ergebniswerten auf der PISA-Metrik

stellt und mit einem χ_2-Test auf signifikante Unterschiede getestet. Ein Vergleich der prozentualen Einordnungen in die Kompetenzstufen wird in Tabelle 17 aufgezeigt. Insgesamt werden sowohl für die Gesamtgruppe, als auch für die Subgruppe Geschlecht keine signifikanten Unterschiede zwischen den Verteilungen auf die Kompetenzstufen gefunden. Die größten Unterschiede zwischen den Verteilungen basierend auf den PISA-Testwerten und den $NEPS_{PISA}$-Ergebniswerten können zwischen den Schülerinnen auf den Kompetenzstufen drei und vier gefunden werden. Hier beträgt die Abweichung jeweils vier Prozent.

Für einen Vergleich der individuellen Zuordnungen der Schülerinnen und Schüler werden die Kompetenzstufen der individuellen Schülerinnen und Schüler auf Basis des PISA-Tests und der $NEPS_{PISA}$-Ergebniswerten in einer Kreuztabelle gegenübergestellt. Auch diese Verteilungen werden mittels eines χ_2-Test auf Signifikanz getestet. In Abbildung 6 wird die Klassifikationskorrektheit der Zuordnung auf Basis der $NEPS_{PISA}$-Ergebniswerte gemessen an der Zuordnung auf Basis der PISA-Testwerte dargestellt. Als Maß für die Klassifikationskorrektheit werden die PÜ und das Cohens Kappa berechnet. Die individuellen Zuordnungen zu den Kompetenzstufen unterscheiden sich zwischen PISA und den $NEPS_{PISA}$ Ergebniswerten signifikant (χ_2= 467.76, df= 16). Mindestens 30% (Kompetenzstufe 1) und maximal 46% (Kompetenzstufe 3) der Schülerinnen und Schüler werden unabhängig des verwendeten Testwertes der gleichen Kompetenzstufe

Tabelle 16: Verteilungsmerkmale für PISA- und $NEPS_{PISA}$-Ergebnisse

		N	Min	Max	MW	SD	Schiefe	Kurtosis
PISA	Gesamt	1270	280	809	541	78	-0.02	3.09
	Männlich	637	280	809	553	80	-0.15	3.32
	Weiblich	633	313	761	529	74	0.06	2.93
$NEPS_{PISA}$	Gesamt	1270	280	793	541	76	-0.10	2.90
	Männlich	637	313	793	554	74	-0.26	2.85
	Weiblich	633	280	793	528	76	0.06	3.14

Tabelle 17: Prozentuale Verteilung der Schülerinnen und Schüler auf die Kompetenzstufen aus PISA anhand der PISA-Testergebnisse und anhand der $NEPS_{PISA}$-Ergebniswerte

		PISA-Kompetenzstufen							
		<1	1	2	3	4	5	6	gesamt
PISA	gesamt	1	4	16	30	28	15	5	100
	weiblich	1	5	20	34	24	13	3	100
	männlich	1	3	13	26	32	18	7	100
$NEPS_{PISA}$	gesamt	1	4	18	28	29	16	5	100
	weiblich	1	5	21	30	27	11	4	100
	männlich	0	3	14	25	31	21	5	100

zugeordnet. Insgesamt wird eine prozentuale Übereinstimmungen von PÜ = 42% und Cohens Kappa Werte von $\kappa = .26$ gefunden. Nach Pietsch et al. (2009) ist bei einer Reliabilität von 1.00 für den ersten Test und von .80 für den zweiten Test sowie einer Korrelation von $r = .90$ zwischen diesen Tests, eine Klassifikationskorrektheit von PÜ = 42% zu erwarten. Die mittlere Klassifikationskorrektheit der $NEPS_{PISA}$ Ergebniswerte ist mit PÜ = 42% daher als zufriedenstellend einzuschätzen. Insgesamt werden mit den χ_2 Ergebniswerten sehr ähnliche Verteilungen über die Kompetenzstufen erreicht wie mit den Testwerten des originalen PISA-Tests. Die Klassifikationskorrektheit auf Schülerebene ist zwar zufriedenstellend, jedoch weniger reliabel.

Tabelle 18: Prozentuale Übereinstimmung der Klassifikation individueller Schülerinnen und Schüler auf die Kompetenzstufen anhand der PISA- und der $NEPS_{PISA}$-Testwerte

		PISA-Kompetenzstufen						
		1	2	3	4	5	6	gesamt
$NEPS_{PISA}$-Kompetenzstufen	1	30	43	19	6	2	0	100
	2	13	41	35	9	2	0	100
	3	3	20	46	25	5	0	100
	4	1	5	29	43	19	4	100
	5	0	1	8	35	41	15	100
	6	0	0	5	22	37	37	100

Linking der Mathematiktests aus NEPS und LV

In Tabelle 19 werden die Verteilungsmerkmale der Ergebnisse aus dem NEPS- und LV-Test in der Linkingstudie gegenübergestellt. Die Schiefe und Kurtosis des LV unterscheiden sich nicht signifikant von 0 bzw. 3 und deuten auf eine symmetrische Gaußsche Normalverteilung der Testergebnisse (siehe auch Abbildung 7). Die Testwerte aus dem NEPS-Test haben im Vergleich zum LV-Test eine tendenziell höhere positive Schiefe, welche sich jedoch nicht signifikant von einer symmetrischen Verteilung unterscheidet. Die Kurtosis ist hingegen signifikant flacher als eine gaußförmige Verteilung (siehe auch Abbildung 8).

Tabelle 19: Verteilungsmerkmale der LV und NEPS Mathematiktests

	Min	Max	MW	SD	Schiefe	Kurtosis
LV	273	859	580	92	-0.01	3
NEPS	-3.47	3.75	0.54	1.32	0.11	2.56

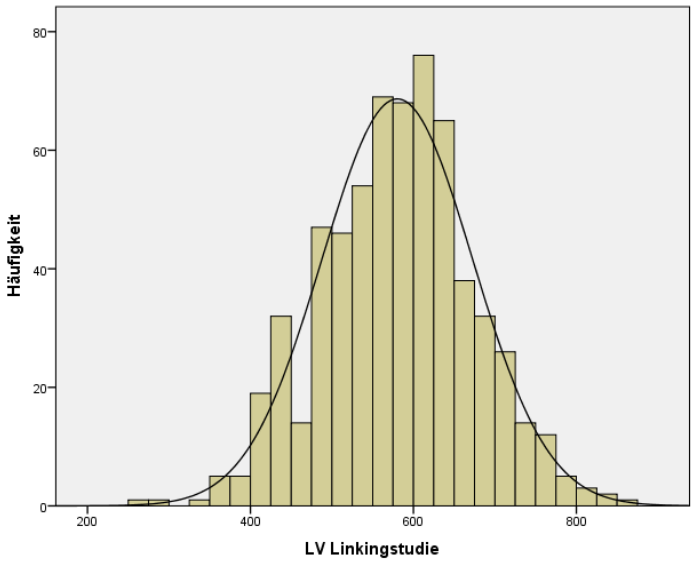

Abbildung 7: Verteilung der Mathematikergebnisse des LV-Tests

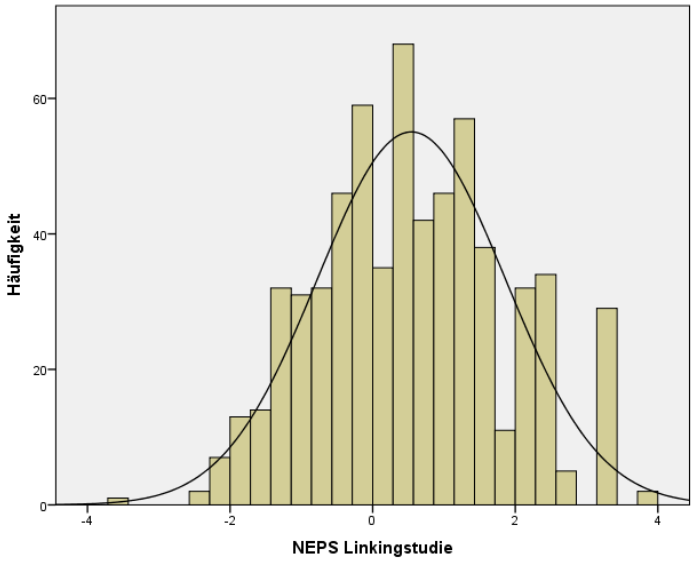

Abbildung 8: Verteilung der Mathematikergebnisse des NEPS-Tests

In Abbildung 9 wird der Zusammenhang der mathematischen Testergebnisse aus NEPS und LV abgebildet. Mit einer beobachtete Korrelation von $r = .74$ und einer latente Korrelation von $r = .92$ zeigen die Tests einen moderaten bis starken Zusammenhang. Trotz des starken konzeptionellen und korrelativen Zusammenhanges der Studien wird in Abbildung 9 deutlich, dass die Testwerte nicht untereinander austauschbar sind. Beispielsweise erreichen Schülerinnen und Schüler mit einem WLE von Θ = -1 bei NEPS im LV Testwerte von ca. 380 bis 640.

