Assessments verknüpfen – neue Aussagen ermöglichen

Verlinkung der mathematischen Kompetenzmessungen im Primarbereich des Nationalen Bildungspanels mit TIMSS und dem Ländervergleich

Von der Fakultät Bildung

der Leuphana Universität Lüneburg zur Erlangung des Grades

Doktorin der Erziehungswissenschaft

- Dr. phil. –

genehmigte Dissertation von

Annika Nissen

geboren am 21.11.1980 in Eutin

Eingereicht am:	22.02.2017
Mündliche Verteidigung (Disputation) am:	06.11.2017
Erstbetreuer und -gutachter:	Prof. Dr. Timo Ehmke
Zweitgutachter:	Prof. Dr. Olaf Köller
Drittgutachter:	Prof. Dr. Dominik Leiß

Danksagung

Die Danksagung zu Beginn einer Dissertation ist zu einer schönen Tradition geworden. Schön deshalb, weil der Prozess von vielen Menschen unterstützt wird, denen man sich auf diese Weise erkenntlich zeigen kann.

Zu besonderem Dank bin ich Prof. Dr. Timo Ehmke verpflichtet. Vielen Dank für die fortwährende Förderung, die konstruktiven Vorschläge sowie die Unterstützung in allen Arbeitsphasen der Dissertation. Prof. Dr. Olaf Köller danke ich für die wertvollen Anregungen und die kritischen Kommentare, die zum Gelingen dieser Arbeit beigetragen haben. Bei Herrn Prof. Dr. Dominik Leiß möchte ich mich dafür bedanken, dass er Interesse an meiner Arbeit gezeigt hat und diese begutachtet hat. Darüber hinaus habe ich das Glück gehabt, viele hilfsbereite Kollegen zu haben (insbesondere Ann-Katrin van den Ham, Sören Odau, Svenja Hammer und Marcus Pietsch), die mir in allen Phasen meiner Arbeit hilfreiche Rückmeldungen und vor allem auch immer wieder Struktur gegeben haben. Mit ihnen konnte ich viele zielführende, methodische und auch inhaltliche Diskussionen führen, die mich in meiner Arbeit immer wieder vorangebracht haben. Weiterhin möchte ich mich bei meinen Kollegen für die tolle Zusammenarbeit bedanken, die mir immer viel Freude bereitet hat. An dieser Stelle möchte ich mich auch bei meinen Kollegen in Kiel und in Berlin bedanken, die an den Validierungsstudien 2011 mitgearbeitet haben und mit denen immer ein Ergebnisaustausch sowie interessante Diskussionen stattfanden. Zudem danke ich dem Verbundprojekt des Zentrums für internationale Bildungsvergleichsstudien (ZIB) und des Leibniz-Instituts für die Pädagogik der Naturwissenschaften und Mathematik (IPN) für die Finanzierung des Projekts. Selbstverständlich danke ich auch den Schülerinnen und Schülern, den Eltern, den Lehrerinnen und Lehrer sowie den Schulleiterinnen und Schulleitern, die an der Validierungsstudie 2011 teilgenommen haben und damit diese Studie erst ermöglicht haben. Mein besonderer Dank gilt meiner Mutter und meinen Freundinnen Isabelle Mahler, Marnie Bove und Maren Preuss. Sie waren immer für mich da und haben mich in allem unterstützt.

Ohne euch alle wäre diese Dissertation nie zustande gekommen. Vielen lieben Dank dafür.

Inhalt

Tabellenverzeichnis

Abbildungsverzeichnis

Einleitung

„Bildung entscheidet maßgeblich über die Chancen des Einzelnen auf gesellschaftliche Teilhabe und die Entwicklung der individuellen Potenziale."
(Bundesministerium für Bildung und Forschung)

Umso wichtiger ist es, ein Bildungssystem guter Qualität bieten zu können. Um diese Qualität überprüfen zu können, werden weltweit viele Schulleistungsstudien durchgeführt, welche die Qualität der Bildung messen und Hinweise über die Stärken und Schwächen des Systems liefern. In Deutschland fand mit der „Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring" (Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland, 2006) ein Paradigmenwechsel statt, von einem einstigen selbstverständlichen Vertrauen auf die Qualität des Bildungssystems hin zum Systemmonitoring, zur Rechenschaftslegung und zur Ergebnisorientierung. Ziel ist das systematische Zusammentragen von Informationen zum Bildungssystem und die Verknüpfung dieser Ergebnisse mit Maßnahmen zur Unterrichts- und Qualitätsentwicklung (Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland, 2006). Die Gesamtstrategie umfasst die stetige Überprüfung des Bildungssystems durch internationale Schulleistungsstudien, nationale Ländervergleiche, nationale Vergleichsarbeiten sowie eine gemeinsame Berichterstattung durch die Bundesrepublik und die Bundesländer. Mit der Teilnahme an internationalen Schulleistungsstudien (z. B. Trends in International Mathematics and Science Study; kurz: TIMSS) soll die Möglichkeit der Einordnung der Leistungsfähigkeit der Schülerinnen und Schüler in Deutschland in einem internationalen Zusammenhang gewährleistet werden. Die Verknüpfung von internationalen Schulleistungsstudien und den nationalen Ländervergleichen erfolgt durch das Institut zur Qualitätsentwicklung im Bildungswesen (IQB) in Berlin, welches ebenfalls für die Durchführung und Auswertung der nationalen Ländervergleiche und der Vergleichsarbeiten (VERA) verantwortlich ist. Ziel dieser Verknüpfung ist die Normierung der Bildungsstandards an den internationalen Maßstäben, damit sich weiterhin alle Bundesländer in einem internationalen Referenzmaßstab verorten können (Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland, 2006). Zusätzlich zu den durch die KMK beschlossenen

Studien (TIMSS, PISA, IGLU, Ländervergleiche und Vergleichsarbeiten) werden in Deutschland noch zahlreiche Studien durchgeführt, die weitere Facetten abdecken sollen, wie z. B. Gesamterhebungen für ein Bundesland oder auch Längsschnittstudien, welche die Kompetenzentwicklung über einen Zeitraum aufzeigen sollen, wie das Nationale Bildungspanel (NEPS).

NEPS ist eine vom Bundesministerium für Bildung und Forschung (BMBF) geförderte Studie, die im Längsschnitt die Kompetenzentwicklung sowie den Bildungsprozess und -verlauf über den gesamten Lebensverlauf in Deutschland untersucht (Blossfeld, Roßbach & Maurice, 2011). Die längsschnittliche Anlegung der Untersuchung erlaubt es, mögliche Einflussfaktoren für die unterschiedlichen Entwicklungen im Bildungsverlauf zu beschreiben. Ein Multi-Kohorten-Sequenz-Design mit sechs Startstichproben – orientiert an den Bildungsübergängen – ermöglicht eine möglichst schnelle Erhebung der Daten. Das Hauptaugenmerk der Studie liegt dabei auf fünf Säulen: Kompetenzentwicklung, Lernumwelten, Bildungsentscheidungen, Migrationshintergrund und Bildungsrenditen.

Im NEPS wird, wie in den durch die KMK beschlossenen Studien auch, u. a. mathematische Kompetenz gemessen. In den Rahmenkonzeptionen steht hierbei explizit, dass eine Anschlussfähigkeit an bereits bestehende nationale und internationale Testkonzeptionen angestrebt wurde. Hiermit wurde das Ziel verfolgt, bereits bestehende Aufgabenpools nutzen zu können sowie eine Voraussetzung für eine Verbindung der Kompetenzskalen im NEPS mit anderen Large Scale Assessments (LSA) zu schaffen (Ehmke et al., 2009). Eine Verbindung von zwei oder mehreren Tests kann auf zwei Weisen erfolgen. Im Rahmen von Äquivalenzstudien (1) wird die Ähnlichkeit der Testinstrumente, der Rahmenkonzeptionen und der Testspezifikationen untersucht, um Gemeinsamkeiten und Unterschiede herauszustellen, die bei der gemeinsamen Interpretation der Ergebnisse der Studien wichtige Informationen liefern können. Eine Äquivalenzstudie bietet damit u. a. die Möglichkeit, Unterschiede zwischen den Ergebnissen mehrerer Studien zu erklären (Cartwright, Lalancette, Mussio & Xing, 2003; Wu, 2010). Dies kann nicht nur für die Interpretation der Ergebnisse nützlich sein, sondern auch, um ein komplexeres Bild der Kompetenzen der Schülerinnen und Schüler zu erhalten als es mit nur einem Test möglich wäre (Neidorf, Binkley, Gattis & Nohara, 2006). Weiterhin lässt sich mit einer

Äquivalenzstudie überprüfen, ob die Standards der eigenen Studie höher oder niedriger sind als es in anderen Studien der Fall ist (Cartwright et al., 2003; Hambleton, Sireci & Smith, 2009). Die Ergebnisse aus der Äquivalenzuntersuchung bilden darüber hinaus die Interpretationsgrundlage für die Ergebnisse einer Linkingstudie (2), da der Grad an Übereinstimmungen zwischen den Studien Auswirkungen auf die Stabilität und Exaktheit des Linking haben kann und damit auch auf die möglichen Interpretationen der Ergebnisse. In einer Linkingstudie werden die Tests auf eine gemeinsame Metrik gebracht, so dass sich weitere bildungspolitische Interpretationsmöglichkeiten ergeben. Cartwright et al. (2003) nennt einen weiteren Vorteil einer Verlinkung von nationalen und internationalen Studien: LSA-Studien sind meist sehr teuer in der Entwicklung und Durchführung. Durch eine Verlinkung einer nationalen Studie mit einer LSA-Studie kann u. a. dem Versprechen Rechnung getragen werden, die Vielfalt und die Kosteneffektivität zu verbessern, indem die internationalen Benchmarks für andere Untersuchungen genutzt werden, da eine Entwicklung von Kompetenzstufenmodellen sehr kostenintensiv ist (Cartwright et al., 2003). Hartig und Frey (2012) stellen einen weiteren Vorteil einer solchen Verknüpfung heraus. Sie nutzen einen Vergleich mit einer bestehenden LSA-Studie für die Validierung eines neuen Testinstruments.

Der Fokus der vorliegenden Untersuchung ist der im NEPS eingesetzte Mathematiktest für die fünfte Jahrgangsstufe. Dieser wurde ausgewählt, weil zum einen die mathematische Kompetenz als eine der Schlüsselkompetenzen für eine kulturelle Teilhabe an der Gesellschaft gezählt wird und für die individuelle und gesellschaftliche Entwicklung als höchst bedeutsam angesehen wird (u. a. National Council of Teachers of Mathematics, 2005; Sälzer, Reiss, Schiepe-Tiska, Prenzel & Heinze, 2013). Zum anderen wurde der Mathematiktest für die fünfte Jahrgangsstufe ausgewählt, weil der Übergang von der Primarstufe in die Sekundarstufe einer der wichtigsten Übergänge im deutschen Bildungswesen ist (Berendes et al., 2011) und weil die Überprüfung der Kompetenz zu Beginn der fünften Jahrgangsstufe deutlich machen kann, ob die basalen Fertigkeiten in der Grundschulzeit erworben wurden, um in den folgenden Jahrgängen der weiterführenden Schulformen auf vorhandenes Vorwissen aufbauen zu können (u. a. KMK, 2005; Einsiedler, 2003).

Bezogen auf den NEPS-Mathematiktest der fünften Jahrgangsstufe ergeben sich bei einer Verbindung der Kompetenzskalen mit anderen LSA-Studien die folgenden Möglichkeiten:

(1) Eine Verknüpfung der Ergebnisse des NEPS-Mathematiktests mit einer internationalen Schulleistungsstudie, würde eine Verortung der Ergebnisse in einem internationalen Referenzmaßstab ermöglichen. Die Auswahl einer internationalen Schulleistungsstudie für den Vergleich erfolgte anhand dreier Kriterien: Die internationale Studie sollte

1. eine ähnliche Definition des mathematischen Konstrukts aufweisen,
2. in etwa zur gleichen Zeit stattfinden und
3. in etwa die gleiche Jahrgangsstufe erheben.

Die Erhebung der mathematischen Kompetenz in der fünften Jahrgangsstufe fand im Rahmen von NEPS im Herbst 2010 statt. Daher bot sich für einen internationalen Vergleich die TIMSS-Studie an, die im Frühjahr 2011 international die mathematischen Kompetenzen am Ende der vierten Jahrgangsstufe untersuchte (Bos, Wendt, Köller & Selter, 2012). TIMSS wird seit 1995 von der *International Association for the Evaluation of Educational Achievement* (kurz: IEA) durchgeführt und untersucht international die mathematische und naturwissenschaftliche Kompetenz der Schülerinnen und Schüler mit dem Ziel, die Lehre und das Lernen in diesen Fächern zu verbessern. Sie nutzen den internationalen Vergleich, um potenzielle Einflussfaktoren zu analysieren, die die unterschiedlichen Leistungen erklären können. TIMSS findet alle vier Jahre statt und erfasst die Kompetenzen in der vierten und achten Jahrgangsstufe in mehr als 60 Ländern. Deutschland beteiligte sich im Jahr 2011 an der Untersuchung in der vierten Jahrgangsstufe.

(2) Weiterhin bietet sich eine Verknüpfung mit dem Ländervergleich an, um bei der Interpretation der Ergebnisse des NEPS ebenfalls einen nationalen Referenzmaßstab ansetzen zu können (Stanat, Pant, Böhme & Richter, 2012). Die Auswahl der Studie erfolgte ebenfalls anhand der drei bereits unter (1) genannten Aspekte. Der Ländervergleich ist eine nationale Studie in Deutschland, die das Ziel verfolgt, das Erreichen der national verbindlichen Bildungsstandards der Kultusministerkonferenz (KMK) zu überprüfen und festzustellen, in welchen Bereichen noch weiterer Steuerungsbedarf besteht. Untersucht

werden alternierend sprachliche und mathematisch/naturwissenschaftliche Kompetenzen in der vierten Jahrgangsstufe alle fünf Jahre sowie in der neunten Jahrgangsstufe alle drei Jahre. 2011 wurden die mathematischen und naturwissenschaftlichen Kompetenzen in der vierten Jahrgangsstufe erfasst. Die Studie wird vom *Institut zur Qualitätsentwicklung im Bildungswesen* (IQB) in Berlin durchgeführt und ausgewertet.

(3) Eine Verknüpfung von NEPS mit den beiden nationalen und internationalen Studien bietet darüber hinaus die Möglichkeit der Übertragung der kriteriumsbasierten Kompetenzstufenmodelle aus diesen Studien, da für das NEPS bis zum jetzigen Zeitpunkt keine eigenen Kompetenzstufenmodelle entwickelt wurden.

(4) Darüber hinaus können die Ergebnisse aus der Äquivalenzuntersuchung der drei Studien Informationen zur Validität der Testwertinterpretationen des NEPS liefern.

Im Rahmen der vorliegenden Untersuchung soll daher aufgrund der Ähnlichkeit der Konstrukte, der zeitlichen Kongruenz sowie der Nähe bezüglich der untersuchten Jahrgangsstufen eine Verbindung des Mathematiktests für die fünfte Jahrgangsstufe des NEPS 2010 mit dem Ländervergleich Mathematik Primar 2011 (vierte Jahrgangsstufe) und TIMSS 2011 (vierte Jahrgangsstufe) erfolgen[1]. Ziel ist, die Voraussetzungen für eine Verlinkung der NEPS-Studie mit der nationalen Studie Ländervergleich sowie der internationalen Studie TIMSS im Sinne einer Äquivalenzstudie zu untersuchen sowie die anschließende Verlinkung der Studien vorzunehmen (Linkingstudie). Die Ergebnisse der Linkingstudie schaffen eine Vergleichbarkeit auf nationaler und internationaler Ebene, ermöglichen darüber hinaus eine Einordnung der NEPS-Testwerte in Kompetenzstufen und damit weitere bildungspolitische Interpretationsmöglichkeiten der NEPS-Testwerte. Beispielsweise könnten die Schülerinnen und Schüler, die die Mindeststandards nicht erreicht haben bzw. die die Regelstandards übertroffen haben, im Längsschnitt untersucht sowie mögliche Bedingungsfaktoren hierfür analysieren werden.

[1] Diese Studie wurde vom Zentrum für internationale Bildungsvergleichsstudien (ZIB) und dem Bundesministerium für Bildung und Forschung (BMBF) gefördert.

1 Theoretischer Hintergrund: Äquivalenz- und Linkingstudien

In dieser Arbeit soll das mathematische Testinstrument sowie die Verteilung der Testwertergebnisse für die Klassenstufe 4 von NEPS 2010 mit denen vom Ländervergleich 2011 und von TIMSS 2011 im Rahmen einer Äquivalenzuntersuchung auf ihre Ähnlichkeit hin untersucht werden. Die Gemeinsamkeiten und Unterschiede liefern wichtige Hinweise für eine gemeinsame Interpretation der Ergebnisse der Studien. Zudem sind diese Ergebnisse grundlegend für die Auswahl des Linking-Verfahrens, sowie für die Stabilität und Exaktheit des anschließenden Linking des NEPS-Mathematiktests mit den Mathematiktests vom Ländervergleich und TIMSS. In Kapitel 1.1 wird daher zunächst eine Einführung in das Vergleichen und Verlinken von Tests gegeben und es werden Herangehensweisen von anderen Äquivalenz- und Linkingstudien aufgezeigt und zusammengeführt sowie daraus die weitere Vorgehensweise für die nachfolgenden Analysen erarbeitet. Wie daran anschließend eine Verlinkung erfolgen kann, wird abschließend in Kapitel 1.2 erläutert.

1.1 Äquivalenz- und Linkingstudien vergleichen: Eine Einführung

Die Ergebnisse von nationalen und internationalen Studien lassen sich nicht ohne Weiteres miteinander in Beziehung setzen, weil die Testergebnisse auf unterschiedlichen Skalen berichtet werden. Das Problem lässt sich anhand unterschiedlicher Maßeinheiten von Temperatur (Temperaturskalen) beschreiben: Die Temperatur wird in unterschiedlichen Ländern in unterschiedlichen Maßeinheiten ausgegeben (z. B. Celsius und Fahrenheit). Möchte man nun die Temperatur in Berlin/Deutschland (gemessen in Celsius) mit der Temperatur in New York/USA (gemessen in Fahrenheit) vergleichen, muss eine Umrechnung der Temperaturskala Fahrenheit in die Temperaturskala Celsius vorgenommen werden (oder andersherum). Ohne diese Umrechnung ist ein Vergleich nicht möglich, da sich die beiden Temperaturskalen voneinander unterscheiden. Für die Umrechnung wird die folgende Formel verwendet: $\text{Temperatur}_C = (\text{Temperatur}_F - 32) \cdot 5/9$.

Die Transformation eines Ergebnisses auf der Skala 1 zu einem Ergebnis auf der Skala 2 wird in der Bildungsforschung Verlinkung (Linking) genannt (Holland, 2007). Die Ergebnisse von den drei ausgewählten Schulleistungsstudien (vgl. Einleitung) lassen sich jedoch nur miteinander verlinken, wenn ihre Ergebnisse auch tatsächlich auf einer

gemeinsamen Skala abgebildet werden können. Dies setzt jedoch voraus, dass die Studien ‚das Gleiche' auf ‚die gleiche Weise' messen, sonst können die Ergebnisse in einem Test keine akkurat geschätzten Ergebnisse auf der Skala des anderen Tests bereitstellen (Cartwright et al., 2003). Desto größer die Übereinstimmung zwischen den Studien ist, desto höher ist auch die Güte der Verlinkung.

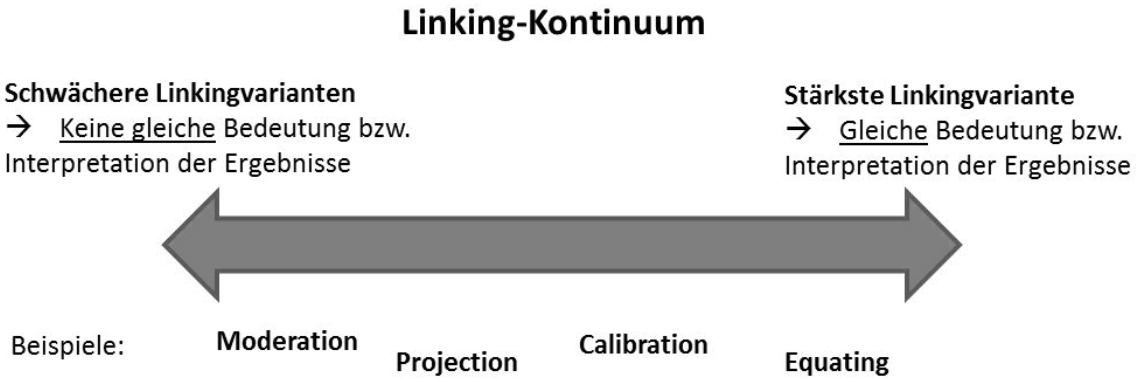


Abbildung 1.1: Linking-Kontinuum (in Anlehnung an Ryan & Brockmann, 2009)

Die stärkste Form des Verlinkens von zwei oder mehr Testvarianten ist das Equating (vgl. Abbildung 1.1: Linking-Kontinuum (in Anlehnung an Ryan & Brockmann, 2009)). Diese Form umfasst ausschließlich das Verlinken von parallelen Testvarianten (American Educational Research Association, American Psychological Association & National Council on Measurement in Education, 1999). Häufig ist es jedoch ebenfalls nützlich, Tests zu vergleichen, die nicht-parallel sind. Hierbei gibt es noch unterschiedliche Abstufungsgrade, z. B. in Calibration, Projection oder Moderation (Linn, 1993; Mislevy, 1992). Dies sind jedoch alles schwächere Linkingvarianten, die nur die jeweiligen Ziele einzelner Forschungsvorhaben erreichen oder für eine bestimmte Subgruppe gelten. Es kann jedoch nicht automatisch angenommen werden, dass die Ergebnisse über die Zeit stabil oder unveränderlich über verschiedene Subgruppen sind. Weiterhin ist unsicher, ob die Testwerte beim Nutzen unterschiedlicher Testvarianten gleich präzise sind. Nur bei einem Equating können die Ergebnisse untereinander austauschbar verwendet werden. Soll eine Verlinkung von zwei nicht-parallelen Tests vorgenommen werden, ist es daher entscheidend, die Güte des Linking zu untersuchen, indem die Unterschiede und Übereinstimmungen zwischen den Studien aufgezeigt werden sowie das Linking im Anschluss zu evaluieren und die Grenzen der Studien aufzuzeigen (Kapitel 1.2.3).

Im Folgenden wird basierend auf der geschichtlichen Entwicklung zunächst eine Einführung in das Verlinken von zwei oder mehreren Testvarianten gegeben. Es werden unterschiedliche Linking-Verfahren vorgestellt, die sich vor allem bezüglich der Stärke bzw. der Güte der Verlinkung unterscheiden. Anschließend werden die Konzeptionen und Ergebnisse von Linking-Studien vorgestellt, um Informationen darüber zu erhalten, wie in praktischen Studien das Vergleichen bzw. das Verlinken von zwei oder mehreren Testformen vorgenommen wird und welche Ergebnisse zu erwarten sind.

1.1.1 Das Verlinken von Tests – geschichtliche Entwicklung

Das Verlinken von Ergebnissen aus verschiedenen Testformen hat in den Bildungswissenschaften und in der Testbranche eine lange Tradition (u. a. Von Davier, 2011; Pommerich & Dorans, 2004; Kolen, 2004). Relevant wurde das Verlinken durch das vermehrte Einsetzen unterschiedlicher Testformen (von Davier, 2011). Diese wurden u. a. verwendet, um die Testsicherheit zu gewährleisten (Kolen & Brennan, 2010). Werden beispielsweise zwei Gruppen zu unterschiedlichen Zeitpunkten getestet, wäre die Testsicherheit verletzt, wenn beide Gruppen exakt denselben Test vorgelegt bekommen würden, da ein Austausch möglich wäre. Ebenso verhält es sich bei einer wiederholten Testung derselben Gruppe. Hier könnten Erinnerungseffekte auftreten und dadurch die Ergebnisse beeinflussen. Die Lösung ergab sich in der Erstellung von sogenannten Paralleltests (*alternate*, *parallel* or *equivalent forms*), die so äquivalent wie möglich sein sollen. Damit die Testvarianten möglichst kongruent sind, müssen sie u. a. auf den gleichen Testspezifikationen basieren und bei der Erstellung der Testvarianten müssen mehrere Aspekte berücksichtigt werden, z. B. Positionseffekte (Flanagan, 1951; Angoff, 1971). Dennoch sind die Testvarianten selten bzw. wenn überhaupt jemals tatsächlich äquivalent und unterscheiden sich beispielsweise leicht in der Schwierigkeit (Angoff, 1971; von Davier, 2011). Um dies auszugleichen wird das technische Verfahren bzw. der Prozess des *Equating* verwendet. Damit werden die Versionen einander angepasst, so dass die Ergebnisse aus den beiden Versionen nach der Umwandlung austauschbar (interchangeable scores) sind (Angoff, 1971). Das Equating erlaubt somit etwa die Messung von Zuwächsen und das Aufzeigen von Trends, allgemeiner gesagt, erlaubt das Equating das Zusammenführen von Daten bzw. Ergebnissen, die mit verschiedenen Testformen gemessen wurden (Angoff, 1971). Wäre der Prozess der Testerstellung perfekt, müsste das Verfahren des Equating

nicht verwendet werden (von Davier, 2011). Ein Equating basiert auf strengen Annahmen und folgt strikten Regeln bzw. Vorgaben. Dorans und Holland (2000) fassen fünf Annahmen zusammen, die für ein Equating gelten müssen, damit es erfolgreich ist und die Testscores austauschbar sind: (1) gleiches Konstrukt, (2) gleiche Reliabilitäten, (3) Symmetrie (die Equating-Funktion für das Verlinken von Test X zu Test Y sollte invers sein zum Verlinken von Test Y zu Test X), (4) Gerechtigkeit bzw. Testfairness (es sollte keine Rolle spielen, ob ein Schüler oder eine Schülerin an Test X oder Test Y teilnimmt) sowie (5) Populationsinvarianz. Diese Annahmen wurden jedoch häufig kritisiert und in der Praxis als unmöglich zu erreichen bzw. als schwierig zu evaluieren betitelt (u. a. Dorans & Holland, 2000; Livingston, 2004). Daher kam immer mehr der Wunsch nach weniger Grenzen und mehr Möglichkeiten auf (Pommerich, 2007).

In den 1950er und 1960er Jahren wurde vermehrt die Möglichkeit des Verlinkens von nicht-parallelen Testformen diskutiert (u. a. Angoff, 1957; Lindquist, 1964). Lindquist (1964) veranschaulicht anhand eines Linking von dem American College Testing Program (ACT) und der Scholastic Apitude Test (SAT) die vielen Probleme, die bei einem Linking auftreten können, wenn es sich um nicht-parallele Testformen handelt. Angoff (1971) deutet darauf hin, dass die Tauglichkeit eines solchen Linkings von zwei Fragen abhängt: (1) Wie ähnlich sind sich die beiden Tests, deren Testwerte verlinkt werden sollen? (2) Inwieweit ist die Gruppe, auf welcher die Vergleichswerte basieren, angemessen, hinsichtlich der Zielgruppe, für die der Vergleich genutzt werden soll. Für das Verlinken von nicht-parallelen Testformen wurden unterschiedliche Begriffe eingeführt bzw. definiert. Angoff (1971) beispielsweise behält den Begriff Equating für das Verlinken von Paralleltests bei. Als *Calibration* bezeichnet er die Verlinkung von Tests mit demselben Inhalt, aber unterschiedlicher Schwierigkeit oder Reliabilität. Sollen zwei Tests mit unterschiedlichen Konstrukten verlinkt werden, nutzt Angoff (1971) den Begriff *Comparability*.

Seit den 1980er Jahren wird dem Test-Linking vermehrte Aufmerksamkeit gewidmet. Kolen und Brennan (2010) lokalisieren für diesen Trend mehrere Entwicklungen, z. B. zeige der andauernde Anstieg in der Anzahl und der Vielfalt von Testprogrammen, die unterschiedliche Testformen nutzen, die Relevanz auf, diese Formen auch ordnungsgemäß zu verlinken. Ebenfalls benennen Kolen und Brennan den immer mehr in den Vordergrund rückenden Aspekt der Testfairness als Ursache für das erhöhte Auftreten von Testequating.

Anfang der 1990er Jahre veröffentlichen Mislevy (1992) und Linn (1993) ihre selbst entwickelten Rahmenkonzepte zum Linking von mehreren Testformen. Sie unterscheiden die statistischen Prozesse des *Equating*, *Calibration, Statistical Moderation* und *Projection* (bzw. *Prediction*). Sie setzen die Tradition fort und verwenden den Begriff Equating nur für das Verlinken von Paralleltests. Ebenfalls nutzen sie den Begriff Calibration in Übereinstimmung mit Angoff (1971) für das Verlinken von Tests, die auf demselben Konstrukt basieren, jedoch eine unterschiedliche Reliabilität und Schwierigkeit aufweisen. Hierzu zählt beispielsweise auch das Vertical Scaling. Die Linking-Variante Comparability nach Angoff (1971), die bei Tests mit unterschiedlichen Konstrukten Anwendung findet, unterteilen Mislevy (1992) und Linn (1993) in Statistical Moderation und Projection. Stastistical Moderation ist ein Prozess, bei dem die Testwerte jedes Tests mit einer dritten, einer sogenannten Moderatorvariablen, verlinkt werden. Unter Projection wird hingegen der Prozess verstanden, bei dem die Testwerte einer Testform die Testwerte einer anderen Testform etwa unter Verwendung der Regressionsmethode vorhersagen. Diese Rahmenkonzepte erlauben jedoch keine Unterscheidung zwischen dem Verlinken von zwei Tests mit sehr ähnlichen Konstrukten im Gegensatz zu zwei Tests mit sehr unterschiedlichen Konstrukten (Kolen, 2004).

Ende der 1990er Jahre veröffentlichen Feuer, Holland, Green, Bertenthal und Hemphill (1999) einen Bericht, in dem sie die Möglichkeit diskutieren, die Ergebnisse von Tests, die in den jeweiligen US-Staaten verwendet wurden, mit den Ergebnissen des *National Assessment of Educational Progress* (NAEP) zu verlinken. Um Unterschiede in den Tests auszumachen, sollten nach Feuer et al. die Tests auf drei Ebenen verglichen werden:

(1) Rahmenkonzept (framework definition): Hierunter wird der Rahmen der jeweiligen Studien verstanden, z. B. die Definition der Inhaltsbereiche und die Festlegung der erwarteten Fähigkeiten.
(2) Testspezifikationen (test specification oder auch blueprint): Zu den Testspezifikationen zählen Aspekte wie Itemformat, Anzahl von Aufgaben und die Regeln der Punkteverteilung.
(3) Itemauswahl (Itemselection): Die Items sollten so ausgewählt sein, dass sie die Testspezifikationen so genau wie möglich repräsentieren.

Diese drei Ebenen charakterisieren die sogenannte Domäne (domain), z. B. Mathematik in der vierten Jahrgangsstufe. Hieraus ergeben sich nach Feuert et al. die folgenden drei Linkingvarianten:

(1) Gleiches Rahmenkonzept und gleiche Testspezifikationen
(2) Gleiches Rahmenkonzept und unterschiedliche Testspezifikationen
(3) Unterschiedliches Rahmenkonzept und unterschiedliche Testspezifikationen

Mit dieser Unterteilung gehen Feuer et al. überwiegend konform mit dem Konzept von Mislevy (1992) und Linn (1993). Die erste Linkingvariante entspricht dem Equating bei Mislevy und Linn. Wenn das gleiche Rahmenkonzept und unterschiedliche Testspezifikationen vorliegen, liegt nach Mislevy und Linn Calibration vor. Die dritte Variante unterteilen Mislevy und Linn weiter in Moderation und Projection.

In 1999 erscheint eine Neuauflage der 'Standards for educational and psychological testing' (American Educational Research Association et al., 1999). Auch hier wird der Begriff des Equating festgelegt für das Verlinken von Paralleltests. Es wird jedoch zusätzlich eingeräumt, dass das Verlinken nicht-paralleler Tests oftmals nützlich ist. Diese Form wird als schwächere Form von Linking beschrieben, die zwar für einige Ziele bzw. für eine Subgruppe zufriedenstellende Ergebnisse liefern kann, jedoch nicht einfach als über die Zeit stabil bzw. als invariant für verschiedene Subgruppen angenommen werden kann. In den Standards für das Verlinken von Testformen wird viel Wert darauf gelegt, dass bei einer Veröffentlichung nicht nur die Linking-Ergebnisse, sondern vielmehr auch der Prozess des Linking detailliert beschrieben werden soll, z. B. sollten die genutzten Prozeduren, die verwendete Linking-Methode sowie – vor allem bei dem Verlinken von nicht parallelen Testformen – die Grenzen der Untersuchung aufgezeigt werden (American Educational Research Association et al., 1999).

Dorans (2004) greift das bereits zuvor vielfach benannte Problem auf, dass bei zwei zu verlinkenden Tests die Unterschiede in den Konstrukten größer und kleiner sein können. Er nutzt den Begriff Concordance, um Linkingvarianten von Tests zu beschreiben, deren Konstrukte zwar nicht gleich, sich jedoch sehr ähnlich sind. In dem Fall sollte das Linking auch zu annähernd gleichen Testwertverteilungen führen. Neben den ähnlichen Konstrukten sollte der Inhalt auch als vergleichbar beurteilt werden, die Testwerte sollten

ebenfalls hoch korrelieren und die Linkingbeziehungen sollten sich von einer Gruppe Testteilnehmer zu einer anderen nur wenig unterscheiden. Liegt keine Concordance vor, sollten nach Dorans Regressionsmethoden für ein Linking verwendet werden.