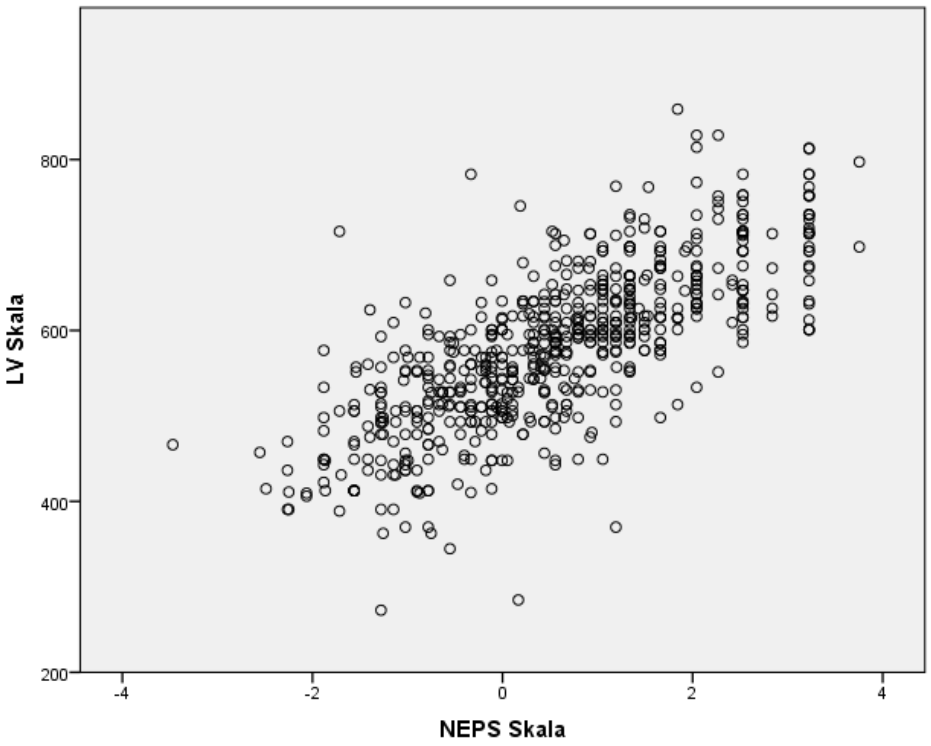

Abbildung 9: Streudiagramm der Testergebnisse für LV und NEPS

Für das Linking wurden in einem ersten Schritt die Perzentilränge der Testwerte im LV-Mathematiktest festgestellt. In Tabelle 20 ist eine Auswahl an niedrigen, mittleren und hohen Testergebnissen im LV-Test, die Häufigkeit, mit der diese von den Schülerinnen und Schülern erreicht wurden, kumulative Häufigkeit, die prozentuale und kumulative prozentuale Häufigkeit sowie die dazugehörigen Perzentilränge angegeben.

Tabelle 21 zeigt die Ergebnisse des Linking für einige niedrige, mittlere und hohe Kompetenzwerte. Den NEPS-WLEs werden die positiv transformierten Werte, die Häufigkeit und kumulative Häufigkeit mit der X erreicht wurde sowie die prozentuale und kumulative prozentuale Häufigkeit mit der X erreicht wurde zugeordnet. In den letzten beiden Spalten werden die Perzentilränge der NEPS-WLEs und $NEPS_{LV}$ Werte aus dem Equipercentile Equating im LV dargestellt.

Tabelle 20: Häufigkeiten der Testwerte und Perzentilwerte für den LV 2012-Test

Scores LV	Häufigkeit	Kumulative Häufigkeit	Prozentuale Häufigkeit	Prozentuale, kumulative Häufigkeit	Perzentilrang
273	1	1	0.00157	0.00157	0.079
285	1	2	0.00157	0.00314	0.236
345	1	3	0.00157	0.00472	0.393
363	2	5	0.00314	0.00786	0.629
370	3	8	0.00472	0.01258	1.022
389	1	9	0.00157	0.01415	1.336
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
577	15	309	0.02358	0.48585	47.406
579	1	310	0.00157	0.48742	48.664
585	7	317	0.01101	0.49843	49.292
586	12	329	0.01887	0.51730	50.786
591	3	332	0.00472	0.52201	51.965
593	13	345	0.02044	0.54245	53.223
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
797	1	630	0.00157	0.99057	98.978
812	1	631	0.00157	0.99214	99.135
813	1	632	0.00157	0.99371	99.292
814	1	633	0.00157	0.99528	99.450
828	2	635	0.00314	0.99843	99.686
859	1	636	0.00157	1	99.921

Tabelle 21: Score-Häufigkeiten und Perzentilwerte für NEPS und $NEPS_{LV}$

NEPS WLE	NEPS Score	Häufigkeit	Kumulative Häufigkeit	Prozentuale Häufigkeit	Prozentuale, kumulative Häufigkeit	Perzentil-rang	$NEPS_{LV}$ Wert
-3.47	153	1	1	0.00157	0.00157	0.079	273
-2.55	245	1	2	0.00157	0.00314	0.236	285
-2.49	251	1	3	0.00157	0.00472	0.393	345
-2.26	274	3	6	0.00472	0.00943	0.708	363
-2.25	275	2	8	0.00314	0.01258	1.101	370
-2.06	294	2	10	0.00314	0.01572	1.415	391
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
-0.11	489	23	216	0.03616	0.33962	32.154	533
-0.06	494	2	218	0.00314	0.34277	34.119	544
-0.01	499	19	237	0.02987	0.37264	35.770	551
0.05	505	5	242	0.00786	0.38050	37.657	553
0.07	507	1	243	0.00157	0.38208	38.129	553
0.1	510	12	255	0.01887	0.40094	39.151	557
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
2.27	727	8	566	0.01258	0.88994	88.365	693
2.42	742	3	569	0.00472	0.89465	89.230	698
2.53	753	31	600	0.04874	0.94340	91.903	713
2.84	784	5	605	0.00786	0.95126	94.733	730
3.22	822	29	634	0.04560	0.99686	97.406	758
3.75	875	2	636	0.00314	1	99.843	859

Für die Kontrolle der Robustheit des Linking wurde die Linkingfuntkion der Gesamtgruppe mit den Linkingfunktionen für die Subgruppen Geschlecht und Highest International Socio-Economic Index of Occupational Status (HISEI) verglichen. Dabei wurde die Unterteilung des HISEI in zwei Subgruppen so gewählt, dass die Subgruppe HISEI-Low mit bis zu 49 Punkten unter dem in PISA 2012 ermittelten Mittelwert von 50.6 Punkten (OECD, 2013) und die Gruppe HISEI-high gleich bzw. über dem Mittelwert liegt. Für den Vergleich wurden die Linkingfunktionen zunächst graphisch abgebildet (vgl. Abbildung 10 und 11). Anschließend wurden die Verteilungsmerkmale der LV Testwerte für die Gesamtgruppe und die Subgruppen den Verteilungsmerkmalen der $NEPS_{LV}$ Werte gegenübergestellt (Tabelle 22). Die Gesamtgruppe unterscheidet sich in Bezug auf den HISEI und das Geschlecht, da für 72 Schülerinnen und Schüler der HISEI nicht bekannt ist. Die Gesamtgruppe Geschlecht beinhaltet N = 636, die Gesamtgruppe HISEI hingegen nur N = 564 Schülerinnen und Schüler. Das Linking zeigt große Übereinstimmungen zwischen der Gesamtgruppe, den Jungen und den Mädchen. Kleinere Abweichungen werden im oberen Kompetenzbereich (über einem NEPS-WLE von $\Theta = 2.8$) deutlich. Für diese Werte erhalten die Mädchen etwas niedrigere $NEPS_{LV}$ Ergebniswerte als die Gesamtgruppe oder die Jungen. Im unteren Kompetenzbereich (unter einem NEPS-WLE $\Theta = -1.8$) lassen sich etwas stärkere Abweichungen erkennen. In diesem Bereich erhalten die Mädchen höhere und die Jungen niedrigere $NEPS_{LV}$ Ergebniswerte als die Gesamtgruppe. Die Abweichungen gelten allerdings nur für die drei Prozent der Schülerinnen und Schüler mit den niedrigsten und die sechseinhalb Prozent der Schülerinnen und Schüler mit den höchsten Kompetenzwerten.