Kolen und Brennan (2010) benennen in ihrem Buch ‚Test equating, scaling und linking' ebenfalls den Grad an Übereinstimmung (Degrees of Similarity) als alternative Herangehensweise. Dabei sollten sich Äquivalenzuntersuchungen nach Kolen und Brennan (2010) auf mindestens vier Aspekte beziehen: Schlussfolgerungen (Inferences), Konstrukte (Constructs), Stichprobe (Population) und Messbedingungen (Measurement characteristics/conditions). Im Rahmen einer Äquivalenzanalyse kann untersucht werden, in welchem Maße sich die zu vergleichenden Studien bezüglich der angegebenen Aspekte ähneln bzw. unterscheiden. Im Folgenden werden nach Kolen und Brennan (2010) beispielhaft Fragen formuliert:

(1) Schlussfolgerungen: Welche Ziele verfolgen die Studien? Welche Schlussfolgerungen lassen sich aus den Testergebnissen ziehen? Haben die Studien unterschiedliche Funktionen bzw. Messintentionen?
(2) Konstrukte: Messen die Tests dasselbe Konstrukt bzw. zu welchem Grad? Beispielsweise messen alle Tests im Rahmen der vorliegenden Arbeit mathematische Kompetenz, aber die Frage ist, ob das Konstrukt in den drei Studien tatsächlich dasselbe ist. Ziel ist es daher Gemeinsamkeiten und Unterschiede zwischen den Rahmenkonzeptionen aufzuzeigen.
(3) Stichprobe: Für wen sind die Tests konstruiert wurden (z. B. für welches Alter)? Gibt es unterschiedliche Ausschlusskriterien für die Teilnahme an den Studien?
(4) Messeigenschaften und -bedingungen: Unter welchen Umständen hat die Messung stattgefunden (z. B. Durchführungsbedingungen) bzw. wie sind die Tests aufgebaut (z. B. Testlänge, Aufgabenformat)?

Van de Vijver (1998) konnte im Rahmen von kulturvergleichenden Studien zeigen, dass sich der Aspekt der Konstruktäquivalenz noch weiter unterteilen lässt. Van de Vijver betont die Relevanz der Identifikation von Konstrukt-, Methoden- oder Itembias bei Vergleichen von mehreren Studien. Die Bias entstehen, wenn die Studien nicht äquivalent sind. Wird

also das Ziel verfolgt, zwei Studien miteinander zu vergleichen, sollte die Konstruktäquivalenz auf drei Ebenen untersucht werden (van de Vijver, 1998; He & van de Vijver, 2012; Pietsch, Böhme, Robitzsch & Stubbe, 2009; Maehler & Schmidt-Denter, 2013):

a) Konzeptioneller Vergleich (construct equivalence)*:* Sind die Assessments der Studien konzeptionell vergleichbar?
b) Dimensionaler Vergleich (measurement unit equivalence)*:* Ist die faktorielle Struktur des Konstrukts gleichwertig?
c) Skalenbezogener Vergleich (scalar equivalence)*:* Erlauben die Skalenkennwerte eines Konstrukts eine gleichwertige Interpretation?

Unterschiede zwischen den Studien bezüglich der genannten Aspekte können zu systematischen Verschiebungen und zu Fehlinterpretationen führen. In solch einem Fall wäre der Vergleich der Studien nicht angemessen (van de Vijver, 1998).

Tabelle 1.1: Maße an Übereinstimmungen nach Kolen und Brennan sowie dem Schema von Mislevy und Linn (Kolen & Brennan, 2010, S.435)

Kategorie	**Inferenzen**	**Konstrukte**	**Stichprobe**	**Mess-bedingungen**
Equating	gleich	gleich	gleich	gleich
Vertical scaling	gleich	gleich/ ähnlich	ungleich	gleich/ ähnlich
Concordance	gleich	gleich	gleich/ ähnlich	(un)gleich
Projection	(un)gleich	(un)gleich	gleich	ungleich
Stat. moderation	(un)gleich	(un)gleich	(un)gleich	ungleich

Bezugnehmend auf die vorherigen Entwicklungen lassen sich nach Kolen und Brennan (2010) einige Gemeinsamkeiten feststellen. Beispielsweise sei die Anforderung des gleichen Rahmenkonzepts nach Feuer et al. (1999) im Wesentlichen vergleichbar mit der Frage nach der Ähnlichkeit der Konstrukte und der gleichen Testspezifikationen mit den Messbedingungen. Weiterhin greifen Kolen und Brennan einige der Kategorien von Mislevy (1992) und Linn (1993) bzw. ähnlich bezeichnete Kategorien auf und stellen sie ihren Kriterien der Maße an Übereinstimmungen gegenüber (vgl. Tabelle 1.1), wobei die

Zuordnung nicht immer eindeutig erfolgt. Dies liegt nach Kolen und Brennan (2010) daran, dass die Kategorien nach Mislevy (1992) und Linn (1993) zum Teil weiter gefasst sind und mehrere Möglichkeiten umfassen, wodurch keine eindeutige Spezifikation bei den Maßen an Übereinstimmung gegeben werden kann. Dennoch gibt die Tabelle 1.1 einen grundlegenden und guten Überblick über die beiden genannten Perspektiven des Linking.

1.1.2 Äquivalenz- und Linkingstudien – Herangehensweisen und Befunde

Unabhängig davon, nach welchem Linking-Verfahren in Äquivalenzuntersuchungen vorgegangen wird, werden meist zunächst die Übereinstimmungen und Unterschiede verschiedener Tests aufgezeigt und Berechnungen zu empirischen Zusammenhängen durchgeführt, um anschließend eine Übertragung der Skalen vorzunehmen. Je nachdem, welches Ziel mit einer solchen Untersuchung verfolgt wird, werden jedoch unterschiedliche Schwerpunkte gesetzt bzw. einzelne Schritte außen vor gelassen. So streben manche Untersuchungen nur einen Vergleich auf der Ebene inhaltlicher Gegenüberstellungen an, andere Studien berechnen zusätzlich empirische Zusammenhänge und einige Studien evaluieren darüber hinaus die statistische Exaktheit und Stabilität ihrer Ergebnisse.

Der Fokus in diesem Abschnitt liegt darauf, was in den Linkingstudien genau verglichen wurde, welche Methoden Verwendung fanden, welche Probleme aufgetreten sind und welche Verbesserungsvorschläge gemacht werden. Eine Zusammenfassung der Ergebnisse soll ermöglichen Schlussfolgerungen daraus zu ziehen, welche der untersuchten Aspekte relevant für die eigene Untersuchung sein könnten und welche Ergebnisse zu erwarten sind. Die Auswahl der Linking-Studien ist nur exemplarisch und auf Untersuchungen begrenzt, die sich mit der Äquivalenz von zwei oder mehreren überwiegend mathematischen Kompetenztests beziehen. Einen Überblick über die im Folgenden beschriebenen Studien liefert Tabelle 1.2. Weitere Übersichten über Äquivalenz- und Linkingstudien geben beispielsweise Ehmke (2014) und Feuer et al. (1999).

Ein Beispiel für Concordance liefert der Vergleich vom College Board's Scholastic Aptitude Test (SAT) und American College Testing Program (ACT), der bereits häufig gemacht wurde und der in der Literatur ebenfalls häufig zur Erläuterung der Theorie der Concordance herangezogen wird (u. a. Pommerich, 2007; Kolen, 2004; Lindquist, 1964; Marco, Abdel-fattah & Baron, 1992; Dorans, 2004). Beide Tests sind College-

Aufnahmetests in den USA. Das Problem besteht darin, dass die Schülerinnen und Schüler sowohl den ACT als auch den SAT bearbeiten können und die Colleges auch beide Tests zulassen (Kolen, 2004). Unklar bleibt jedoch, wie die Testergebnisse zueinander stehen und ob ein Schüler, der in dem einen Test mit erreichten 1200 Punkten den Zugangsvoraussetzungen entspricht, ebenfalls die Zugangsvoraussetzungen erfüllt hätte, die für den anderen Test gelten. Abhilfe würde eine Konkordanz-Tabelle schaffen, die offenlegt, wie sich die beiden Tests zueinander verhalten und welche Testwerte vergleichbar sind. Dorans (2004) stellt die Testinstrumente aus dem SAT und ACT zunächst auf einer inhaltlichen Ebene gegenüber, da die beiden Tests weder auf den gleichen Testspezifikationen basieren noch dieselben Konstrukte messen und vergleicht daher u. a. die erfassten Kompetenzen und die jeweilige Aufgabenanzahl. Bei seinen weiteren Analysen bezieht er sich auf eine Studie bei der 103 525 Schülerinnen und Schüler sowohl den ACT als auch den SAT bearbeitet haben. Um die beiden Tests zu verlinken, wurde das equipercentile Linkingverfahren verwendet (Dorans, Lyu, Pommerich & Houston, 1997). Anschließend stellt Dorans (2004) die empirischen Zusammenhänge zwischen den beiden Tests dar, die sich je nach Untertest unterscheiden. Die höchste Korrelation ergab sich mit r = .89 zwischen dem ACT Mathematiktest und dem SAT I Mathematiktest. Des weiteren untersucht Dorans die Invarianz von Subgruppen (Geschlecht) und stellt auch hier wiederum Unterschiede zwischen den jeweiligen Untertests heraus, z. B. sind der verbale Test des SAT I und der ACT Lesetest ähnlicher ausgerichtet als der ACT Englischtest.

Blum et al. (2004) stellen die Messwerte des nationalen Programme for International Student Assessment (PISA) Mathematiktests 2003 dem internationalen PISA Mathematiktest 2003 gegenüber. Hierzu vergleichen sie die empirischen Schwierigkeiten der beiden Tests, berichten Korrelationen zwischen den nationalen und internationalen Subskalen der Inhaltsbereiche und vergleichen die nationalen und internationalen Mathematikaufgaben hinsichtlich ihrer curricularen Validität (Stoff bis 9. Klasse behandelt, Vertrautheit mit der Aufgabenstellung und Bedeutsamkeit der Förderung der Kompetenz). Sie fanden heraus, dass der nationale Ergänzungstest etwas schwieriger ist als der internationale Test. Die Aufgaben, die die Kompetenz des begrifflichen Modellierens erfassen, sind – im Vergleich zum internationalen Test – die Schwierigsten. Die latente Korrelation zwischen den Tests beträgt r = .92. Zusätzlich wurden Korrelationen zwischen den Inhaltsbereichen der beiden Tests berechnet. Es zeigte sich, dass die Subskalen des

nationalen Tests mit den Subskalen des internationalen Tests zusammenhängen, jedoch nicht identisch sind (zwischen r = .82 und r = .90). Blum et al. konkludieren daraus, dass der internationale Mathematiktest bezüglich nationaler Fragestellungen eine hohe Validität aufweist. Bezüglich der curriculuaren Validität der PISA-Tests halten Blum et al. fest, dass der nationale Mathematiktest mit 89.5 % an Aufgaben, deren Stoff nach Einschätzung der Experten bis zur 9. Klassenstufe behandelt wird, etwas lehrplannäher ist als der internationale Test (83.2 Prozent). Unterschiede zwischen den Tests zeigten sich hinsichtlich der Vertrautheit mit der Aufgabenstellung. Die Experten schätzten die nationalen Aufgaben als überwiegend vertraut ein, wohingegen sie die internationalen Aufgaben als eher weniger vertraut einschätzten. Hinsichtlich der Wichtigkeit der Förderung der Kompetenz wurden die Aufgaben in beiden Tests als hoch bedeutsam eingestuft.

Grønmo und Olsen (2007) vergleichen die Anlagen und die Ergebnisse von TIMSS 2003 und PISA 2003, mit dem Ziel, Informationen darüber zu erhalten, wie die Ergebnisse aus den beiden Studien zusammenfassend interpretiert werden können, um ein ganzheitlicheres Bild der mathematischen Kompetenzen der Schülerinnen und Schüler aus den unterschiedlichen Ländern zu erhalten. Da viele Länder an beiden Studien teilnehmen, ergibt sich die Möglichkeit, die Gemeinsamkeiten und Unterschiede der mathematischen Bildung in einem internationalen Vergleich näher zu analysieren. Hierzu untersuchen Grønmo und Olsen (2007) zum einen die Rahmenkonzeptionen der Studien und zum anderen formale Charakteristika der Mathematikitems. In diesem Zusammenhang zeigen sie auf, dass die Studien zum Teil ein unterschiedliches Konstrukt mathematischer Kompetenz operationalisieren. Dadurch bedingt können sich die Ergebnisse eines Landes sowie die Interpretation der Ergebnisse in den beiden Studien unterscheiden. Grønmo und Olsen zeigen auf, dass die TIMSS-Studie vor allem Basiswissen und mathematische Prozeduren testet (pure mathematics) wohingegen die PISA-Studie das Lösen mathematischer Probleme in Realsituationen in den Vordergrund stellt (applied mathematics). Hieraus ergibt sich ein weiterer Unterschied. Die meisten TIMSS-Aufgaben kommen überwiegend ohne einen Kontextbezug aus und sind damit rein innermathematisch, die PISA-Aufgaben weisen hingegen durch den Realitätsbezug stets einen Kontext auf. Anschließend werden die unterschiedlichen Zielsetzungen der Studien gegenübergestellt. Grønmo und Olsen zeigen u. a. auf, dass die TIMSS-Studie das Ziel hat

zu analysieren, ob die teilnehmenden Länder die curricularen Zielsetzungen erreichen (Curriculum-Modell). Die PISA-Studie hingegen analysiert, ob die Schülerinnen und Schüler mathematische Kompetenzen im Lösen von Alltagsproblemen aufweisen. In einem weiteren Schritt analysieren sie 22 Länder und Regionen, um die Unterschiede in den Ergebnissen der beiden Studien zu erklären. Hierzu vergleichen sie zunächst die unterschiedlichen Ränge, die die Länder in den beiden Studien erreichen. Anschließend versuchen sie mögliche Ursachen für diese Unterschiede zu identifizieren, indem sie die Aufgaben der Studien hinsichtlich formaler Merkmale (z. B. das Itemformat oder das Vorhandensein von grafischen Repräsentationsformen) klassifizieren sowie eine Klassifikation der PISA-Aufgaben in die Rahmenkonzeptionen der TIMSS-Studie vornehmen. Hinsichtlich des Rankings stellen Grønmo und Olsen fest, dass die Länder-Rankings der beiden Studien eine Übereinstimmung von r = .76 aufweisen. Für fünf der 22 Länder und Regionen zeigen sie zusätzlich relative Stärken und Schwächen in den Inhaltsbereichen der Studien auf und stellen u. a. fest, dass sich die Länder in den TIMSS-Inhaltsbereichen stärker unterscheiden als in PISA. Bezüglich der Vergleiche der Aufgaben halten Grønmo und Olsen fest, dass PISA mehr Aufgaben beinhaltet, die das Lesen, Interpretieren und Evaluieren von Daten erfordern. Weiterhin ergaben die Analysen, dass die PISA-Aufgaben mehr grafische Repräsentationsformen nutzen und die Aufgaben häufiger mit einem offenen Antwortformat gestellt werden. Aus den Ergebnissen konkludieren Grønmo und Olsen, dass hohe Kompetenzen in der angewandten Mathematik allerdings nur erreicht werden können, wenn die Grundlagen vorhanden sind. Daher darf für die Schulmathematik das mathematical Literacy Konzept nicht als eine Alternative zu dem pure mathematics Konzept angesehen werden.

Hambleton et al. (2009) verlinken NAEP 2003 mit TIMSS 2003 und zusätzlich mit PISA 2003, um herauszufinden, ob die Ansprüche der nationalen Studie höher bzw. niedriger sind als in internationalen Studien. Bei ihrer Untersuchung konzentrieren sie sich auf die Mathematiktests für die achten Jahrgangsstufen (NAEP und TIMSS) bzw. für die 15-Jährigen (PISA), die in 2003 eingesetzt wurden. NAEP wird von dem National Center for Education Statistics (NCES) durchgeführt. Die Studie hat zum Ziel, darüber zu informieren, welche Fähigkeiten die Schülerinnen und Schüler im Elementarbereich (Klassenstufe 4) und im Sekundarbereich (Klassenstufe 8 und 12) u. a. in den Fächern Mathematik und Naturwissenschaften in den einzelnen Bundesstaaten bzw. in großen Schuldistrikten und

in der Nation USA haben. NAEP unterscheidet drei Kompetenzstufen: Grundkenntnisse (Basic), solides schulisches Wissen (Proficient) und fortgeschrittenes Wissen (Advanced). Hambleton et al. (2009) verwenden die Methode des Equipercentilen Linking, um die Kompetenzstufen der NAEP Studie mit den entsprechenden internationalen Skalen zu verlinken. Dies erlaubt eine Ermittlung der prozentualen Verteilung der Schülerinnen und Schülern in den teilnehmenden Ländern zu den NAEP-Kompetenzstufen. Als Ergebnis halten sie fest, dass einige andere Länder einen höheren prozentualen Anteil an Schülerinnen und Schüler haben, die die höchste Kompetenzstufe erreichen, als in der NAEP-Stichprobe selbst. Übergreifend schlussfolgern sie jedoch, die NAEP Kompetenzstufen seien vom Niveau her nicht zu hoch angesetzt. Hinsichtlich der Limitationen dieser Untersuchung verweisen Hambleton et al. darauf, dass die Studien beispielsweise unterschiedliche Exklusionsregeln anwenden, die Erhebungen nicht zur gleichen Zeit durchgeführt wurden und die Schülerinnen und Schüler in der PISA-Studie etwas älter sind als in den anderen beiden Studien. Und obwohl andere Studien festgestellt haben, dass sich die drei Studien NAEP, TIMSS und PISA hinsichtlich ihrer erfassten Inhalte und der Aufgabenformate ähneln, kein Equating, sondern vielmehr Concordance vorliegt, d. h. die Ergebnisse erlauben zwar kein präzises Ranking der Länder, dies war jedoch auch nicht das Ziel dieser Untersuchung. Hartig und Frey (2012) untersuchen die Validität des Ländervergleichs in Deutschland (Normierungsstichprobe in 2006), indem sie diesen PISA 2006 gegenüberstellen. Der Ländervergleich überprüft die deutschlandweit gültigen Bildungsstandards, um festzustellen, inwieweit die Bildungsstandards erreicht werden und ob bzw. wo noch weiterer Steuerungsbedarf besteht. Diese Schulleistungsstudie erlaubt dabei einen Vergleich auf Ebene der Bundesländer in Deutschland. Die Studie von Hartig und Frey bezieht sich auf die mathematischen Kompetenztests aus dem Ländervergleich (9. Klassenstufe) und von PISA 2006. Zunächst stellen sie auf einer inhaltlichen Ebene die Rahmenkonzeptionen, den Aufgabenentwicklungsprozess sowie die Bezüge zu schulischen Inhalten in den beiden Studien einander gegenüber. Hartig und Frey stellen hohe Übereinstimmungen zwischen den Studien fest, wobei sie jedoch anmerken, dass die Bildungsstandards durch die nationale Orientierung sich mehr an den Schulinhalten orientieren als die Rahmenkonzeption von PISA. Sie führen zudem unterschiedliche statistische Berechnungen durch.

Tabelle 1.2: Überblick über aktuelle Äquivalenz- und Linkingstudien

		Vergleich von Studien-merkmalen		Vergleich der Konstrukte			
Autoren	Analysierte Schulleistungsstudien	Schluss-folgerungen	Population	Mess-bedingungen	Konzeption, Test-framework	Dimensionale Äquivalenz	Skalen-bezogene Äquivalenz
Blum et al., 2004	PISA 2003 national und international					x	
Dorans, 2004	ACT, SAT			x	x	x	x
Grønmo & Olsen, 2007	TIMSS 2003, PISA 2003	x	x	x	x		
Hambleton et al., 2009	NAEP 2003, TIMSS 2003 und PISA 2003				x		x
Hartig & Frey, 2012	LV (Normierungsstichprobe) und PISA 2006			x	x	x	
National Center for Education Statistics, 2013	NAEP 2011 und TIMSS 2011	x	x	x	x		x
Neidorf et al., 2006	NAEP 2003, TIMSS 2003 und PISA 2003	x	x	x	x		
Nohara & Goldstein, 2001	NAEP 2000, TIMSS-R 1999 und PISA 2000	x	x		x		
Wu, 2010	PISA 2003 und TIMSS 2003	x	x	x	x	x	(x)

Sie stellen beispielsweise latente korrelative Zusammenhänge zwischen den mathematischen Kompetenztests der beiden Studien fest (r = .94), wobei sie zusätzlich die Schulformunterschiede mit berücksichtigen (r = .88 bis .92).Die Überprüfung, ob ein vierdimensionales Modell (Ländervergleich Mathematik, PISA Mathematik, PISA Lesen und PISA Naturwissenschaften) oder ein dreidimensionales Modell, bei dem die Items des Ländervergleichs und PISA zusammen auf einer Dimension laden, ergab, dass das dreidimensionale Modell eine bessere Passung aufweist. Hartig und Frey argumentieren in ihrem Artikel, dass Korrelationen alleine jedoch ein unvollständiges Bild liefern. Daher analysieren sie darüber hinaus die Varianz zwischen Schulen und zwischen Schulformen unter Verwendung von Mehrebenen-Raschmodellen. Die Ergebnisse fallen erwartungsgemäß aus. Die Varianzaufklärung ist in beiden Fällen im Mathematiktest der Bildungsstandards höher als in dem PISA-Test. Hartig und Frey argumentieren, dass aufgrund der höheren Varianz zwischen Schulen für den Ländervergleich-Test als für den PISA-Test die Annahme gestützt wird, dass die curriculare Validität des Ländervergleichs-Tests höher ist. Weiterhin verweisen sie auf die Relevanz, in kommenden Studien vermehrt die Effekte von Hintergrundmerkmalen und die Zusammenhänge der in den Tests erbrachten Leistung mit den Schulnoten zu analysieren, um weitere Hinweise darüber zu erhalten, wie nah die Tests an den Curricula orientiert sind.

Das Institute of Education Sciences (IES) hat im Jahr 2011 eine Linkingstudie durchgeführt, welche die nationale Schulleistungsstudie NAEP, durchgeführt in den Vereinigten Staaten von Amerika, mit der internationalen Schulleistungsstudie TIMSS (National Center for Education Statistics, 2013) zusammenführt. Hierfür nutzen sie Kompetenzdaten der Achtklässlerinnen und Achtklässler in Mathematik und in den Naturwissenschaften. Ein Ziel der NAEP-TIMSS-Linkingstudie ist, die NAEP-Skala mit der TIMSS-Skala zu verknüpfen, um die Fähigkeiten der Schülerinnen und Schüler in der nationalen Studie mit denen von Schülerinnen und Schüler anderer Länder vergleichen zu können. Sie stellen fest, dass der Mittelwert in der Mathematikleistung für öffentliche Schulen in 36 Bundesländern höher ist, als der TIMSS Mittelwert von 500. Der Mittelwert in der Leistung in den Naturwissenschaften ist in 47 Bundesstaaten höher als der TIMSS Mittelwert. Darüber hinaus wurden die Bundesstaaten den TIMSS-Kompetenzstufen zugeordnet. Dabei zeigte sich, dass die Schülerinnen und Schüler von 51 Bundesstaaten in Mathematik auf bzw. über dem Benchmark von 475 Kompetenzpunkten liegen. Dies entspricht einer Kompetenzstufe von 3 und höher. In den

Naturwissenschaften erreichen ebenfalls die Schülerinnen und Schüler von 51 Bundesstaten die Kompetenzstufe drei oder höher. Hinsichtlich der genannten Ergebnisse verweisen die Autoren jedoch darauf, dass u. a. Unterschiede in der Administration, den definierten Konstrukten, den Testzeiten und den erlaubten Hilfsmitteln dazu führen können, dass die vorhergesagten TIMSS-Ergebnisse fehlerbehaftet sind. Daher seien die vorhergesagten TIMSS-Ergebnisse nicht als tatsächliche TIMSS-Ergebnisse zu interpretieren. Ein weiteres Ziel der Untersuchung ist, drei unterschiedliche Linking-Methoden (Calibration, Statistical Projection, Statistical Moderation) zu evaluieren, welche sich für die Verknüpfung von Studien eignen, die zu einem gewissen Grad etwas unterschiedliches Messen. Die Akkuratheit der vorhergesagten TIMSS-Werte evaluieren sie dabei anhand von neun Bundesstaaten, die an beiden Untersuchungen teilgenommen haben. Hinsichtlich des Vergleichs der drei Methoden stellen die Autoren fest, dass die drei angewendeten Linking-Methoden in dem vorliegendem Fall im Wesentlichen zu gleichen Ergebnissen führen. Sie entscheiden sich, die Ergebnisse auf Grundlage der Statistical Moderation Methode zu berichten, da es sich hierbei um eine einfache Methode mit wenigen Parametern handelt, die zudem keine zusätzlich zusammengestellten Testheftvarianten voraussetzt. Dies spare zukünftig die Erstellung und Durchführung dieser zusätzlichen Testheftvariation, wie bei den anderen beiden Linking-Methoden.

Neidorf et al. (2006) haben ebenfalls eine Untersuchung zum Vergleich nationaler und internationaler Studien durchgeführt. Sie haben die Frameworks und die Items vom NAEP, TIMSS und PISA 2003 verglichen, um herauszufinden, inwiefern ein Vergleich und eine gemeinsame Interpretation der Ergebnisse zulässig sind. Hierzu haben sie die NAEP und TIMSS-Mathematiktests für Viert- und Achtklässler sowie den Mathematiktest von PISA für 15-Jährige herangezogen. Neidorf et al. haben in ihrer Studie auf einer inhaltlichen Ebene die Rahmenkonzeptionen hinsichtlich des Aufbaus und der Definitionen der inhaltlichen und prozessorientierten Fähigkeiten sowie anderer Aspekte (Aufgabenformate, kognitive Prozesse und Itemkontexte) nebeneinander gestellt. Die Vergleiche der Aufgaben basieren auf den Einordnungen der NAEP und TIMSS-Aufgaben in die jeweils gegenseitigen Rahmenkonzeptionen, sowie der Einordnung der PISA Aufgaben in die NAEP Rahmenkonzeption. Weiterhin wurden die Aufgaben von PISA und TIMSS hinsichtlich ihrer mathematischen Komplexität in die in NAEP definierten Komplexitätsstufen (niedrige, mittlere und hohe Komplexität) klassifiziert. Neidorf et al. fassen übergreifend zusammen,

dass sich die drei Studien in vielen Aspekten sehr ähnlich sind, jedoch könne nicht angenommen werden, dass sie dieselben Inhalte tatsächlich auf die gleiche Weise erfassen. Der Vergleich von NAEP und TIMSS ergab u. a., dass auf der Oberfläche viele Ähnlichkeiten bestehen und eine gute Zuordnung zu den gegenseitigen Rahmenkonzeptionen möglich sei, ein detaillierterer Vergleich jedoch viele Unterschiede offenbare. Werden die Unterkategorien der Inhaltsbereiche mit berücksichtigt, könnten 20 Prozent der Aufgaben für die vierten Jahrgangsstufe und etwa 15 Prozent der Aufgaben für die achte Jahrgangsstufe nicht in den gegenseitigen Rahmenkonzeptionen verortet werden. Neidorf et al. schließen daraus, dass diese Aufgaben wahrscheinlich nicht in dem anderen Test vorkommen würden und daher die beiden Studien nicht den gleichen mathematischen Inhalt erfassen. PISA verhält sich komplementär zu NAEP und TIMSS. Im Gegensatz zu den beiden anderen Studien hat PISA zum Ziel, die Problemlösefähigkeiten der Schülerinnen und Schüler in realen Situationen zu erfassen. Zudem ist die Stichprobe in PISA alters- und nicht klassenbasiert. Dennoch lassen sich die PISA Aufgaben gut in die Rahmenkonzeption von NAEP einordnen, sogar auf Ebene der Unterkategorien der Inhaltsbereiche (mehr als 90 Prozent). Neidorf et al. konkludieren aus den Ergebnissen, dass genauere Analysen der Items weitere wichtige Unterschiede aufklären könnten. Jedoch zeigen die Analysen den komplementären Charakter der drei Studien. Vor allem bezüglich der Fähigkeiten in den Inhaltsbereichen inklusive der Unterkategorien könnten die jeweiligen Ergebnisse der Studien je nach Fragestellung mal mehr und mal weniger informativ sein.

Nohara (2001) stellen die Inhalte der Studien NAEP 2000, TIMSS-R 1999 und PISA 2000 gegenüber. Ziel ist, die Gemeinsamkeiten und die Unterschiede der drei Studien darzustellen, um aufzuzeigen, was jeweils genau gemessen wird und was dies für die Kompetenzen der U. S. Schülerinnen und Schüler in den jeweiligen Studien bedeutet, um die Unterschiede in den Ergebnissen in den drei Studien erklären zu können. Nohara spezialisiert sich hierbei auf die Mathematik- und Naturwissenschaftstests für die achte Jahrgangsstufe (NAEP und TIMSS-R) bzw. die 15-jährigen Schülerinnen und Schüler (PISA). Für den Vergleich von NAEP mit TIMSS-R und PISA hat Nohara die Testaufgaben hinsichtlich mehrerer Kriterien durch Experten klassifizieren lassen. Zunächst wurden alle Aufgaben der Studien den in der NAEP-Rahmenkonzeption definierten Inhaltsbereichen zugeordnet. Die Ergebnisse der Aufgabenklassifikation der Mathematikaufgaben zeigen, dass die meisten Aufgaben von NAEP und TIMSS-R dem Inhaltsbereich ‚Number sense, properties, and operations' zugeordnet

werden konnten (NAEP: 32 %; TIMSS-R: 46%) wohingegen dieser Inhaltsbereich bei PISA den geringsten Teil an Aufgaben ausmacht (9%). Bei PISA wurden die meisten Aufgaben (31%) dem Inhaltsbereich ‚data analysis, statistics, and probability' zugeordnet, bei TIMSS-R wurden indessen nur 11% und bei NAEP nur 14% der Aufgaben diesem Inhaltsbereich zugeordnet. In einem zweiten Schritt wurden die Items hinsichtlich ihres Antwortformats klassifiziert. Hierbei zeigte sich, dass das Antwortformat Multiple-Choice am häufigsten bei NAEP (60%) und bei TIMSS-R (77%) Verwendung fand. In PISA wurden nur 34% der Aufgaben als Multiple-Choice Items klassifiziert. Das am häufigsten verwendet Antwortformat bei PISA sind Kurzantworten (50%). Weiterhin wurden die Aufgaben bezüglich des Kontextes analysiert. Unterschieden wurden Aufgaben, die rein mathematisch sind und Aufgaben, die in einen Kontext eingebunden sind, der relevant für das Leben außerhalb von Schule ist. Bei PISA wurde nur eine Aufgabe als nicht kontextbezogen eingestuft wohingegen bei NAEP nur 48% und bei TIMSS nur 44% der Aufgaben als kontextbezogen klassifiziert wurden. Zusätzlich wurden die Items klassifiziert hinsichtlich ihrer Komplexität der benötigten Berechnungen und ob für die Berechnung mehrere Schritte durchdacht werden mussten. Nohara (2001) fand heraus, dass TIMSS-R mit 34% die meisten Aufgaben hat, die eine höhere Komplexität der Berechnung erfordern (NAEP: 27%; PISA: 25%). TIMSS-R hat indessen mit 31 % der Aufgaben weniger Items, die mehrere Schritte zur Lösung erfordern als NAEP und PISA (NAEP: 41%; PISA: 44%). Zudem werden in PISA in fast allen Aufgaben Repräsentationsformen verwendet (91%) und in NAEP und TIMSS-R sind es mit 56% und 45% deutlich weniger solcher Aufgaben. Nohara (2001) konkludiert aus den Ergebnissen, dass PISA die schwerste der drei Studien ist. Dies leitet er daraus ab, dass 59 % der PISA-Aufgaben zwei oder mehr der genannten Kriterien erfüllen. Bei NAEP und TIMSS-R sind es nur 39 % und 24 % der Aufgaben.