Tabelle 22: Verteilungsmerkmale für LV- und $NEPS_{LV}$-Ergebnisse

		N	Min	Max	MW	SD	Schiefe	Kurtosis
LV	Gesamtgruppe Geschlecht	636	273	859	580	92	-0.01	2.99
	männlich	329	273	859	590	94	-0.1	3.11
	weiblich	307	345	828	570	90	0.06	2.93
	Gesamtgruppe HISEI	564	273	859	582	93	-0.03	3.04
	HISEI-Low	279	345	783	560	85	-0.04	2.63
	HISEI-High	285	273	859	603	95	-0.17	3.4
$NEPS_{LV}$	Gesamtgruppe Geschlecht	636	273	859	580	91	-0.06	2.9
	männlich	329	370	859	593	89	-0.12	2.66
	weiblich	307	273	859	565	92	0.03	3.23
	Gesamtgruppe HISEI	564	273	859	582	91	-0.08	2.98
	HISEI-Low	279	273	859	560	90	0.03	3.39
	HISEI-High	285	345	859	603	88	-0.18	2.78

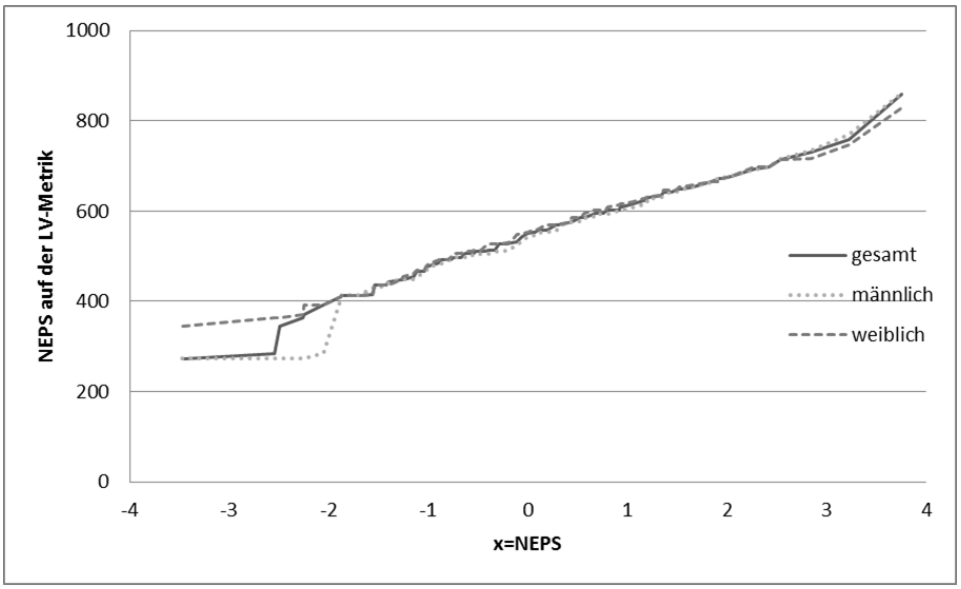

Abbildung 10: Ergebnisse des Equipercentile Equating zwischen NEPS-Scores und LV-Score-Equivalents differenziert nach Gesamtgruppe und Geschlecht

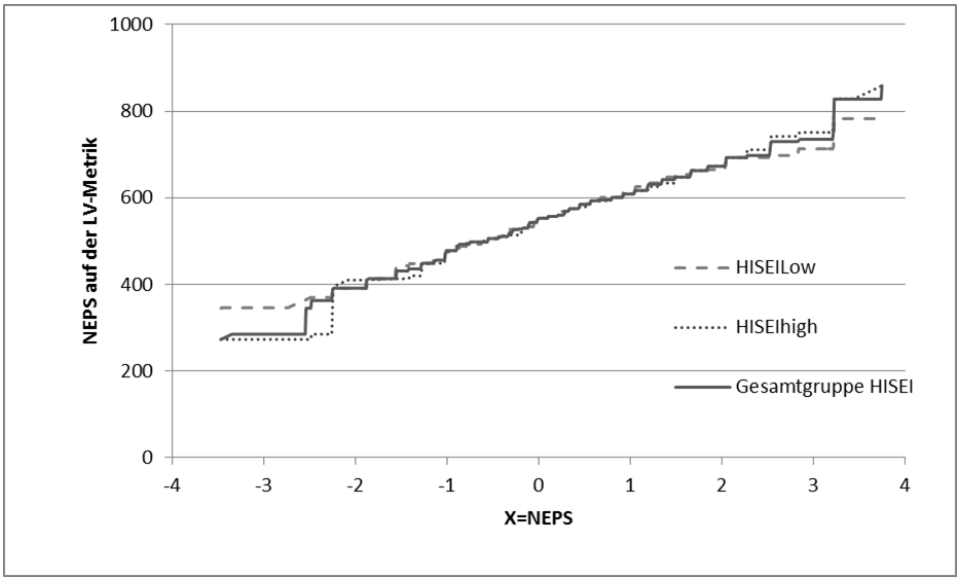

Abbildung 11: Ergebnisse des Equipercentile Equating zwischen NEPS-Scores und LV-Score-Equivalents differenziert nach Gesamtgruppe und sozioökonomischem Status (HISEI)

Die kontinuierliche Kompetenzskala des LV ist in voneinander abgrenzende Abschnitte eingeteilt, welche als Kompetenzstufen bezeichnet werden. Diese Kompetenzstufen beschreiben die kognitiven Fähigkeiten der Schülerinnen und Schüler, die das jeweilige Kompetenzniveau erreicht haben. Die Schülerinnen und Schüler können anhand ihrer Testwerte in dem LV-Test und anhand der $NEPS_{LV}$-Ergebniswerte in die Kompetenzstufen des LV eingeordnet werden. Für einen Vergleich der Zuordnungen in die Kompetenzstufen mit dem LV-Test und den $NEPS_{LV}$ Ergebniswerten werden die mit beiden Tests entstehenden prozentualen Kompetenzstufenverteilungen gegenübergestellt und mit einem χ_2-Test auf signifikante Unterschiede getestet. Ein Vergleich der prozentualen Einordnungen in die Kompetenzstufen wird in Tabelle 23 aufgezeigt. Insgesamt werden sowohl für die Gesamtgruppe, als auch für die Subgruppen Geschlecht und HISEI keine signifikanten Unterschiede zwischen den Verteilungen auf die Kompetenzstufen gefunden. Die größten Unterschiede zwischen den Verteilungen basierend auf den LV-Testwerten und den $NEPS_{LV}$ Ergebniswerten können zwischen den Schülerinnen auf den Kompetenzstufen zwei und vier gefunden werden. Hier beträgt die Abweichung jeweils fünf Prozent.