Wu (2010) vergleicht die Rahmenkonzeptionen, Aufgabenmerkmale und die Erhebungsmethoden der mathematischen Kompetenztests, die im Rahmen von PISA 2003 und TIMSS 2003 eingesetzt wurden. Das übergeordnete Ziel der Vergleichsstudie ist, mit Hilfe von Regressionsanalysen Unterschiede in den Ergebnissen aufzuklären. Im Detail untersucht Wu (2010) hierzu vier Aspekte: (1) die Übereinstimmungen zwischen den beiden Studien, um die zugrundeliegenden Botschaften zu bekräftigen und Aussagen über die Validität der Ergebnisse treffen zu können, (2) die Unterschiede und/oder Widersprüche zwischen den beiden Studien sowie mögliche Erklärungen, (3) Aspekte, die nur in einer Studie untersucht werden und (4) die Lehren, die aus den Ergebnissen eines Vergleichs gezogen werden können,

um z. B. Aussagen darüber treffen zu können, wie die jeweiligen Studien verbessert werden können. Hierzu stellt Wu zunächst auf einer inhaltlichen Ebene die Methoden, die Rahmenkonzeptionen und die Testspezifikationen gegenüber. In einem zweiten Schritt vergleicht sie die Ergebnisse von PISA und TIMSS und identifiziert Faktoren, die die unterschiedlichen Ergebnisse der Länder in PISA und TIMSS erklären können. Weiterhin vergleicht Wu die Unterschiede auf der Ebene von Subgruppen (Geschlecht, Sozioökonomischer Hintergrund und Einstellungen zur Mathematik, z.B. Selbstkonzept). Wu fand u. a. heraus, dass vor allem Unterschiede in der prozentualen Verteilung der Aufgaben auf die Inhaltsbereiche, in der Zeitspanne der Beschulung sowie in der erforderlichen Leseleistung für das Lösen der Mathematikaufgaben die Ergebnisse auf Länderebene beeinflussen können. Für die Analysen des Einflusses der Verteilung auf die Inhaltsbereiche wurden die PISA-Items re-klassifiziert und den Inhaltsbereichen von TIMSS zugeordnet. Die Unterschiede in der prozentualen Verteilung machen 66 Prozent der Streuung der Unterschiede der Ländermittelwerte zwischen PISA und TIMSS aus. Laut Wu suggerieren diese Ergebnisse, dass der Umfang der mathematischen Curricula einen signifikanten Einfluss auf die Fähigkeiten der Schülerinnen und Schüler hat. Ein wesentlicher Unterschied zwischen den beiden Studien besteht darin, dass die Schülerinnen und Schüler der Länder, die eine hohe Kompetenz in PISA erreichen, hohe Fähigkeiten in der alltäglichen Mathematik haben, wohingegen diejenigen, die in TIMSS eine hohe Kompetenz erreichen, hohe Fähigkeit in der Schulmathematik aufweisen. Das zweite Ergebnis der Studie war, dass die Lesekompetenz einen guten Prädiktor für die Unterschiede zwischen TIMSS und PISA liefert. Da die PISA Aufgaben mehr Leseleistung erfordern als die TIMSS-Aufgaben schlussfolgert Wu u. a., dass anzunehmen ist, dass die Schülerinnen und Schüler, die eine niedrige Kompetenz im Lesen aufweisen, in PISA und TIMSS nicht gleich abschneiden würden. Bezüglich der Ergebnisse zu dem Einfluss der Anzahl der Schuljahre auf die Ergebnisse merkt Wu an, dass die Anzahl der Schuljahre bei PISA berücksichtigt werden sollten, weil ein weiteres Schuljahr den Mittelwert der Fähigkeit um 20 bis 40 Punkte erhöhen kann. Abschließend verweist Wu auf die Relevanz einer sorgfältigen und vorsichtigen Interpretation der Ergebnisse, da es viele länderspezifische Faktoren gibt, die die Fähigkeiten der Schülerinnen und Schüler beeinflussen können. Weiterhin sollte beim Vergleich der Ergebnisse von TIMSS und PISA berücksichtigt werden, dass TIMSS curriculumbasiert ist, wohingegen PISA sich auf die Fähigkeit des Lösens von alltagsspezifischen Problemen konzentriert.

1.1.3 Zwischenfazit

Der vorangegangene Überblick zeigt, dass es zwar verschiedene Herangehensweisen gibt, Linking-Varianten zu unterscheiden, diese jedoch große Überschneidungen aufweisen. Gemeinsam ist den Definitionen, dass der Begriff des Equating nur für das Verlinken von Paralleltests genutzt wird. Die Differenzierung von Linking-Typen bei nicht-parallelen Tests erfolgt unterschiedlich, allerdings hängt die Differenzierung immer von der Ähnlichkeit der Tests z. B. hinsichtlich der Testspezifikationen und der Rahmenkonzeptionen ab. Hier wurde zunächst nur unterschieden, ob die Tests dieselben Testspezifikationen oder dieselbe Rahmenkonzeption haben. Außer Acht gelassen wurden nach dieser Unterscheidung jedoch solche Tests, die zwar unterschiedliche Testspezifikationen und/oder Rahmenkonzeptionen haben, welche sich aber in hohem Maße ähneln. Daher wurden von mehreren Autoren Herangehensweisen aufgezeigt, mit denen der Grad an Übereinstimmung zwischen den Tests hinsichtlich verschiedener Kriterien analysiert werden kann, um Aussagen darüber treffen zu können, wie ähnlich bzw. wie unähnlich sich die Tests tatsächlich sind, beispielsweise hinsichtlich ihrer definierten Konstrukte. Hier werden jedoch meist nur Oberbegriffe genannt bzw. vereinzelt Beispiele gegeben, die viel Spielraum lassen, welche Faktoren gegenübergestellt werden können bzw. sollten. Im Folgenden wird daher ein zusammenfassender Überblick darüber gegeben, welche Aspekte in unterschiedlichen Beiträgen genannt werden, sowohl in den theoretischen Texten als auch in den vorgestellten Studien (u. a. Kolen & Brennan, 2010; Feuer et al., 1999; American Educational Research Association et al., 1999; Kolen, 2004; Holland & Dorans, 2006; Pommerich, 2007; Hambleton et al., 2009; Kolen, 2004; Lindquist, 1964; Marco et al., 1992; Dorans, 2004).

Die zu vergleichenden Aspekte lassen sich grundlegend zwei Ebenen zuordnen:

(1) Der Vergleich auf Grundlage der Hauptuntersuchungen: Hier werden die Daten genutzt, die von den jeweiligen Studien selbst zur Verfügung gestellt werden. Mit Hilfe dieser Daten können z. B. ein Vergleich der Rahmenkonzeptionen der Tests und ein Vergleich der Aufgabenmerkmale durch ein Expertenreview erfolgen. Für diese Analysen sind keine quantitativen Daten aus einer zusätzlich durchgeführten Linking-Studien notwendig (Dorans, 2004).

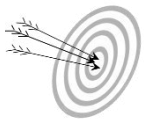
(2) Der Vergleich auf Grundlage einer Linkingstudie: Für die Analyse der dimensionalen und skalenbezogenen Zusammenhänge sowie der Evaluation der Ergebnisse des Linking hingegen werden Daten aus einer Linking-Studie benötigt, d. h. die zu vergleichenden Tests müssen entweder gleiche Items beinhalten oder von einer Stichprobe bearbeitet worden sein (Kapitel 1.2).

Eine Untersuchung der Äquivalenz lässt sich nach Kolen und Brennan (2010) weiter unterteilen in den Vergleich (1) der Schlussfolgerungen, die aus den Tests gezogen werden können, (2) der Stichprobe, für die die Tests konstruiert sind, (3) der Messbedingungen sowie (4) den Konstrukten. Wie bereits beschrieben, lässt sich der letzte Aspekt nach van de Vijver (1998) weiter unterteilen in einen konzeptionellen, dimensionalen und skalenbezogenen Vergleich. Zusätzlich wird im Rahmen dieser Arbeit noch ein weiterer Vergleich hinzugefügt: der methodische Vergleich. Hierzu zählen die in den Studien verwendeten Auswertungsmethoden.

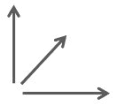
Die Tabelle 1.3 gibt im oberen Bereich einen zusammenfassenden Überblick darüber, welche unterschiedlichen Aspekte bei einem Vergleich von zwei oder mehr Tests analysiert werden können (nach Kolen & Brennan, 2010; Feuer et al., 1999; American Educational Research Association et al., 1999; Kolen, 2004; Holland & Dorans, 2006; Pommerich, 2007; Hambleton et al., 2009; Lindquist, 1964; Marco et al., 1992; Dorans, 2004) eingeordnet in die Bereiche nach Kolen und Brennan (2010) und van de Vijver (1998). Empfohlen und durchgeführt werden zudem von mehreren Autoren und Autorinnen ein detaillierter Vergleich der jeweiligen Aufgaben sowie eine re-klassifikation der Aufgaben in die jeweils anderen Rahmenkonzeptionen. Holland und Dorans (2006) merken ergänzend an, dass eine Experteneinschätzung zu der Gleichheit der Testinhalte nützlich sein könnte. Des Weiteren enthält die Tabelle 1.3 eine Aufführung von Faktoren, die von den oben genannten Autoren und Autorinnen – neben den bereits genannten Aspekten – der Evaluation eines Linking dienen.

Zusammenfassend lässt sich festhalten, dass nur, wenn ein Equating vorliegt, die Ergebnisse der Studien als austauschbar verwendet werden können und damit auch eine Interpretation auf Individualebene erfolgen kann. Liegt eine schwächere Form des Linking vor, kann nicht davon ausgegangen werden, dass sich die Linking Ergebnisse eins zu eins übertragen lassen. Daher sollte auf Interpretation auf Individualebene verzichtet werden. Sind

Tabelle 1.3: Aspekte, die für die Bestimmung der Güte eines Linkings analysiert werden können sowie Kriterien zur Evaluation der Exaktheit und Stabilität des Linking

	Aspekte der Inhaltlichen Gegenüberstellung	
Ebene 1: Auf Grundlage der Hauptuntersuchungen	Anlage und Ziele 	Messintention Mögliche Schlussfolgerungen Referenzrahmen National vs. international Untersuchungsdesign Kompetenzbereiche Grundlegende Konzeption
	Stichprobe 	Population versus Stichprobe Alters- oder Jahrgangsbasiert Stichprobenziehung Größe der Stichprobe Ausschlusskriterien
	Messbedingungen 	Messzeitpunkt Testzeit Testdesign Erlaubte Hilfsmittel
	Konstrukte: konzeptionell 	Theoretische Konzeption der Konstrukte Inhaltsbereiche Kognitive Anforderungsbereiche Prozedurale Fähigkeiten Aufgabenmerkmale (formal und sprachlich)
	Konstrukte: methodisch 	Skalierungsmodelle (z. B. 1-PL- vs. 3-PL-Modell)

(Fortsetzung der Tabelle auf der folgenden Seite)

Ebene 2: Auf Grundlage einer Linkingstudie	**Überprüfung der dimensionalen und skalenbezogenen Zusammenhänge**	
	Konstrukte: dimensional 	Faktorielle Struktur der Tests
	Konstrukte: skalenbezogen 	Reliabilität Korrelative Zusammenhänge zwischen den Tests
	Evaluation der statistischen Exaktheit und Stabilität	
	deskriptive Kennwerte (Mittelwerte, Standardabweichung, Schiefe, Kurtosis) Klassifikationskorrektheit zu den Kompetenzstufen Stabilität über Subgruppen	

die Linkingergebnisse reliabel, können jedoch auf Gruppenebene Interpretationen vorgenommen werden. Wichtig ist, dass Unterschiede zwischen den Studien hinsichtlich der genannten Aspekte zu Fehlern in den Ergebnissen der Verlinkung der Tests führen können, und daher ausführlich untersucht werden sollten.

1.2 Äquivalenz- und Linkingstudien verknüpfen: Stichprobendesigns

Unabhängig von der vorliegenden Linking-Variante sollte beim Linking-Prozess darauf geachtet werden, dass die unterschiedlichen Fähigkeiten der teilnehmenden Schülerinnen und Schüler Berücksichtigung finden, denn die Ergebnisse werden immer von zwei Faktoren beeinflusst (Holland, 2007): (1) von der jeweiligen Schwierigkeit der beiden Tests und (2) von der jeweiligen Fähigkeit der beiden Gruppen in den Tests. Holland (2007) unterscheidet zwei Wege, hiermit umzugehen: (A) das Nutzen einer gleichen Stichprobe für beide Testvarianten, um die Unterschiede in der Fähigkeit der Teilnehmer und Teilnehmerinnen zu kontrollieren oder (B) die Verwendung gleicher Items (sogenannte Anker-Items), um die Unterschiede in den Testschwierigkeiten zu kontrollieren.

Je nachdem welcher Weg gewählt wird, sind unterschiedliche Formen von Stichprobendesigns möglich, die für einen Vergleich gewählt werden können. Wichtig hierbei zu beachten ist, dass nicht mit allen Designs auch alle Linking-Varianten und Linking-

Methoden verwendet werden können. Sollen zwei unterschiedliche Testformen X und Y miteinander verglichen werden, gibt es unterschiedliche Stichprobendesigns (vgl. Abbildung 1.2), die sich hierfür anbieten. In der Literatur werden zumeist drei bzw. vier Stichprobendesigns unterschieden, die bereits in vielen Studien Verwendung fanden (von Davier, Carstensen & von Davier, 2008; Holland & Dorans, 2006; Kolen & Brennan, 2010; Muraki, Hobo & Lee, 2000; Ryan & Brockmann, 2009): (1) ‚equivalent group design' oder auch ‚random group design' genannt (EG), (2) ‚single group design' (SG), (3) ‚counterbalanced design' (CB) und (4) ‚non-equivalent groups with anchor test design' oder auch ‚common-item nonequivalent groups design' genannt (NEAT).

Diese vier Stichprobendesigns lassen sich zwei Gruppen zuordnen. Bei der ersten Gruppe gehören alle Testteilnehmerinnen und Testteilnehmer zu einer Stichprobe (EG, SG und CB). Demgegenüber gibt es bei der zweiten Gruppe zwei unterschiedliche Stichproben, die über gleiche Items miteinander in Beziehung gesetzt werden sollen (NEAT). Diese beiden Stichproben müssen jedoch nicht vergleichbar sein. Dafür gibt es im NEAT-Design gleiche Items

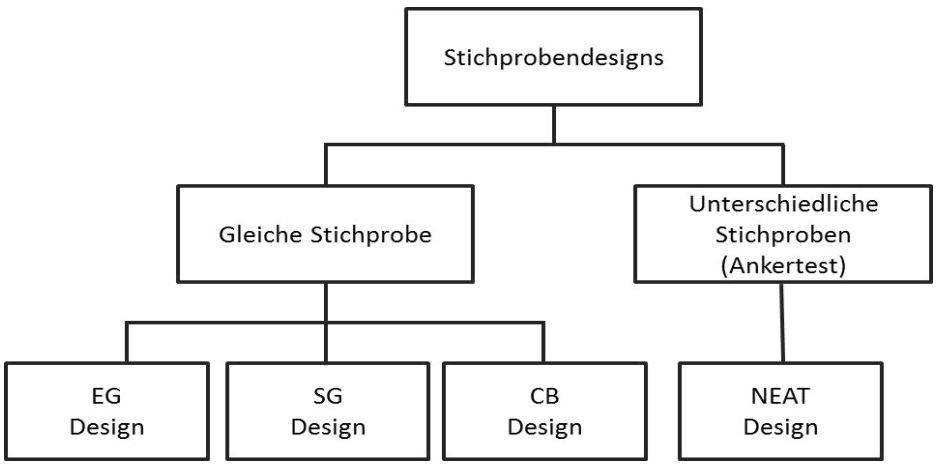


Abbildung 1.2: Stichprobendesigns

(Ankeritems) – in den beiden Testteilen. Im Folgenden werden die vier oben genannten Stichprobendesigns kurz vorgestellt. Außer Acht gelassen wird hierbei, dass sich die vier Stichprobendesigns zudem auch miteinander kombinieren lassen (z. B. Von Davier et al., 2008. Zudem könnten, wenn beispielsweise eine IRT-Transformation für einen Vergleich herangezogen werden soll, weitere Designs verwendet werden, die ein unvollständiges Datendesign abbilden (z. B. Multi-Matrix-Design).

(1) Im EG-Design (vgl. Tabelle 1.4) gibt es eine Stichprobe (S), die in zwei Substichproben (1 und 2) aufgeteilt wird. Wichtig ist, die Gruppen so aufzuteilen, dass die Substichproben äquivalent sind. Eine Möglichkeit, die vermehrt Anwendung findet ist, die zufällige Aufteilung der Teilnehmerinnen und Teilnehmer in zwei Subgruppen beispielsweise beim Austeilen der Tests. Der erste Teilnehmer bekommt den Testteil X, der zweite Teilnehmer den Testteil Y, der dritte Teilnehmer wieder Testteil X usw. Letztendlich gibt es damit zwei Substichproben, die einer gemeinsamen Stichprobe entstammen. Die Substichprobe 1 bearbeitet nur den Testteil X und die Substichprobe 2 bearbeitet nur den Testteil Y. Die zufällige Zuordnung zu den Gruppen führt dazu, dass die Ergebnisse theoretisch vergleichbar sind (Muraki et al., 2000). Bei dieser Variante ist zu beachten, dass – wenn beide Subgruppen gleichzeitig in einem Raum getestet werden – die Instruktion die gleiche sein muss und dass die Tests die gleiche Länge haben müssen (Holland & Dorans, 2006). Eine weitere Möglichkeit besteht darin, vorher zwei Substichproben zufällig aus der Stichprobe zu ziehen und dann jede Gruppe einzeln zu testen (Holland & Dorans, 2006). Welche Möglichkeit auch gewählt wird, die beiden Substichproben sollten möglichst zur gleichen Zeit getestet werden. Finden die Testungen der beiden Substichproben zu unterschiedlichen Zeitpunkten statt, ist eine Vergleichbarkeit der beiden Gruppen schwieriger zu gewährleisten (Feuer et al., 1999). Für dieses Design ist eine relativ große Stichprobe nötig (von Davier et al., 2008; Ryan & Brockmann, 2009). Das EG-Design wird von den Methoden des linearen und equipercentilen Linking unterstützt sowie von verschiedenen IRT-Ansätzen (Ryan & Brockmann, 2009).

Tabelle 1.4: Equivalent Groups (EG) Design

Stichprobe	Substichprobe	X	Y
S	1	✓	
S	2		✓

(2) Im SG-Design (vgl. Tabelle 1.5) bearbeiten alle Teilnehmer und Teilnehmerinnen beide Testteile X und Y. Dies ist mit einem größeren Zeitaufwand verbunden, dafür ist im SG-Design – im Gegensatz zum EG-Design – nur eine kleinere Stichprobe nötig. Es ist jedoch möglich nach einem ersten Durchlauf mit den kompletten Testteilen den Test zu minimieren, indem einige Items nachträglich hinausgenommen werden. Dann kann anschließend der kürzere Test mit den ursprünglichen Testteilen verlinkt werden. Das SG-Design bietet die

direkteste Methode für Linking bzw. Equating. Interkorrelationen zwischen den Tests können Aussagen darüber erlauben, zu welchem Maße die beiden Tests äquivalente Inhalte erfassen (Feuer et al., 1999). Das SG-Design wird beispielsweise bei der Linking-Variante Projektion vorausgesetzt. Nicht kontrolliert werden bei dem SG-Design jedoch mögliche Positionseffekte, die dadurch entstehen könnten, dass der Testteil Y immer an zweiter Stelle nach Testteil X erfolgt und beispielsweise Ermüdungseffekte vor allem bei jüngeren Teilnehmerinnen und Teilnehmern auftreten könnten (Robitzsch, 2009; Muraki et al., 2000). Daher findet dieses Design in der Praxis selten Verwendung (Ryan & Brockmann, 2009).

Tabelle 1.5: Single Group (SG) Design

Stichprobe	**Substichprobe**	**X**	**Y**
S	1	✓	✓

(3) Die Positionseffekte finden im CB-Design (vgl. Tabelle 1.6) Berücksichtigung. Falls Positionseffekte zu erwarten sind, sollte daher das CB-Design gewählt werden. Hierbei gibt es zwei unterschiedliche Testvarianten für zwei Substichproben. Die erste Substichprobe bearbeitet zunächst den Testteil X und im Anschluss den Testteil Y. Bei der zweiten Substichprobe wird die Reihenfolge verändert.

Nach Holland und Dorans (2006) kann das CB-Design bei den Linking-Formen Concordances, Equating, Predicting und Calibration Verwendung finden. Sowohl das SG-Design wie auch das CB-Design unterstützt die Methoden des linearen, equipercentilen und IRT-Linking (Ryan & Brockmann, 2009).

Tabelle 1.6: Counterbalanced (CB) Design

Stichprobe	**Substichprobe**	**X_1**	**Y_1**	**X_2**	**Y_2**
S	1	✓			✓
S	2		✓	✓	

(4) Im NEAT-Design werden zwei unterschiedliche Stichproben S und T gezogen. Diese Stichproben müssen dabei nicht äquivalent sein. Daher bietet sich dieses Design beispielsweise bei der Linking-Variante Moderation (Mislevy; 1992 und Linn; 1993) an (Kapitel 1.2.1). Die Stichprobe S bearbeitet den Testteil X und die Stichprobe T bearbeitet den Testteil

Y. Zusätzlich bearbeiten beide Stichproben S und T den Testteil A – den sogenannten Ankertest. Der Ankertest bietet die Möglichkeit die Unterschiede zwischen X und Y zu quantifizieren (Holland & Dorans, 2006). Stellt sich heraus, dass die beiden Stichproben äquivalent sind, wird von einem EG-Design mit Anker-Items gesprochen.

Tabelle 1.7: Non Equivalent Groups with Anchor Test (NEAT)

Stichprobe	**Substichprobe**	**X**	**Y**	**A**
S	1	✓		✓
T	2		✓	✓

Es kann unterschieden werden in interne (internal) und externe (external) Ankertests. Beim internen Ankertest Design sind die Ankeritems meistens sowohl Bestandteil des Testteils X als auch Bestandteil des Testteils Y, sie werden also über die beiden Testteile verteilt. Zudem werden die internen Ankeritems in die Berechnung des Gesamtscores der Teilnehmerinnen und Teilnehmer mit eingerechnet. Beim externen Ankertest Design werden die externen Ankeritems nicht im Gesamtscore mitgerechnet und die Ankeritems sind meistens in einem separaten Block angelegt, d. h. sie sind weder in Testteil X noch in Testteil Y vorhanden (Kolen & Brennan, 2010; von Davier et al., 2008; Holland & Dorans, 2006). Bei dieser Variante bekommen die Testteilnehmerinnen und Testteilnehmer extra Zeit, um die Anker-Aufgaben zu bearbeiten. Jedoch kann die Motivation geringer sein, bedingt dadurch, dass die Items nicht in das Gesamtergebnis mit eingerechnet werden. Zudem kann es auch hierbei zu Problemen kommen, die durch die Position des Anker-Blocks entstehen (z. B. Ermüdungseffekte). Zu berücksichtigen ist bei beiden Varianten die Position der Ankeritems bzw. des Anker-Blocks, sie sollte in beiden Testteilen gleich sein (Ryan & Brockmann, 2009). Wichtig ist ebenfalls, dass der Wortlaut exakt der gleiche ist, das Itemformat nicht verändert wird und dass keine anderen Veränderungen vorgenommen werden. Zudem sollten die Ankeritems – sofern möglich – die Inhalte der beiden Testteile X und Y repräsentieren und sie müssen in beiden Testteilen gleich sein – sie sollten eine sogenannte Mini-Version der beiden Testteile abbilden (Muraki et al., 2000; Kolen & Brennan, 2010). Bei dem Erstellen einer Mini-Version sollte u. a. darauf geachtet werden, dass die prozentuale Verteilung auf die Inhaltsbereiche ebenso gleich bleibt wie die Verteilung der Itemschwierigkeiten und der Itemformate (Ryan & Brockmann, 2009). Jedoch bleibt die Frage offen, inwieweit die Ankeritems für beide Gruppen tatsächlich

gleich sind, weil – selbst wenn die Ankeritems in beiden Testteilen an der gleichen Stelle auftreten – die Items um die Ankeritems herum nicht dieselben sind (Ryan & Brockmann, 2009).

Die Testzeit ist insgesamt meist etwas länger als beim EG-Design bzw. CB-Design, aber meist kürzer als beim SG-Design abhängig von der Menge der Ankeritems wobei der Test dadurch aber oftmals weniger reliabel ist (Feuer et al., 1999). Vorteil bei diesem Design ist, dass bei einem Vergleich keine äquivalenten Gruppen vorausgesetzt werden (Ryan & Brockmann, 2009). Das NEAT-Design findet Verwendung beim Equating, Vertical Scaling und Anchor Scaling nach Holland und Dorans (2006).

2 Forschungsfragen

Sowohl national als auch international gibt es eine Reihe von Schulleistungsstudien (z. B. TIMSS, der Ländervergleich Mathematik Primar (Ländervergleich) und NEPS), die am Ende der Grundschulzeit die mathematischen Kompetenzen von Schülerinnen und Schülern erfassen (Bos et al., 2012; Stanat et al., 2012; Blossfeld, Maurice & Schneider, 2011). Die genannten drei Studien messen, neben weiteren Domänen, alle mathematische Kompetenz am Ende der Grundschulzeit bzw. zu Beginn der Sekundarstufe. Unterschiede zwischen den drei Testverfahren bestehen aber beispielsweise in der konzeptionellen Orientierung. So lehnt sich TIMSS an die Curricula der teilnehmenden Nationen an, der Ländervergleich an den Bildungsstandards und das nationale Bildungspanel geht von dem Konzept der mathematischen Grundbildung (Mathematical Literacy) aus. Es bleibt daher offen, inwieweit in den drei Studien eine vergleichbare Messung von mathematischer Kompetenz erreicht wird.

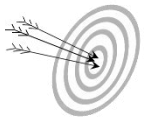
Ein Ziel dieser Arbeit ist daher der Frage nachzugehen, ob die NEPS Mathematikstudie K5 hinreichend ähnlich im Vergleich zu TIMSS und dem Ländervergleich ist, so dass eine Einordnung der Kompetenzmessungsergebnisse in einen nationalen und internationalen Rahmen interpretiert werden könne sowie dass die Kompetenzstufenmodelle der nationalen und internationalen Studien in der NEPS-Studie möglichst exakt und stabil verankert werden können. Im Folgenden werden drei Forschungsfragen formuliert, deren Ergebnisse eine akkurate Einschätzung der übergeordneten Fragestellung ermöglichen sollen. Die Forschungsfragen basieren auf dem zusammengefassten Überblick über mögliche zu untersuchende Kriterien zur Bestimmung der Güte der Übereinstimmung von zwei oder mehreren Tests (Tabelle 1.3). Dabei wird unterschieden zwischen einer inhaltlichen Gegenüberstellung (Forschungsfrage 1), welche alle Analysen umfasst, die sich auf die Rahmenkonzeptionen und die Testspezifikationen der Studien beziehen, und der Analyse dimensionaler und skalenbezogener Zusammenhänge (Forschungsfrage 2), für deren Berechnung eine zusätzliche Linkingstudie durchgeführt wurde (Kapitel 3.1). Die dritte Forschungsfrage umfasst Analysen hinsichtlich der Stabilität und der Exaktheit der Ergebnisse aus dem Verlinken der Studien. Einen Überblick über die verschiedenen Forschungsfragen liefert Abbildung 2.1.

Forschungsfrage 1: Ist eine inhaltliche Vergleichbarkeit der Hauptuntersuchungen (HU) $NEPS_{HU}$ Mathematikstudie K5 (2010) mit den Studien $TIMSS_{HU}$ (2011) und $Ländervergleich_{HU}$ Mathematik Primar (2011) gegeben?

Ist das Ziel zwei oder mehrere unterschiedliche Tests miteinander zu verlinken, die nicht per se auf den gleichen Rahmenkonzeptionen basieren und denen unterschiedliche Testkonzeptionen zu Grunde liegen, sollte das Maß an Übereinstimmung zwischen den Tests bestimmt werden. Damit können erste Aussagen über die Güte des Linking getroffen werden. Grundsätzlich gilt, desto größer die Übereinstimmungen zwischen den Tests sind, desto exakter kann das Linking sein und desto mehr Aussagen können aus den Ergebnissen des Linking gewonnen werden (vgl. hierzu u. a. Kolen und Brennan, 2010; Feuer et al., 1999; Mislevy, 1992; Linn, 1993).

Auf Grundlage der Tabelle 1.3 soll daher im Rahmen der ersten Forschungsfrage untersucht werden, ob der $NEPS_{HU}$ Mathematiktest der fünften Jahrgangsstufe mit den Mathematiktests aus der $TIMSS_{HU}$ Grundschulstudie und dem $Ländervergleich_{HU}$ Primar hinsichtlich der folgenden Merkmale vergleichbar ist:

(1a) Anlage und Ziele



Der erste Aspekt, der nach Kolen & Brennan (2010) hinsichtlich des Maßes an Übereinstimmung untersucht werden sollte, sind die Schlussfolgerungen (Inferences), die aus den Studien gezogen werden können. Dieser Aspekt wurde auf Grundlage weiterführender Literatur, die zusätzlich weitere Aspekte gegenüberstellen (vgl. u. a. Artelt et al., 2008; Cartwright, 2003; Ginsburg, 2005; Grønmo, 2007; NCES, 2013; Johnson, 2003; Neidorf et al., 2006; Nohara et al., 2001, Wu, 2010), erweitert und wird nun unter dem Überbegriff „Anlage und Ziele“ der Studien zusammengefasst. Untersucht werden soll, ob $NEPS_{HU}$ K5 ähnliche oder gleiche

- Messintentionen verfolgt,
- Schlussfolgerungen aus den Ergebnissen zieht,
- Untersuchungsdesigns vorliegen,
- Kompetenzbereiche erfasst und
- Testkonzeptionen zugrunde liegen

wie in der $TIMSS_{HU}$ Grundschulstudie und im Ländervergleich Mathematik Primar.

(1b) Stichprobe

Der zweite Aspekt, der nach Kolen & Brennan (2010) vergleichend gegenübergestellt werden sollte ist die Definition der Stichprobe. Auch dieser Aspekt wurde in vielen Linkingstudien untersucht und beinhaltet meistens mehrere Teilbereiche (vgl. u a. Artelt et al., 2008; Cartwright, 2003; Ginsburg, 2005; Grønmo, 2007; NCES, 2013; Johnson, 2003; Neidorf et al., 2006; Nohara et al., 2001; Wu, 2010). Auf Grundlage der Literatur soll dementsprechend verglichen werden, ob die Stichproben in $NEPS_{HU}$ K5, in der $TIMSS_{HU}$ Grundschuluntersuchung und im Ländervergleich Mathematik Primar

- die Population oder eine Stichprobe abdecken,
- alters- oder jahrgangsbasiert sind,
- unter ähnlichen oder gleichen Bedingungen gezogen wurden,
- ähnlich oder gleich groß sind und
- unter der Bedingung ähnlicher oder gleicher Ausschlusskriterien gezogen wurden.

(1c) Messbedingungen

Unter diesem Oberbegriff werden nach Kolen und Brennan (2010) Vergleiche zusammengefasst, die die Durchführungsbedingungen betreffen. Das Literaturreview hat unterschiedliche Aspekte aufgezeigt, die hinsichtlich der Messbedingungen vergleichend gegenübergestellt werden können (vgl. u. a. Artelt et al., 2008; Cartwright, 2003; NCES, 2013; Johnson, 2003; Nohara et al., 2001; Wu, 2010). Es soll daher untersucht werden, ob

- der Messzeitpunkt,
- die Testzeit,
- das Testdesign und
- die erlaubten Hilfsmittel

in der $NEPS_{HU}$ K5 Studie vergleichbar sind mit denen der $TIMSS_{HU}$ Grundschuluntersuchung und der Untersuchung im $Ländervergleich_{HU}$ Primar.

(1d) Konstrukte: konzeptionell

Alle drei Studien erfassen das Konstrukt mathematischer Kompetenz. Es stellt sich jedoch die Frage, ob die drei Studien tatsächlich das Konstrukt mathematischer Kompetenz auch gleich oder ähnlich definieren (vgl. u. a. Artelt et al, 2008; Cartwright, 2003; Ginsburg, 2005; Grønmo, 2007; NCES, 2013; Neidorf et al., 2006; Nohara et al., 2001; Wu, 2010). Daher soll untersucht werden, ob in dem $NEPS_{HU}$ K5 Mathematiktest die gleichen oder ähnliche

- Inhaltsbereiche,
- kognitiven Anforderungsbereiche sowie
- prozeduralen Fähigkeiten

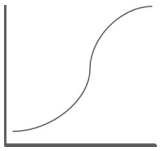
abgedeckt werden wie in den Mathematiktests von $TIMSS_{HU}$ K4 und $Ländervergleich_{HU}$ Primar. Eine detaillierte Analyse der untersuchten Kompetenzbereiche ist wichtig, da es zum einen möglich ist, dass die Kompetenzbereiche in den Studien gleich heißen, aber dennoch unterschiedliches abdecken („jingle fallacy“), zum anderen ist es ebenfalls möglich, dass die Kompetenzbereiche in den Studien unterschiedlich benannt werden, aber gleiches abdecken („jangle fallacy“) (vgl. Marsh, 1994). Die Linkingstudien von Grønmo (2007), Neidorf (2006) und Nohara (2001) haben u. a. aufgezeigt, wie sinnvoll ein detaillierter Vergleich der definierten Konstrukte ist und dass sich erst bei einer genaueren Betrachtung (beispielsweise durch eine Kreuzklassifikation der Aufgaben) Unterschiede aufdecken lassen.

Neben Unterschieden oder Gemeinsamkeiten in Bezug auf die erfassten mathematischen Kompetenzen, können auch die Mathematikaufgaben an sich unterschiedliche Aufgabenmerkmale aufweisen (z. B. Aufgabenformate oder Repräsentationsformen). Dies wiederum kann Einfluss auf die in den Studien erreichten Kompetenzen haben und sollte daher bei einem Vergleich ebenfalls berücksichtigt werden (vgl. u. a. Grønmo, 2007; Neidorf, 2006; Nohara, 2001; Wu, 2010). Grønmo (2007) und Wu (2010) nehmen zudem die benötigte Leseleistung bei der Bearbeitung der Aufgaben in den Blick. Zusammenfassend soll daher der Frage nachgegangen werden, ob sich der $NEPS_{HU}$ K5 Mathematiktest in Bezug auf die

- formalen und
- sprachlichen

Aufgabenmerkmale von den beiden Mathematiktests der $TIMSS_{HU}$ Grundschulstudie und des $Ländervergleich_{HU}$ Primar unterscheidet.