Tabelle 23: Prozentuale Verteilung der Schülerinnen und Schüler auf die Kompetenzstufen des LV anhand der LV-Testergebnisse und anhand der $NEPS_{LV}$-Ergebniswerte

		Kompetenzstufen					
		1	2	3	4	5	gesamt
LV	Gesamtgruppe	6	21	28	31	15	100
	weiblich	7	22	31	29	11	100
	männlich	5	20	24	33	18	100
	HISEI-Low	7	26	29	30	9	100
	HISEI-High	5	14	26	33	22	100
$NEPS_{LV}$	Gesamtgruppe	6	22	27	30	16	100
	weiblich	7	27	30	24	13	100
	männlich	5	18	24	35	19	100
	HISEI-Low	9	25	29	27	10	100
	HISEI-High	3	16	25	34	23	100

Für einen Vergleich der individuellen Zuordnungen der Schülerinnen und Schüler werden die Kompetenzstufen der individuellen Schülerinnen und Schüler auf Basis des LV-Tests und der $NEPS_{LV}$-Ergebniswerten in einer Kreuztabelle gegenübergestellt. Auch diese Verteilungen werden mittels eines χ_2-Test auf Signifikanz getestet. In Tabelle 24 wird die Klassifikationskorrektheit der Zuordnung auf Basis der $NEPS_{LV}$-Ergebniswerte gemessen an der Zuordnung auf Basis der LV-Testwerte dargestellt. Als Maß für die Klassifikationskorrektheit werden die PÜ und das Cohens Kappa berechnet. Die individuellen Zuordnungen zu den Kompetenzstufen unterscheiden sich zwischen LV und den $NEPS_{LV}$-Ergebniswerten signifikant (χ_2= 467.76, df= 16). Mindestens 43% (Kompetenzstufe 1 und 3) und maximal 54% (Kompetenzstufe 5) der Schülerinnen und Schüler werden unabhängig des verwendeten Testwertes der gleichen Kompetenzstufe zugeordnet. Insgesamt wird eine prozentuale Übereinstimmungen von PÜ = 48% und Cohens Kappa Werte von $\kappa = .31$ gefunden. Nach Pietsch et al. (2009) ist bei einer Reliabilität von 1.00 für den ersten Test und von .80 für den zweiten Test sowie einer Korrelation von $r = .90$ zwischen diesen Tests, eine Klassifikationskorrektheit von PÜ = 42% zu erwarten. Die mittlere Klassifikationskorrektheit der $NEPS_{LV}$ Ergebniswerte ist mit PÜ = 48% daher als sehr zufriedenstellend einzuschätzen.

Tabelle 24: Prozentuale Klassifikationskorrektheit der individuellen Schülerinnen und Schüler auf die Kompetenzstufen

		LV-Kompetenzstufen					
		1	2	3	4	5	Gesamt
$NEPS_{LV}$	1	43	49	6	0	3	100
	2	13	44	36	6	1	100
	3	1	25	43	26	4	100
	4	1	6	25	52	17	100
	5	0	0	3	43	54	100

Insgesamt werden mit den $NEPS_{LV}$ Ergebniswerten sehr ähnliche Verteilungen über die Kompetenzstufen erreicht wie mit den Testwerten des LV-Tests, welcher die originalen LV Aufgaben beinhaltet. Die Klassifikationskorrektheit auf Schülerebene ist zwar zufriedenstellend, jedoch weniger reliabel. Ziel dieser Studie ist es zu überprüfen, ob die Mindest-, Regel- und Optimalstandards des LV auf das NEPS übertragen werden können. Aus diesem Grund werden, ebenso wie schon bei den Kompetenzstufen, die

aus dem Linking entstehende Verteilungen in den Standards den auf dem LV-Test basierenden Verteilungen gegenübergestellt (Tabelle 25) und die Klassifikationskorrektheit (Tabelle 26) berechnet. Die Verteilungen der Schülerinnen und Schüler auf die Standards basierend auf LV und $NEPS_{LV}$ sind sowohl für die Gesamtgruppe, als auch für die Subgruppen sehr ähnlich. Es konnten keine signifikanten Unterschiede gefunden werden.

Die individuellen Zuordnungen zu den Mindest-, Regel- und Optimalstandards unterscheiden sich zwischen LV und den $NEPS_{LV}$ Ergebniswerten dahingegen signifikant (χ_2 = 296.8, df= 4). Über die Hälfte der Schülerinnen und Schüler werden unabhängig des verwendeten Testwertes dem gleichen Standard zugeordnet. Insgesamt wird eine prozentuale Übereinstimmungen von PÜ = 68% und Cohens Kappa Werte von κ = .43 gefunden.

Tabelle 25: Prozentuale Verteilung der Schülerinnen und Schüler auf die Mindest-, Regel- und Optimalstandards des LV anhand der LV Testergebnisse und anhand der $NEPS_{LV}$ Ergebniswerte

		LV			
		Mindeststandard	Regelstandard	Optimalstandard	Gesamt
LV	gesamt	26.60	58.50	14.90	100
	weiblich	28.70	59.90	11.40	100
	männlich	24.60	57.10	18.20	100
	HISEI-Low	32.60	58.80	8.60	100
	HISEI-High	19.30	58.90	21.80	100
$NEPS_{LV}$	gesamt	27.50	56.40	16.00	100
	weiblich	33.20	54.10	12.70	100
	männlich	22.20	58.70	19.10	100
	HISEI-Low	33.30	56.30	10.40	100
	HISEI-High	18.60	58.90	22.50	100

Die Untersuchungen zum Linking machen deutlich, dass der NEPS-K9-Mathematiktest nicht die gleiche mathematische Kompetenz misst, wie der PISA- oder der LV-Mathematiktest. Jedoch messen diese Tests sehr ähnliche Kompetenzen. Nach Kolen und Brennan (2010) messen Tests bei einem Linking von Testergebnissen, welche menschliche Merkmale oder

Tabelle 26: Prozentuale Klassifikationskorrektheit der individuellen Schülerinnen und Schüler auf die Kompetenzstufen anhand der $NEPS_{LV}$

		LV			
		Mindest-standard	Regel-standard	Optimal-standard	Gesamt
$NEPS_{LV}$	Mindeststandard	63	35	1	100
	Regelstandard	16	73	11	100
	Optimalstandard	0	46	54	100

Fähigkeiten widerspiegeln, fast immer zumindest geringfügig verschiedene Konstrukte, auch wenn diese ähnliche Namen tragen. Daher ist davon auszugehen, dass Unterschiede im Testergebnis, sowohl auf Konstruktverschiedenheit als auch auf Messfehlern beruhen (Kolen & Brennan, 2010). Dies spiegelt sich auch in den Ergebnissen der Linkingstudie wider. Das Linking ist vor allem in den Randbereichen ungenau, sowohl für die Gesamtgruppe, als auch für die Subgruppen Geschlecht und HISEI. Die individuelle Zuordnung der Kompetenzstufen ist sogar deutlich fehlerbehaftet. Das bedeutet laut Kolen und Brennan (2010) allerdings nicht, dass ein Linking zweier solcher Tests nicht bestimmt werden könne. Allerdings ist kein Linking, welches auf zwei Tests mit unterschiedlichen Konstrukten beruht, für jegliche Zwecke und alle Populationen adäquat, egal wie vollständig es durchgeführt wurde. Aufgrund der Linkingfehler können die $NEPS_{PISA}$ und $NEPS_{LV}$ Testergebnisse nicht mit der Glaubwürdigkeit austauschbarer Scores bzw. eines Linking zweier näher zusammenhängender Assessments interpretiert werden. Das Linking dieser Studie führt jedoch im mittleren Kompetenzbereich sowohl für die Gesamtgruppe, als auch für die Subgruppen Geschlecht und HISEI zu robusten Ergebniswerten und Verteilungen auf die Standards und Kompetenzstufen. Diese durch das Linking entstandenen Verteilungen können daher für globale Interpretationen genutzt werden. Beispielsweise können die hohen und niedrigen Kompetenzwerte des NEPS mit den Kompetenzstufen in PISA und den Standards im LV verglichen werden.

2.1.4 Ergebnisse zu den Effekten des BLK-Modellversuchsprogramms SINUS

In der Studie wurde unter anderem der Frage nachgegangen, welche Auswirkungen die zu untersuchenden Module *Aus Fehlern lernen* und *Verantwortung für das eigene Lernen stärken* im Gegensatz zu den anderen möglichen Modulen des Programms SINUS-Transfer auf den Mathematikunterricht haben. Die für diese Fragestellung herangezogenen Fragebögen beschränkten sich auf die der Lehrkräfte und der Schülerinnen und Schüler, bei denen das SINUS-Transfer-Programm im Mathematikunterricht eingesetzt wurde. Für die Gruppe der Lehrkräfte für das Fach Mathematik ergab sich daraus eine Stichprobe von n = 54 und für die Gruppe der Schülerinnen und Schüler von n = 1.127. Es handelte sich dabei jeweils um die neunte Klassenstufe, wobei 23 Klassen den Bildungsabschluss des Abiturs anstrebten und 31 Klassen andere Bildungsabschlüsse. Für die Auswertung wurden die Antworten auf ausgewählte Fragebogenitems zwischen den Schülerinnen und Schülern bzw. Lehrerinnen und Lehrern, welche die obengenannten Module belegten mit den Antworten der übrigen Schülerinnen und Schülern bzw. Lehrerinnen und Lehrern verglichen. Die zentralen Ergebnisse für das Modul Aus Fehlern lernen haben ergeben, dass sich diese Methode in zwei Bereichen besonders bewähren konnte. Zum einen wird eine negative Einstellung der Schülerschaft zum Mathematikunterricht gemindert. Zum anderen unterliegen dem Unterricht im Zusammenhang mit diesem Modul weniger Störungen. Das Modul Verantwortung für das eigene Lernen stärken grenzt sich insofern von den übrigen Modulen ab, als dass das Interesse der Lehrerinnen und Lehrer, die sich für das Modul entschieden haben, an der Schülerschaft stärker ausgeprägt ist. Dabei geht es um das Interesse, dass sich die Schülerschaft während des Mathematikunterrichts wohlfühlt, Spaß an der Mathematik hat und motiviert ist. Insofern geht es um basale Voraussetzungen, die die Lernaktivitäten fördern. Insgesamt wird für alle SINUS-Schulen eine geringe negative Einstellung der Schülerschaft gegenüber dem Mathematikunterricht, ein großes Interesse der Lehrkräfte an der Schülerschaft und häufige Aufgabenstellungen bzw. Vorgehensweisen, die der konstruktivistischen Didaktik gefunden. Teilweise sind dies auch die Bereiche, durch die sich die untersuchten Module hervorheben.