(1e) Konstrukte: methodisch



In viele Schulleistungsstudien findet die Auswertung auf Basis der Item-Response-Theory (IRT) statt. Jedoch gibt es hier mehrere unterschiedliche Ansätze, welches Skalierungsmodell bevorzugt wird bzw. welche statistischen Methoden zum Einsatz kommen. Die unterschiedliche Herangehensweise an die Skalierung der Daten kann jedoch ebenfalls Auswirkungen auf die Ergebnisse der Studien haben und damit auch auf die Vergleichbarkeit.

Beispiele für einen methodischen Vergleich von zwei oder mehreren Studien lassen sich bei Böhme et al. (2014), Hutchison et al. (2006), Johnson (2003), Pietsch et al. (2009) und Wu (2010) finden. Zudem wurden bereits vielfach die Auswirkungen unterschiedlicher Analysemethoden auf einen Datensatz und die Ergebnisse untersucht (Bonsen, Lintorf, Bos, Frey, 2008; Brown, Micklewright, Schnepf & Waldmann, 2005; Pietsch et al., 2009; Robitzsch, 2009; Stone, Weissman, Lane, 2005; Winkelmann & Robitzsch, 2009). Daher soll untersucht werden, ob die

- statistischen Methoden und
- die Skalierungsmodelle

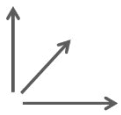
in der $NEPS_{HU}$ K5 Mathematikstudie gleich bzw. ähnlich wie in den Mathematikgrundschulstudien von $TIMSS_{HU}$ und dem $Ländervergleich_{HU}$ sind.

Forschungsfrage 2: Zeigen sich hohe dimensionale und skalenbezogene Zusammenhänge zwischen den in der Linkingstudie (Link) eingesetzten Mathematiktests der $NEPS_{Link}$ K5 Studie und den Studien $TIMSS_{Link}$ K4 und $Ländervergleich_{Link}$ Primar?

Neben der Bestimmung der inhaltlichen Übereinstimmungen stellt sich die Frage, ob sich auch empirische Zusammenhänge zwischen den Tests aufzeigen lassen (Kolen & Brennan, 2010; Dorans & Holland, 2000; Holland & Dorans, 2006; Dorans & Walker, 2007, van de Vijver, 1998), um Aussagen darüber treffen zu können, inwiefern die Tests das gleiche bzw. ein unterschiedliches Konstrukt mathematischer Kompetenz messen. Hinsichtlich der zweiten Forschungsfrage soll daher anhand der Daten einer zusätzlich durchgeführten Linkingstudie

(Kapitel 3.1) untersucht werden, ob eine dimensionale (Forschungsfrage 2a) und eine skalenbezogene (Forschungsfrage 2b) Äquivalenz zwischen dem $NEPS_{Link}$-Mathematiktest K5 und den Mathematiktests von $TIMSS_{Link}$ K4 und dem $Ländervergleich_{Link}$ Primar gegeben ist (Tabelle 1.3).

(2a) Konstrukte: dimensional

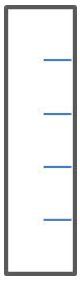


Hinsichtlich der dimensionalen Vergleichbarkeit soll nach van de Vijver (1998) überprüft werden, ob die faktorielle Struktur des $NEPS_{Link}$ Mathematiktests K5 ähnlich ist wie in den Mathematiktests von $TIMSS_{Link}$ K4 und vom $Ländervergleich_{Link}$ Primar.

Die drei Studien unterscheiden verschiedene mathematische Inhaltsbereiche. Eine inhaltliche Gegenüberstellung erfolgte bereits im Rahmen der Untersuchung der Forschungsfrage 1d. Es stellt sich jedoch die Frage, ob sich die inhaltlichen Gemeinsamkeiten und Unterschiede auch empirisch nachweisen lassen. Daher soll zum einen die faktorielle Struktur innerhalb der Tests verglichen werden und zum anderen zwischen den Tests (vgl. u. a. Blum et al., 2004; Winkelmann, Robitzsch, Stanat & Köller, 2012). Diesbezüglich soll – anhand der Ergebnisse einer gemeinsamen Stichprobe (Linkingstudie; die Schülerinnen und Schüler bearbeiten jeweils beide Testinstrumente) – den Fragen nachgegangen werden, ob

- die Korrelationen der Teildimensionen im $NEPS_{Link}$ Mathematiktest K5 vergleichbar ist mit den Korrelationen der Teildimensionen in den Mathematiktests von $TIMSS_{Link}$ K4 und dem $Ländervergleich_{Link}$ Primar?
- hohe Korrelationen zwischen den Teildimensionen der Tests und damit hohe Überschneidungen in den Teildimensionen der Tests aufgezeigt werden können?

(2b) Konstrukte: skalenbezogen



Hinsichtlich der skalenbezogenen Vergleichbarkeit soll in Anlehnung an van de Vijver (1998) überprüft werden, ob – bei gleicher Stichprobe (Daten aus der Linkingstudie) – die Personenfähigkeiten, die mit den Tests geschätzt wurden, zwischen den Studien (nahezu) äquivalent sind und damit eine Linearität zwischen den Schülerinnen und Schülern in den unterschiedlichen Tests besteht.

Dies bedeutet, es soll untersucht werden ob Schüler A in beiden Tests in etwa gleich abschneidet. Dadurch können Rückschlüsse gezogen werden, die sich auf die Vergleichbarkeit der mathematischen Konstrukte in den Tests beziehen. In einem IRT-Kontext und unter Verwendung eines Single Group Designs kann ein Korrelationskoeffizient Hinweise dazu liefern, ob die Tests vergleichbares messen (Kolen & Brennan, 2010; Dorans & Holland, 2000). Zudem kann mithilfe von Modellvergleichstests überprüft werden, ob ein eindimensionales Modell (beide Tests werden auf einer Dimension abgebildet) oder ein zweidimensionales Modell (jeder Test wird auf einer eigenen Dimension abgebildet) besser auf die Daten passt.

Eine weitere Voraussetzung für ein stabiles Equating ist, dass die Reliabilität der beiden Tests adäquat und annähernd äquivalent sein sollte (Kapitel 1.1.1; Dorans & Holland, 2000; Holland & Dorans, 2006; Dorans & Walker, 2007). Auch wenn hier kein Equating, sondern ein Linking angestrebt wird, ist die Höhe und die Ähnlichkeit der Reliabilität entscheidend, weil sie Auswirkungen auf die Güte des Linking, beispielsweise hinsichtlich der Invarianz über Subgruppen, haben kann (Dorans & Holland, 2000).

Daher soll zusätzlich überprüft werden, ob die Reliabilitäten des $NEPS_{HU}$ K5 Mathematiktests adäquat und vergleichbar sind, mit denen der Mathematiktests von $TIMSS_{HU}$ und $Ländervergleich_{HU}$.

Forschungsfrage 3: Zeigt sich in der Linkingstudie hinsichtlich des Linking der $NEPS_{Link}$ Mathematikstudie K5 mit den beiden Studien $TIMSS_{Link}$ Mathematik K4 und $Ländervergleich_{Link}$ Mathematik Primar eine hohe statistische Exaktheit und Stabilität?

Das Verlinken zweier oder mehrerer Tests ist immer eine statistische Schätzung. Statistische Schätzungen können aus unterschiedlichen Gründen messfehlerbehaftet sein. Beispielsweise können Unterschiede zwischen bestimmten Subpopulationen auftreten. Aus diesem Grund sollte ein Linking immer hinsichtlich seiner Exaktheit und Stabilität untersucht werden (vgl. u. a. American Educational Research Association et al., 1999; Holland & Dorans, 2006; Dorans & Holland, 2000; Yin, Brennan & Kolen, 2004; Kolen & Brennan, 2010; Livingston, 2004.

Daher sollen sowohl die Vergleichbarkeit hinsichtlich zentraler Tendenzen – also deskriptiver Messwerte – abgeschätzt werden (z. B. Mittelwerte und Standardabweichungen), als auch für die Verteilung auf die Kompetenzstufen, da diesbezüglich unterschiedliche

Ausmaße an Verzerrungen zu erwarten sind (vgl. u. a. Brown et al., 2005; DeMars, 2001; Pietsch et al., 2009).

Demgemäß soll im Rahmen der dritten Forschungsfrage untersucht werden (vgl. Tabelle 1.3) ob eine hohe Übereinstimmung hinsichtlich zentraler Tendenzen (Gruppenebene) sowie eine hohe Übereinstimmung in der Zuordnung zu den Kompetenzstufen besteht (Individualebene) und damit eine hohe Exaktheit des Linking sowie ob die Ergebnisse des Linking über Subgruppen (hier am Beispiel des Geschlechts) stabil sind und damit Gruppeninvarianz aufweisen.

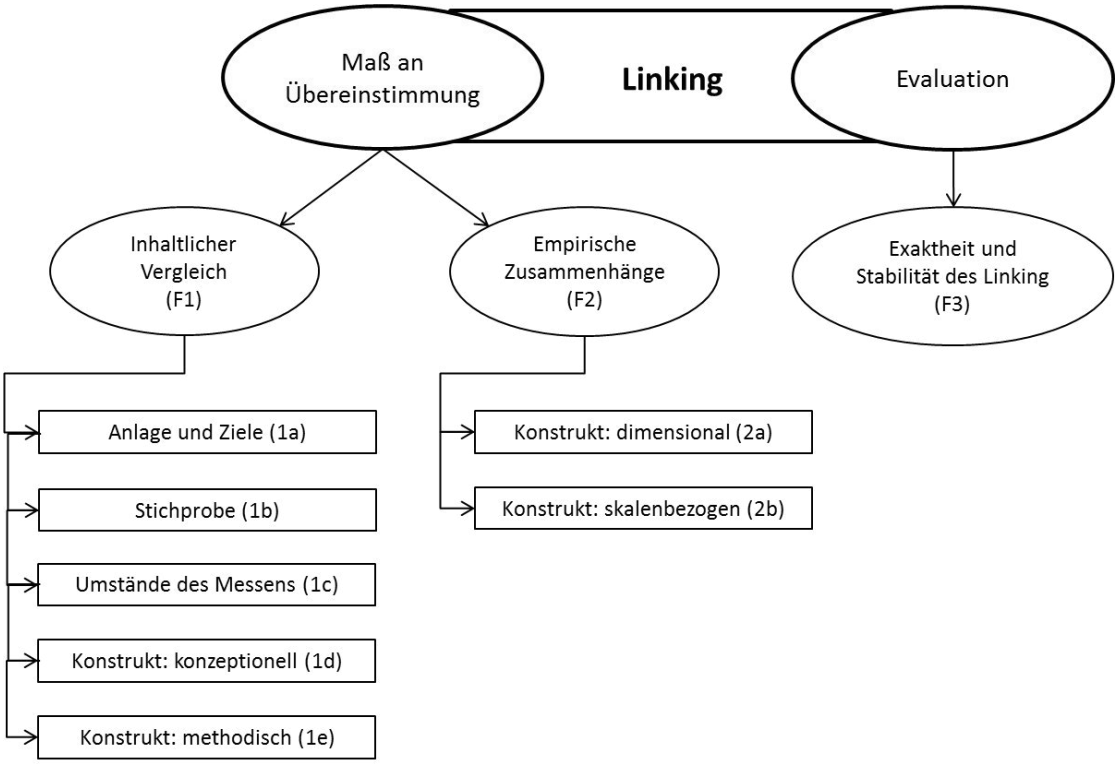


Abbildung 2.1: Übersicht über die Forschungsfragen

3 Methode

Im Folgenden sollen die Methoden zur Beantwortung der im vorherigen Kapitel vorgestellten Forschungsfragen aufgezeigt werden. Für die Beantwortung der ersten Forschungsfrage zur inhaltlichen Vergleichbarkeit der in den drei Studien eingesetzten Mathematiktests wurden die zur Verfügung stehenden Arbeiten aus den jeweiligen Studien, also beispielsweise die Rahmenkonzeptionen der Studien, die Ergebnisberichterstattungen und die jeweiligen Testhefte, herangezogen (vgl. Kapitel 3.6.1). Für die Analysen zu den dimensionalen und skalenbezogenen Zusammenhängen (Forschungsfrage 2) zwischen den Tests und die anschließende Verlinkung (Forschungsfrage 3) wurde eine zusätzliche Studie durchgeführt, da für eine Verlinkung von zwei oder mehreren Tests die Tests entweder gleiche Aufgaben beinhalten müssen oder es muss Schülerinnen und Schüler geben, die beide Tests bearbeitet haben. Da dies bei den Hauptuntersuchungen TIMSS, Ländervergleich und NEPS nicht der Fall ist, wurde für die Untersuchung der Forschungsfragen 2 und 3 in 2011 eine zusätzliche Studie durchgeführt, die sogenannte Linking-Studie 2011. Daher soll zunächst die Linking-Studie näher beschrieben werden, insbesondere hinsichtlich der Zielsetzung der Studie (Kapitel 3.1). Zudem wird in dem vorliegenden Kapitel die Stichprobenziehung in der Linking-Studie 2011 detailliert dargelegt. Dies umfasst sowohl den Prozess bzw. den Ablauf der Stichprobenziehen als auch die Ausschlusskriterien der Studie (Kapitel 3.2). Anschließend wird die Durchführung der Linking-Studie 2011 beschrieben. Diesbezüglich wird der zeitliche Ablauf der beiden Testtage vorgestellt (Kapitel 3.3). In Kapitel 3.4 wird die Teilnahmequote – bezogen auf die jeweils eingesetzten Testinstrumente an den beiden Testtagen – der Linking-Studie dargestellt. Das Testheftdesign der drei eingesetzten Testhefte wird detailliert in Kapitel 3.5 beschrieben. Abschließend werden die wichtigsten Schritte des methodischen Vorgehens bei der Datenauswertung dargestellt und erläutert (Kapitel 3.6).

3.1 Beschreibung der Linking-Studie 2011

Die Linking-Studie 2011 ist ein Kooperationsprojekt des Leibniz-Instituts für die Pädagogik der Naturwissenschaften und Mathematik (IPN) in Kiel und der Leuphana Universität in

Lüneburg[2]. Ziel ist, die Ergebnisse der NEPS-Mathematiktests für die fünfte Jahrgangsstufe in einen nationalen und internationalen Kontext einordnen zu können. Voraussetzung hierfür ist eine hohe Übereinstimmung zwischen den Studien (vgl. Kapitel 1.2.1).

Im Jahr 2011 ergibt sich erstmals die Möglichkeit, den NEPS-Mathematiktest für die fünfte Jahrgangsstufe national und international zu verorten, da sowohl TIMSS (Bos et al., 2012) als auch der Ländervergleich in der Primarstufe (Stanat et al., 2012) in diesem Jahr durchgeführt werden (vgl. Abbildung 4.2) und diese beiden Studien ebenfalls die mathematische Kompetenz von Schülerinnen und Schülern am Übergang von der Primarstufe in die Sekundarstufe messen.