2.2 Positionen des zahlenmäßigen Nachweises

Die Positionen des zahlenmäßigen Nachweises befinden sich im Verwendungsnachweis, der zum Ende des Berichtzeitraums eingereicht wird.

2.3 Notwendigkeit und Angemessenheit der geleisteten Arbeit

Das Projekt untersuchte die Möglichkeit die Mathematik- und Naturwissenschaftstests aus NEPS mit den Studien PISA 2012 und Ländervergleich 2012 zu verankern. Der Forderung nach einer wechselseitigen Verbindung von nationalen und internationalen Testinstrumenten aus Large-Scale-Assessments (LSA) in Deutschland wurde damit Rechnung getragen. Für die Mathematik- und Naturwissenschaftstests aus dem NEPS wurden keine kriterialen Bezugsrahmen für die Interpretationen der Ergebnisse entwickelt. Diese Studie untersuchte, ob der nationale Bezugsrahmen aus LV und der internationale Bezugsrahmen aus PISA auf das NEPS übertragen werden können, um die Interpretationen der NEPS-Ergebnisse zu erweitern. Die bisherigen Untersuchungen zum BLK-Modellversuchsprogramm SINUS geben Hinweise für Effekte des Programms. Die Untersuchung hat einen wissenschaftlichen Erkenntnisfortschritt im Bereich der Large Scale Assessments ergeben, besonders im Bereich der Testwertinterpretationen der Mathematik- und Naturwissenschaftstests aus NEPS, PISA und LV sowie im Bereich der Linking-Methoden.

2.4 Voraussichtlicher Nutzen

Im Rahmen des Verwertungsplans wurden wissenschaftliche und anwendungsbezogene Verwertungsziele formuliert. Die Erreichung und Fortschreibung dieser Ziele wird in den folgenden Abschnitten dargelegt.

2.4.1 Wissenschaftliche Verwertungsziele

Das Projekt verfolgt das Ziel, die Kompetenzskalen für Mathematik und Naturwissenschaften von den Studien PISA 2012 und den länderübergreifenden Bildungsstandards in NEPS zu verankern. Somit lassen sich künftig die Ergebnisse der Kompetenzmessungen in NEPS auch an einem internationalen Referenzmaßstab oder im Vergleich mit den Ergebnissen einzelner Länder verorten. Es ist zu erwarten, dass die Ergebnisse der Studie durch die Publikationen und Vorträge auf Kongressen und Fachtagungen eine breite Öffentlichkeit erreichen werden. Des Weiteren wird erwartet, dass die Ergebnissen in die Interpretation der Daten aus den Naturwissenschafts- und Mathematiktests aus NEPS einfließen werden. Bei künftiger Auswertung der NEPS-Daten durch Wissenschaftlerinnen und Wissenschaftler können so gezielt Schülergruppen analysiert werden, die national bzw. international anerkannte Kompetenzerwartungen nicht erreichen bzw. übertreffen. Auf diese Weise können Risikofaktoren aber auch unterstützende Ressourcen identifiziert werden, welche die mathematische und naturwissenschaftliche Entwicklung der Schülerinnen und Schüler beeinflussen. Die vorliegende Studie kann darüber hinaus auch als Beispiel für weitere Linkingstudien mit anderen nationalen und internationalen Tests fungieren.

Eine Qualifizierung des wissenschaftlichen Nachwuchses erfolgte durch Einbindung von Studierenden in Form von zu erstellenden Abschlussarbeiten und studentischen Hilfskräften. Im Forschungsprojekt wurden insgesamt zwei studentische Arbeiten angefertigt. Diese Personen tragen die gewonnenen Erkenntnisse in andere Forschungseinrichtungen und in Schulen weiter.

Im Projekt bzw. mit Bezug auf die wissenschaftliche Fragestellung verfolgen außerdem zwei Mitarbeiterinnen derzeit ein Promotionsvorhaben zu folgenden Themen (Arbeitstitel):

(I) Ann-Katrin van den Ham (Lüneburg): Analysen zur Äquivalenz von PISA, den Bildungsstandards und NEPS.
(II) Helene Wagner (Kiel): Äquivalenz von Large-Scale-Assessments naturwissenschaftlicher Kompetenzen.

2.4.2 Anwendungsbezogene, bildungspolitische Verwertungsziele

Es ist zu erwarten, dass die empirischen Befunde über die langfristigen Effekte des Modellversuchprogramms SINUS bzw. SINUS-Transfer dazu beitragen können, künftige Schul- und Unterrichtprogramme zu verbessern.

Die nationalen Bildungsstandards der Kultusministerkonferenz stellen eine länderübergreifende, verbindliche Richtschnur dar und spezifizieren, welche Anforderungen die Schülerinnen und Schüler zu bewältigen in der Lage sein sollten. Längsschnittliche Studien, etwa zu Risikogruppen, welche die Bildungsstandards nicht erreichen oder aber auch zu Gruppen, welche diese Standards übertreffen sind besonders wichtig, um förderliche und hinderliche Bedingungsfaktoren zu identifizieren. Die wissenschaftlichen Ergebnisse durch die Übertragung der kriterialen Bezugsrahmen auf die längsschnittliche NEPS-Studie machen solche Analysen möglich. Es wird daher erwartet, dass diese Studie langfristig evidenzbasierte bildungspolitische Einflüsse haben wird (beispielsweise eine Durchführung Maßnahmen zur Stärkung von bildungsrelevanten Ressourcen).

2.5 Fortschritt bei anderen Stellen

Der Verbindung von Large Scale Assessments wurde in den letzten Jahren eine erhöhte Aufmerksamkeit in der Forschung zuteil. Zwischen diesem Forschungsprojekt und einem weiteren Projekt der Leuphana Universität Lüneburg in Kooperation mit dem Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik (IPN) gab es Synergieeffekte. Das Ziel dieses weiteren Forschungsprojektes war es, die Testergebnisse des NEPS Mathematiktests für die vierte Klassenstufe in die Kompetenzstufen des TIMSS-Rahmenkonzeptes für die vierte Klassenstufe einzuordnen.

Ein Vergleich der Rahmenkonzepte beider Studien wies auf große Übereinstimmungen zwischen den Tests, deckte jedoch auch einige Unterschiede auf. Beide Tests haben einen leicht unterschiedlichen Fokus. Während der NEPS-Test auf dem Konzept der Mathematical Literacy basiert, orientiert sich der TIMSS-Test an einem Curriculum Modell (Foy, Brossman & Galia, 2012; Weinert et al., 2011). Dennoch werden große Überschneidungen

zwischen den inhaltlichen und kognitiven Domänen gefunden. In einer weiteren Analyse wurden die Linking-Ergebnisse aus zwei statistischen Methoden verglichen, dem Equipercentile Equating und dem IRT-Linking. Beide Methoden führen zu ähnlichen Populationsmittelwerten und Schiefe der geschätzten äquivalenten Verteilung. Die Methode des Equipercentile Equating führt jedoch zu annähernd gleicher Standardabweichung und Kurtosis zwischen den TIMSS-Scores und den äquivalenten Ergebniswerte. Bei einem Vergleich der Verteilungen der äquivalenten Ergebniswerte und der Testergebnisse aus dem TIMSS-Tests auf die TIMSS-Kompetenzstufen werden mit beiden Methoden zu zufriedenstellende Konsistenzen gefunden (κ um 0.35). Mit dem Equipercentile Equating wird im Mittel mit PÜ = 44% eine höhere Klassifikationskorrektheit als mit dem IRT Linking (PÜ= 33%) gefunden. Bei einer Korrelation von r= .90 zwischen den Tests und Reliabilitäten von α = .82 für den TIMSS-Test (Foy, Martin, Mullis & Stanco, 2012) sowie EAP/PV Relibilität = .80 (Duchhardt & Gerdes, 2012) für den NEPS-Test, beträgt die maximal erwartete Klassifikationskorrektheit im Mittel PÜ = 42% (vgl. Pietsch et al., 2009)). Nach Nissen et al. (eingereicht) können die erreichten Konsistenzen beider Linking-Methoden als gute Annäherung betrachtet werden. Insgesamt führt das Linking somit zu reliablen Deskriptivstatistiken. Die Ergebnisse dieser Studie können für Interpretationen auf Populationsebene genutzt werden, sind jedoch für Interpretationen auf Individualebene ungeeignet.

2.6 Erfolgte oder geplante Veröffentlichungen

2.6.1 Vorstellung der Projektergebnisse auf wissenschaftlichen Tagungen

van den Ham, A-K., Ehmke, T., Müller, K., Sälzer, C., & Schroeders, U. (2013). Äquivalenz der Mathematik-Kompetenztests in der Sekundarstufe zwischen den Studien NEPS, Ländervergleich und PISA. In: *GEBF Abstractband: Bildungsverläufe über die Lebensspanne* (S. 196).

Schöps, K., Wagner, H., Hahn, I., & Pietsch, M. (2013). Drei Tests, ein Konstrukt? Ein Vergleich der Kompetenztests von PISA, den nationalen Bildungsstandards und dem Nationalen Bildungspanel. In: *GEBF Abstractband: Bildungsverläufe über die Lebensspanne*

(S. 197).

van den Ham, A-K., Nissen, A., Ehmke, T., & Richter, D. (2013). Linguistic Determines Mathematics: How Linguistic Item Characteristics Influence the Difficulty of Mathematics Test Items. In: *AEA Abstractband: International surveys, policy borrowing and national assessment* (S. 70).

van den Ham, A-K., Ehmke, Sälzer, C., & Schroeders, U. (2014). PISA, NEPS und BiSta – Sind die Kompetenzmessungen in Mathematik vergleichbar? In: *GEBF Abstractband: Die Perspektiven verbinden* (S. 36).

Wagner, H., Schöps, K., Hahn, I., Pietsch, M., & Rönnebeck, S. (2014). PISA, Bildungsstandards und das Nationale Bildungspanel (NEPS). Vergleich der Rahmenkonzepte und Validierung des NEPS-Testinstruments in den Naturwissenschaften. In: *GEBF Abstractband: Die Perspektiven verbinden* (S. 81).

van den Ham, A-K., Ehmke, T., Sälzer, C., & Schroeders, U. (2014). Validität des NEPS-Mathematiktests für die neunte Klasse. In Güntürkün, O. (Hrsg.): *49. Kongress der deutschen Gesellschaft für Psychologie: Supplement to Psychological Test and Assessment Modeling. 21. bis 25. September 2014 Ruhr-Universität Bochum. Die Vielfalt der Psychologie: Abstracts*. Pabst Science Publishers (S. 181).

Wagner, H., Schöps, K., Hahn, I., Pietsch, M., & Rönnebeck, S. (2014). PISA, Bildungsstandards und das Nationale Bildungspanel (NEPS). Vergleich der Rahmenkonzepte und Validierung des NEPS-Testinstruments in den Naturwissenschaften. In Güntürkün, O. (Hrsg.): *49. Kongress der deutschen Gesellschaft für Psychologie: Supplement to Psychological Test and Assessment Modeling. 21. bis 25. September 2014 Ruhr-Universität Bochum. Die Vielfalt der Psychologie: Abstracts*. Pabst Science Publishers (S.181).

2.6.2 Erfolgte Publikationen

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

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.

van den Ham, A.-K., Nissen, A., Ehmke, T., Sälzer, C., & Roppelt, A. (2014). Mathematische Kompetenz in PISA, IQB- Ländervergleich und NEPS- Drei Studien, gleiches Konstrukt? Unterrichtswissenschaft, 42(4), 321-341.

2.6.3 Geplante Publikationen

Ehmke, T., van den Ham, A.-K., Nissen, A., Sälzer, C., & Heine, J.-H. (in Vorbereitung). Measuring Mathematics in international and national Large Scale Assessment: Linking PISA with NEPS.

van den Ham, A.-K., Ehmke, T., Roppelt, A., & Nissen, A. (in Vorbereitung). Assessments verbinden, Interpretationen erweitern? Lässt sich die Mathematikskala des Ländervergleichs 2012 auf die Mathematikergebnisse aus dem Nationalen Bildungspanel übertragen?

Wagner, H., Schöps, K., Hahn, I. & Köller, O. (in Vorbereitung). Validität des NEPS-Naturwissenschaftstest für die neunte Klassenstufe (Arbeitstitel).

Wagner, H., Schöps, K., Hahn, I. & Köller, O. (in Vorbereitung). Linking the NEPS Science test and the Science test from PISA and the National Assement in Germany (Arbeitstitel).

Ehmke et al. (in Vorbereitung). Effekte des Modellversuchsprogramms SINUS auf das Kompetenzniveau (Arbeitstitel).

Literaturverzeichnis

Abedi, J. & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement In Education*, *14* (3), 219–234. doi: 10.1207/S15324818AME1403_2

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, USA: American Educational Research Association.

Artelt, C., Drechsel, B., Bos, W. & Stubbe, T. C. (2008). Lesekompetenz in pisa und pirls/iglu - ein vergleich. *Zeitschrift für Erziehungswissenschaft*, *10* (Sonderheft 10-08), 35–52.

Barbu, O. C. & Beal, C. R. (2010). Effects of linguistic complexity and math difficulty on word problem solving by english learners. *International Journal of Education*, *2*, 1–19.

B.C. Ministry of Education. (o. J.). *Linking fsa 2008 grade 4 reading results to pirls 2006: Determining pirls school level performance based on school fsa scores.* Zugriff am 09.04.2015 auf `http://www.bced.gov.bc.ca /assessment/nat_int_pubs.htm.`

Blossfeld, H.-P., Schneider, T. & Doll, J. (2009). Die längsschnittstudie nationales bildungspanel: Notwendigkeit, grundzüge und analysepotential. *Pädagogische Rundschau*, *63* (2), 249–259.

Blum, W., Neubrand, M., Ehmke, T., Senkbeil, M., Jordan, A., Ulfig, F. & Carstensen, C. H. (2004). Mathematische kompetenz. In M. Prenzel et al. (Hrsg.), *Pisa 2003. der bildungsstand der jugendlichen in deutschland : Ergebnisse des zweiten internationalen vergleichs* (S. 47–92). Münster: Waxmann.

Braun, H. I. & Holland, P. W. (1982). Observed-score testing equiating: A mathematical analysis of some ets equating procedures. In P. W. Holland & D. B. Rubin (Hrsg.), *Test equating* (S. 9–49). New York: Academic Press.

Cartwright, F. (2012). *Linking the british columbia english examination to the oecd combined reading scale.*

Cartwright, F., Lalancette, D., Mussio, J. & Xing, D. (2003). *Linking provincial student assessments with national and international assessments* (Bd. no. 005). Ottawa: British Columbia Ministry of Education.

Duchhardt, C. & Gerdes, A. (2012). *Neps technical report for mathematics – scaling results of starting cohort 3 in fifth grade: (neps working paper no. 19).* Bamberg.

Foy, P., Brossman, B. & Galia, J. (2012). Scaling the timss and pirls 2011 achievement data. In M. O. Martin & I. Mullis (Hrsg.), *Methods and procedures in timss and pirls 2011.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Foy, P., Martin, M. O., Mullis, I. V. & Stanco, G. (2012). Reviewing the timss and pirls 2011 achievement item characteristics. In M. O. Martin & I. Mullis (Hrsg.), *Methods and procedures in timss and pirls 2011* (S. 1–27). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Haag, N., Heppt, B., Stanat, P., Kuhl, P. & Pant, H. A. (2013). Second language learners' performance in mathematics: Disentangling the effects of academic language features. *Learning and Instruction*, *28*, 24–34. doi: 10.1016/j.learninstruc.2013.04.001

Hartig, J. & Frey, A. (2012). *Validität des tests zur überprüfung des erreichens der bildungsstandards in mathematik* (Nr. 1). Zugriff am 22.06.2012 auf `http://www.psycontent.com/content/ar7757012107r4k0/fulltext.pdf` doi: 10.1026/0012-1924/a000064

Klieme, E., Neubrand, M. & Lüdtke, O. (2001). Mathematische grundbildung: Textkonzeption und ergebnisse. In J. Baumert et al. (Hrsg.), *Pisa 2000* (S. 141–191). Opladen: Leske + Budrich.

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

Kolen, M. J. & Brennan, R. L. (2010). *Test equating, scaling, and linking: Methods and practices* (2. Aufl.). New York, NY: Springer New York.

NCES. (2014). *2011 naep-timss linking study: Technical report on the linking methodologies and their evaluations.*

Neidorf, T., Binkley, M., Gattis, K. & Nohara, D. (2006). *Comparing mathematics content in the national assessment of educational progress (naep), trends in international mathematics and science study (timss), and program for international student assessment (pisa) 2003 assessments.* Washington,

DC: National Center for Education Statistics. Zugriff am 15.06.2012 auf http://eric.ed.gov/PDFS/ED491692.pdf

Nissen, A., Ehmke, T., Köller, O. & Duchhardt, C. (eingereicht). 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*.

OECD. (2013). *Pisa 2012 assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy*. Paris: OECD Publishing. Zugriff auf http://dx.doi.org/10.1787/9789264190511-en

Phillips, G. W. (2009). *The second derivative: international benchmarks in mathematics for u.s. states and school districts*. Washington, D.C.: American Institutes for Research.

Pietsch, M., Böhme, K., Robitzsch, A. & Stubbe, T. C. (2009). Das stufenmodell zur lesekompetenz der länderübergreifenden bildungsstandards im vergleich zu iglu 2006. In D. Granzer, O. Köller & A. Bremerich-Vos (Hrsg.), *Bildungsstandards deutsch und mathematik* (S. 393–416). Weinheim, Basel: Beltz.

Pohl, S. & Carstensen, C. H. (2012). *Neps technical report – scaling the data of the competence tests: Neps working paper no. 14*. Otto-Friedrich-Universität, Nationales Bildungspanel: Bamberg.

Prenzel, M., Artelt, C. et al. (Hrsg.). (2007). *Pisa 2006: Die ergebnisse der dritten internationalen vergleichsstudie*. Münster: Waxmann.

Prenzel, M., Carstensen, C. H., Senkbeil, M., Ostermeier, C. & Seidel, T. (2005). Wie schneiden sinus-schulen bei pisa ab? *Zeitschrift für Erziehungswissenschaft*, *8* (4), 540–561.

Prenzel, M., Häußler, P., Rost, J. & Senkbeil, M. (2002). Der pisa-naturwissenschaftstest: Lassen sich die aufgabenschwierigkeiten vorhersagen? *Unterrichtswissenschaft* (30), 120–135.

Prenzel, M., Rost, J., Senkbeil, M., Häußler, P. & Klopp, A. (2001). Naturwissenschaftliche grundbildung: testkonzeption und ergebnisse. In J. Baumert et al. (Hrsg.), *Pisa 2000* (S. 192–250). Opladen: Leske + Budrich.

Prenzel, M., Schöps, K., Rönnebeck, S., Senkbeil, M., Walter, O., Carstensen, C. H. & Hammann, M. (2007). Naturwissenschaftliche kompetenz im internationalen vergleich. In M. Prenzel et al. (Hrsg.), *Pisa 2006: Die ergebnisse der dritten internationalen vergleichsstudie* (S. 63–105). Münster: Waxmann.

Rost, J. (1996). *Lehrbuch testtheorie, testkonstruktion*. Bern [u.a.]: Huber.

van de Vijver, F. J. (1998). Towards a theory of bias and equivalence. In J. A. Har-

kness (Hrsg.), *Cross-cultural survey equivalence* (Bd. no. 3, S. 41–65). Mannheim: Zentrum für Umfragen, Methoden und Analysen (ZUMA).

Weinert, S., Artelt, C., Prenzel, M., Senkbeil, M., Ehmke, T. & Carstensen, C. H. (2011). Development of competencies across the life span. *Zeitschrift für Erziehungswissenschaft*, *14* (S2), 67–86. doi: 10.1007/s11618-011-0182-7

Wolf, M. K. & Leon, S. (2009). An investigation of the language demands in content assessments for english language learners. *Educational Assessment*, *14* (3-4), 139–159. Zugriff am 25.06.2012 auf `http://content.ebscohost.com/pdf23_24/pdf/2009/7LS/01Jul09/47132359.pdf?T=P&P=AN` doi: 10.1080/10627190903425883

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.

Berichtsblatt

1. ISBN oder ISSN ---	2. Berichtsart (Schlussbericht oder Veröffentlichung) Schlussbericht

3. Titel
PISA, Bildungsstandards und die National Educational Panel Study (NEPS)
Vergleich der Rahmenkonzepte und Validierung der NEPS-Testinstrumente in den Naturwissenschaften und in der Mathematik

4. Autor(en) [Name(n), Vorname(n)]
Ehmke, Timo
Van den Ham, Ann-Katrin
Schöps, Katrin
Wagner, Helene
Hahn, Inga
Rönnebeck, Silke

5. Abschlussdatum des Vorhabens
März 2015

6. Veröffentlichungsdatum
---

7. Form der Publikation
Bericht

8. Durchführende Institution(en) (Name, Adresse)
Leuphana Universität Lüneburg
Scharnhorststr. 1
21335 Lüneburg

Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik (IPN)
Olshausenstr. 62
24118 Kiel.

9. Ber. Nr. Durchführende Institution
3

10. Förderkennzeichen
01LSA009

11. Seitenzahl
65

12. Fördernde Institution (Name, Adresse)

Bundesministerium für
Bildung und Forschung (BMBF)
53170 Bonn

13. Literaturangaben
36

14. Tabellen
26

15. Abbildungen
11

16. Zusätzliche Angaben

Handelt es sich z.B. um einen Fortsetzungsbericht, bitte angeben: Folgebericht zu (ggf. Berichtsnummer, Förderkennzeichen, Titel)

17. Vorgelegt bei (Titel, Ort, Datum)

Wird das Vorhabenergebnis in Ausnahmefällen nicht publiziert, jedoch - oder über die Publikation hinaus -fachlich interessierten Stellen in der Bundesrepublik Deutschland zugänglich gemacht, bitte Namen/Bezeichnung,Ort und Datum z.B. der Konferenz (Fachkongress, Statusseminar etc.) angeben

18. Kurzfassung

Die mathematische und naturwissenschaftliche Kompetenz von Jugendlichen am Ende der Sekundarstufe I wird in Deutschland derzeit in mehreren Schulleistungsstudien gemessen, wie z. B. im Programme for International Student Assessment 2012 (PISA 2012), im IQB-Ländervergleich 2012 (LV 2012) und in der National Educational Panel Study für die neunte Klassenstufe (NEPS K9). Offen bleibt jedoch, inwieweit diese Kompetenzmessungen vergleichbar sind und ob sie das gleiche Konstrukt mathematischer bzw. naturwissenschaftlicher Kompetenz erfassen. In dieser Studie soll daher die konzeptionelle, dimensionale und skalenbezogene Äquivalenz der Tests überprüft werden (vgl. van de Vijfer, 1998, Kolen & Brennan, 2010, Ehmke et al., 2014).

Für die Untersuchung wurde eine Stichprobe von 80 Schulen mit 1965 Schülerinnen und Schülern gezogen. Diese Schülerinnen und Schüler wurden gebeten an zwei Testtagen sowohl Aufgaben des NEPS-Tests als auch Aufgaben aus PISA 2012 und LV 2012 zu bearbeiten. Dabei wurde ein Multimatrixdesign verwendet.

In der vorliegenden Studie wurde zunächst die konzeptionelle Äquivalenz des K9-Mathematik- und Naturwissenschaftstests des NEPS mit den PISA-Tests zur Erfassung naturwissenschaftlicher und mathematischer Grundbildung aus 2012 und den Tests zur Überprüfung des Erreichens der Bildungsstandards in den Fächern Mathematik, Biologie, Chemie und Physik untersucht.

Zu diesem Zweck wurden die NEPS-Items von Expertinnen und Experten den verschiedenen Teilbereichen der Rahmenkonzeptionen von PISA 2012 und den Bildungsstandards aus 2012 zugeordnet. Zusätzlich wurden NEPS-Mathematikitems ebenfalls bezüglich ihrer linguistischen und sprachlichen Merkmale mit Mathematikitems aus dem LV 2012 und PISA 2012 verglichen. Die Ergebnisse zeigen, dass die drei Tests auf Ebene der naturwissenschaftlichen und mathematischen Rahmenkonzeptionen große Gemeinsamkeiten aufweisen. Auf Ebene der Mathematikaufgaben sind die Aufgabenstellungen bei PISA 2012 bezüglich der Wortschwierigkeiten und Komplexität der Satzstrukturen schwieriger als beim NEPS-K9. Die Aufgaben des NEPS-Mathematiktests der K9 unterscheiden sich von denen des LV 2012 vor allem durch mehr geschlossene Aufgabenformate und weniger mathematischer Begriffe.

In einem weiteren Schritt wurde die dimensionale Äquivalenz der Mathematik- bzw. Naturwissenschaftstest untersucht. Dafür wurden die Zusammenhänger der Inhaltsbereiche innerhalb der Tests verglichen und die Zusammenhänge der Inhaltsbereich zwischen den Tests analysiert. Außerdem wurden die Modellgütekriterien für eine gemeinsame Skalierung der Naturwissenschafts- bzw. Mathematiktests den Kriterien für getrennte Skalierungen gegenübergestellt. Insgesamt werden mit $r = .92$ bis $r = .96$ für die Naturwissenschaften und $r = .78$ bis $r = .92$ für Mathematik hohe Korrelationen zwischen den Inhaltsbereichen innerhalb der Tests gefunden. Sowohl für den NEPS-Naturwissenschaftstest als auch für den NEPS-Mathematiktest lassen sich keine Inhaltsbereiche finden, die sich vom PISA-Naturwissenschaftstest bzw. vom LV- und PISA-Mathematiktest abgrenzen lassen. Die Modellgütekriterien sind für die zweidimensionale Skalierung der Naturwissenschaftstests aus NEPS und PISA niedriger als für die eindimensionale Skalierung. Dieses Ergebnis reproduziert sich auch für die Skalierung der Mathematiktests aus NEPS und PISA bzw. NEPS und LV. Insgesamt lässt sich festhalten, dass sich die Naturwissenschaftstests aus NEPS und PISA dimensional sehr ähnlich, jedoch nicht äquivalent sind. Auch für die Mathematiktests aus NEPS und PISA bzw. NEPS und LV kann geschlossen werden, dass zwar eine große dimensionale Ähnlichkeit besteht, es sich jedoch nicht um äquivalente Tests handelt.

In einem letzten Schritt wurde die Möglichkeit einer Übertragung der Kompetenzstufen des PISA-Mathematiktests und der Kompetenzstufen sowie Mindest-, Regel- und Optimalstandards des LV-Mathematiktests auf die Testergebnisse des NEPS-K9-Mathematiktests überprüft. Dafür werden in einem ersten Schritt die Ähnlichkeit der Testwerte im NEPS-Mathematiktest und des PISA- bzw. LV-Mathematiktest statistisch geprüft. In einem zweiten Schritt werden die Skalen des NEPS-K9-Mathematiktests und des Mathematiktests aus PISA-2012 bzw. LV-2012 mit der Methode Equipercentile Equating verlinkt. Es zeigt sich, dass das Linking von NEPS und PISA bzw. NEPS und LV im mittleren Bereich nahezu linear ausfällt. Die Verteilung der Schülerinnen und Schüler auf die Kompetenzstufen und die Standards kann nahezu reproduziert werden. Zusammenfassend kann festgehalten werden, dass das Linking der Testergebnisse auf Gruppenebene stabil ausfällt. Eine Interpretation der NEPS Testergebnisse auf den Metriken von PISA und dem LV für einzelne Schülerinnen und Schüler ist jedoch nicht möglich.

19. Schlagwörter Large Scale Assessment, NEPS, Linking, PISA, Ländervergleich, Äquivalenz	
20. Verlag ---	21. Preis ---

Document Control Sheet

1. ISBN or ISSN ----	2. type of document (e.g. report, publication) Report
3. title PISA, IQB National Assessment Study and the National Educational Panel Study (NEPS) Comparison of the Frameworks and validation of the NEPS Mathematics and Science Tests	
4. author(s) (family name, first name(s)) Ehmke, Timo van den Ham, Ann-Katrin Schöps, Katrin Wagner, Helene Hahn, Inga Rönnebeck, Silke	5. end of project March, 2015 6. publication date --- 7. form of publication Report
8. performing organization(s) (name, address) Leuphana Universität Lüneburg Scharnhorststr. 1 21335 Lüneburg Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik (IPN) Olshausenstr. 62 24118 Kiel	9. originator's report no. 3 10. reference no. 01LSA009 11. no. of pages 65
12. sponsoring agency (name, address) Bundesministerium für Bildung und Forschung (BMBF) 53170 Bonn	13. no. of references 36 14. no. of tables 26 15. no. of figures 11
16. supplementary notes ---	
17. presented at (title, place, date) ---	

18. abstract

There are many national and international large-scale assessment studies which measure Mathematics competencies of primary and secondary school students, for example, the *Programme for International Student Assessment 2012*, the IQB National Assessment Study and the *National Educational Panel Study* in Germany. However, the comparability of the studies and the measured constructs remains unclear. Therefore the conceptual, dimensional and scale equivalence has to be analyzed (eg. van de Vijfer, 1998, Kolen & Brennan, 2010, Ehmke et al., 2014).

This study is based on a sample containing 80 schools and 1965 students. In a two-day assessment the students took items of the PISA 2012, the NA 2012 and NEPS test according to a multi matrix design.

In this study the conceptual equivalence of the NEPS, PISA and NA Mathematics framework, as well as the equivalence of the NEPS, PISA and NA Science Framework was analyzed. Therefore experts assigned the NEPS-items to the PISA 2012 and NA 2012 Mathematics and Science framework. In addition the NEPS Mathematics items were compared to the PISA and NA Mathematics items according to formal and linguistic characteristics. The outcomes show that the NEPS Mathematics and Science framework is very similar to the PISA and the NA Mathematics respective Science framework. According to the language demands the PISA items are more complex on word and sentence level. The NA has fewer items with multiple choice format and the items contain less mathematical terms.

In a further step the dimensional equivalence of the Mathematics respectively the Science tests were analyzed. Therefore the relationship of the content domains within the tests and the relationship of the content domains between the tests were examined. Furthermore the information criteria AIC, BIC and CAIC were calculated for scaling the NEPS Mathematics respectively Science test and the PISA Mathematics respectively Science tests on one dimension and for scaling a two dimensional model with the NEPS Mathematics respectively Science Test and the PISA Mathematics respectively the Science tests on two separate dimensions. The same is calculated for the NEPS an NA Mathematics and Science tests. Overall correlations between $r = .92$ and $r = .96$ for the Science content domains and $r = .78$ and $r = .92$ for the Mathematics content domains indicate a high overlap between the domains within the tests. Overall the content domains in the NEPS Science and Mathematics test cannot be differentiated from the PISA respectively NA domains. The information criteria are lower for the two dimensional scaling of the NEPS and PISA Science test. The same is found for the two dimensional models NEPS and PISA respectively NEPS and NA Mathematics test. Altogether is can be stated that there is a high dimensional overlap between the NEPS and PISA Science tests, but that they are not equivalent. The same can be stated for the NEPS and PISA respectively NEPS and Science Mathematics tests.

In the final step of the study the possibility of linking the NEPS and NA Mathematics tests maintaining the international (PISA) respectively the national (NA) reporting scale was analyzed. Therefore the equivalence of the NEPS and the PISA respectively NA test scores was examined. Afterwards the Mathematics Scale of the NEPS test was linked to the scale of the PISA respectively NA using the Equipercentile Linking method. It can be shown that in the mid-part of the score distribution the linking is quite linear. The linking almost identically reproduced the distribution of the students over the proficiency levels. The distribution to the proficiency levels, however, is less reliable. Overall, it can be stated that inferences are possible on a population level, but the results should not be reported or interpreted on an individual level.

19. keywords

Large Scale Assessment, NEPS, Linking, PISA, National Assessment, Equivalence

20. publisher	21. price
---	---