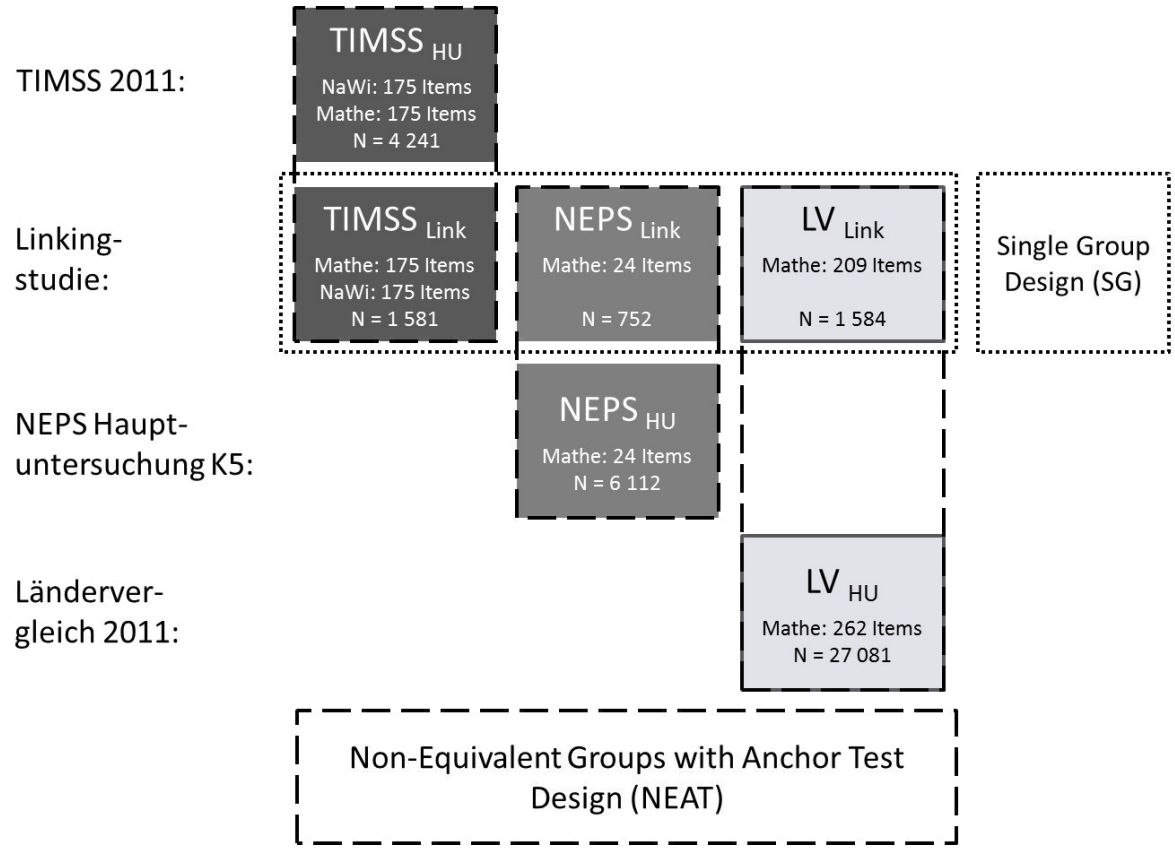


Abbildung 3.1: Datenerhebungsdesign der Linking-Studie

Eine Verlinkung der Studien setzt dabei entweder gleiche Items oder gleiche Stichproben voraus (vgl. Kapitel 1.2.1). Da keine Ankopplung der NEPS-Testinstrumente an die Hauptuntersuchungen von TIMSS und dem Ländervergleich möglich war, wird eine zusätzliche Stichprobe gezogen, ein sogenanntes Oversampling, bei dem die Schülerinnen und Schüler die

[2] Gefördert vom Zentrum für internationale Bildungsvergleichsstudien (ZIB) und dem Bundesministerium für Bildung und Forschung (BMBF)

Testinstrumente bzw. Teile der Testinstrumente der drei Studien bearbeiten (vgl. Abbildung 3.1). Knapp 1 600 Schülerinnen und Schüler bearbeiten die Testhefte von TIMSS und vom Ländervergleich Mathematik Primar und etwa 750 dieser Schülerinnen und Schüler bearbeiten zusätzlich den NEPS-Mathematiktest für die 5. Jahrgangsstufe (SG-Design). Da die Testinstrumente der Hauptuntersuchung bzw. Teile der Testinstrumente eingesetzt wurden, ist eine Verankerung zu den Hauptuntersuchungen von TIMSS und dem Ländervergleich gegeben (NEAT-Design). Hier gibt gleiche Items, jedoch keine gemeinsame Stichprobe.

3.2 Stichprobenziehung

Wie bereits beschrieben wurde eine zusätzliche Stichprobe gezogen. Aufgrund einer Kooperation mit dem IPN und einer am IPN verfolgten weiteren Forschungsfrage (die jedoch nicht Teil dieser Arbeit ist) wurde für die Stichprobenziehung die Teilnahme am Programm SINUS-Grundschule (Demuth, Walther & Prenzel, 2011) vorausgesetzt. Es galt die Grundvoraussetzung, dass die Schulen sowohl an *SINUS-Transfer Grundschule* als auch an *SINUS an Grundschulen* aktiv teilgenommen haben und derzeit noch aktiv am SINUS-Programm mitarbeiten. Ein weiteres stichprobeneinschränkendes Kriterium ist, dass die Klassen im Verlauf ihrer Grundschulzeit mindestens ein Jahr von einer SINUS-Lehrkraft, die sich aktiv am Programm beteiligt, unterrichtet wurden.

Da sowohl TIMSS als auch der Ländervergleich Mathematik Primar die Kompetenzen von Viertklässlerinnen und Viertklässlern erheben (Bos et al., 2012; Stanat et al., 2012), wurden für das Oversampling ebenfalls vierte Klassen ausgewählt. Es wird davon ausgegangen, dass der NEPS-Mathematiktest (Duchhardt & Gerdes, 2012a) – ursprünglich für den Beginn der fünften Klassenstufe konzipiert – ebenfalls die Kompetenzen von Schülerinnen und Schüler am Ende der vierten Klassenstufe messen kann (vgl. hierzu die Analysen in Kapitel 4.1).

Abbildung 3.2: Teilnahmeländer und assoziierte Mitglieder an SINUS an Grundschulen

SINUS an Grundschulen (vgl. u. a. Prenzel; Fischer, Rieck & Dedekind, 2009; Fischer, Rieck, Döring & Köller, 2014) gibt es zurzeit in insgesamt zehn Bundesländern, fünf weitere Bundesländer nehmen als assoziierte Mitglieder teil, d.h. sie können die Entwicklungen direkt mitbeobachten. Thüringen ist zur Zeit der Stichprobenziehung noch ordentliches Mitglied und wird daher ebenfalls berücksichtigt. Dementsprechend nehmen elf Bundesländer an der Untersuchung teil. In den fünfzehn Bundesländern beteiligen sich insgesamt knapp 400 Grundschulen, wobei die Anzahl der teilnehmenden Grundschulen je Bundesland variieren. Von den knapp 400 Grundschulen beteiligen sich nur 113 Schulen aktiv an den beiden Programmen *SINUS-Transfer Grundschule* und *SINUS an Grundschulen*. Weitere 35 Schulen fielen aus der Stichprobe heraus, weil es keine vierte Klasse gab, die von einer SINUS-Lehrkraft unterrichtet wurde oder weil die Schule bereits zur Regelstichprobe in TIMSS oder dem Ländervergleich gehört und die Schülerinnen und Schüler nicht zweimal die gleichen Testhefte ausfüllen sollen. Letztendlich konnten 78 Grundschulen mit 80 vierten Klassen für die Studie gewonnen werden. Dies entspricht einer Gesamtanzahl von N = 1 760 Schülerinnen und Schülern. Des Weiteren wurden 80 Lehrpersonen der getesteten Klassen, 78 Schulleitungen der teilnehmenden Schulen sowie die Eltern der Schülerinnen und Schüler befragt.

Die Tabelle 3.1 zeigt die Verteilung der 78 Schulen innerhalb der Bundesländer. Die Verteilung auf die Bundesländer entspricht in etwa der prozentualen Verteilung in SINUS an Grundschulen. In einer niedersächsischen und einer Thüringer Grundschule wurden jeweils zwei Klassen gezogen.

Tabelle 3.1: Anzahl der teilnehmenden Schulen und Klassen pro Bundesland

Bundesland	**teilnehmende Schulen (N)**	**teilnehmende Klassen (N)**
Baden-Württemberg	1	1
Bayern	17	17
Brandenburg	4	4
Bremen	0	0
Hamburg	11	11
Niedersachsen	9	10
Rheinland Pfalz	11	11
Saarland	3	3
Sachsen-Anhalt	4	4
Schleswig-Holstein	2	2
Thüringen	16	17
Insgesamt	78 Schulen	80 Klassen

3.3 Testablauf

Die Linking-Studie wurde von dem IEA Data Processing and Research Center (DPC) in Hamburg durchgeführt und fand an zwei aufeinander folgenden Tagen unter standardisierten Bedingungen statt. Die Durchführung wurde von schulexternen Testleiterinnen und Testleitern übernommen, die zuvor vom DPC geschult wurden und sich an die vorgegebenen Abfolgen und Instruktionen hielten. Am ersten Testtag (vgl. Tabelle 3.2) bearbeiteten die Schülerinnen und Schüler der insgesamt 80 teilnehmenden Klassen – nach einer zehnminütigen Einweisung in die Bearbeitung der Testhefte – den mathematischen und naturwissenschaftlichen Kompetenztest von TIMSS (siehe Kapitel 3.5.3). Die Testzeit betrug

insgesamt 72 Minuten, nach 36 Minuten gab es eine zehnminütige Pause. Nach dem TIMSS-Kompetenztest gab es eine weitere zehnminütige Pause. Im Anschluss bearbeiteten die Schülerinnen und Schüler einen verbalen kognitiven Fähigkeitstest (KFT 4-12 +R V-Test 3, Form A (verbal); Heller & Perleth, 2000), der sieben Minuten in Anspruch nahm. Die Abkürzung „V“ steht hierbei für den Fähigkeitsbereich *sprachliches Denken*. An den kognitiven Fähigkeitstest anschließend wurden die Kinder in die Bearbeitung des Fragebogens eingewiesen (5 Minuten) und erhielten 30 Minuten Zeit für das Ausfüllen. Dies ergibt für den ersten Testtag eine reine Testzeit von 109 Minuten bzw. mit Pausen 144 Minuten.

Der zweite Testtag war ähnlich aufgebaut (vgl. Tabelle 3.2). Zu Beginn gab es erneut eine Einweisung in die Bearbeitung der Testhefte (10 Minuten). Danach hatten die Schülerinnen und Schüler insgesamt 80 Minuten Zeit, um die Mathematikaufgaben des Ländervergleichs- und Teile des NEPS-Tests zu bearbeiten (vgl. Kapitel 3.5.2 und 3.5.1). Nach 40 Minuten wurde eine Pause von 10 Minuten eingelegt, bevor die Schülerinnen und Schüler weitere 40 Minuten Zeit für die Mathematiktests erhielten. Nach einer zehnminütigen Pause folgte im Anschluss die Durchführung des figuralen kognitiven Fähigkeitstests (KFT 4-12 +R N-Test 3, Form A (figural); Heller & Perleth, 2000). Dabei steht die Abkürzung „N“ für *anschauungsgebundenes Denken*. Hierfür waren 8 Minuten Zeit vorgesehen. Daraufhin wurden die Schülerinnen und Schüler in die Bearbeitung des zweiten Schülerfragebogens eingewiesen (5 Minuten). Für das daran anschließende Ausfüllen der Fragebögen wurden 50 Minuten eingeplant. Dies entspricht einer reinen Testzeit von 138 Minuten und mit Pause einer Gesamtzeit von 173 Minuten.

Um eine hohe Teilnehmerquote zu erreichen, wurden die Schülerinnen und Schüler, die an einem der beiden Testtage gefehlt haben, soweit möglich nachgetestet. Hierfür gab es einen zusätzlichen Termin.

Tabelle 3.2: Testablauf

1. Testtag		**2. Testtag**	
10 Minuten	Einweisung	5 Minuten	Einweisung
36 Minuten	TIMSS Mathematische und Naturwissenschaftliche Kompetenzen	40 Minuten	Ländervergleich & NEPS Mathematische Kompetenzen
10 Minuten	Pause	20 Minuten	Pause
16 Minuten	TIMSS Mathematische und Naturwissenschaftliche Kompetenzen	40 Minuten	Ländervergleich & NEPS Mathematische Kompetenzen
10 Minuten	Pause	10 Minuten	Pause
7 Minuten	KFT Kognitiver Fähigkeitstest (verbal)	8 Minuten	KFT Kognitiver Fähigkeitstest (figural)
5 Minuten	Einweisung	5 Minuten	Einweisung
30 Minuten	Schülerfragebogen Teil I	50 Minuten	Schülerfragebogen Teil II

3.4 Teilnahmequote

Von den insgesamt 1 760 Schülerinnen und Schüler bearbeiteten 1 578 das Testheft am ersten Testtag (vgl. Tabelle 3.3) und damit auch den KFT verbal, da dieser in das erste Testheft integriert war. Von 98 Schülerinnen und Schülern lag keine Elterngenehmigung vor, zwei Schülerinnen und Schüler haben die Schule verlassen und es gab 81 weitere Ausfälle, da die Schülerinnen und Schüler zum Zeitpunkt der Testung nicht anwesend waren. Von einer Teilnehmerin oder einem Teilnehmer wurde im Nachhinein die Genehmigung durch die Eltern widerrufen. Diese Daten wurden gelöscht, so dass dieses Testheft nicht mit in die Analysen eingeht. Im Rahmen eines Nachtests konnten drei weitere Schülerinnen und Schüler das Testheft des ersten Testtages bearbeiten. Damit haben am ersten Testtag insgesamt 1 581 Schülerinnen und Schüler das erste Testheft bearbeitet. Für die Teilnahme am Schülerfragebogen des ersten Testtages lagen weniger Elterngenehmigungen vor und es gab einen weiteren Ausfall, so dass insgesamt 1 493 Schülerinnen und Schüler den Fragebogen ausgefüllt haben.

Am zweiten Testtag bearbeiteten insgesamt 1 575 Schülerinnen und Schüler das zweite Testheft und acht weitere konnten nachgetestet werden, so dass die Teilnahme bei insgesamt 1 584 Schülerinnen und Schülern lag (vgl. Tabelle 3.3). Der KFT wurde von 1 582 Teilnehmerinnen und Teilnehmern bearbeitet und der zweite Schülerfragebogen von 1 493 Schülerinnen und Schülern. Die Stichprobe besteht zu 49.2 % aus Mädchen. Insgesamt haben die Schülerinnen und Schüler ein Durchschnittsalter von 9.7 Jahren (SD = .58) und 10 % sind nicht deutscher Muttersprache.

Tabelle 3.3: Teilnahmequote an der Linking-Studie 2011

		1. Testtag		**2. Testtag**		
Erhebungsinstrument		Testheft	Schüler-fragebogen	Testheft	KFT	Schüler-fragebogen
Gesamt		1760	1760	1760	1760	1760
Haupt-test	Teilnahme	1578	1490	1575	1573	1485
	Ausfall	81	82	96	97	100
	keine Eltern-genehmigung	98	185	86	87	172
	Schule verlassen	2	2	2	2	2
	Widerruf	1	1	1	1	1
Nachtest	Teilnahme	3	3	9	9	8
	Ausfall			1	1	1
	Teilnahme gesamt	1581	1493	1584	1582	1493

3.5 Testheftdesign

Im Folgenden werden die Testheftdesigns der drei Studien vorgestellt, die in der Linking-Studie verwendet wurden. Hierbei soll nicht das Testheftdesign der Hauptuntersuchungen vorgestellt werden (siehe hierfür Kapitel: 4.3), sondern das Design, wie es in der Linking-Studie verwendet wurde. Hierzwischen besteht ein Unterschied, da z. B. nicht alle Testhefte des Ländervergleichs Mathematik Primar verwendet wurden und der NEPS-Mathematiktest in das Design des Ländervergleichs hinein rotiert wurde.

Sowohl der Ländervergleich (Stanat et al., 2012) als auch TIMSS (Bos et al., 2012) verwenden ein Multi-Matrix-Design oder auch Balanced Incomplete Block Design (BIB) genannt (von Davier et al., 2008; Ryan & Brockmann, 2009). Bei diesem Design gibt es unterschiedliche Testformen bzw. Testhefte, die einer Stichprobe vorgelegt werden, so dass jede Schülerin und jeder Schüler zur Bearbeitung nur eine ausgewählte Anzahl an Items erhält. Die Items werden sogenannten Blöcken zugeordnet, die auf mehrere Testhefte verteilt werden. Deutlich wird, dass nicht mehr jedes Testheft alle Items beinhaltet. Jedoch sind die Testhefte untereinander verlinkt, indem die Blöcke mehrmals und vor allem in den unterschiedlichen Testheften an verschiedenen Positionen auftauchen. Die Testhefte werden nun spiralförmig (spiralled) auf die Untersuchungsteilnehmerinnen und -teilnehmer verteilt (vgl. Kapitel 1.2). Das Multi-Matrix-Design minimiert die Testzeit für die Schülerinnen und Schüler, erlaubt aber gleichzeitig die Nutzung von größeren Itempools.

3.5.1 NEPS in der Linking-Studie

Der NEPS-Mathematiktest, der in der Linking-Studie eingesetzt wird, entspricht exakt dem NEPS-Mathematiktest, der in der NEPS-Haupterhebung eingesetzt wird (Duchhardt & Gerdes, 2012a). Es gibt ein Testheft, welches 25 Mathematikaufgaben beinhaltet, für dessen Bearbeitung die Schülerinnen und Schüler 30 Minuten Zeit erhalten. Das NEPS-Testheft wurde am zweiten Testtag der Linking-Studie eingesetzt und in die Testhefte des Ländervergleichs hinein rotiert. Das Rotationsdesign wird im folgenden Kapitel noch näher beschrieben.

3.5.2 Ländervergleich in der Linking-Studie

Das Testheftdesign des zweiten Testtages entspricht nur zu Teilen dem in der Hauptuntersuchung eingesetzten Test aus dem Ländervergleich Mathematik Primar (vgl. Kapitel 4.3; Stanat et al., 2012), da zum einen nicht alle Testhefte des Ländervergleichs genutzt wurden und zum anderen in ausgewählten Testheften das NEPS-Testheft (vgl. Tabelle 3.4) hinein rotiert wurde. Ausgewählt wurden neun von den insgesamt 35 Testheften (für Regelschulen) des Ländervergleichs, in denen das Multi-Matrix-Design verwendet wurde. Die neun ausgewählten Testhefte bilden nun die ersten neun Testhefte der Linking-Studie. Jedes der neun Testhefte besteht dabei aus acht Blöcken, wobei die Blöcke in den unterschiedlichen Testheften mehrmals auftreten, um eventuelle Effekte durch das Design besser kontrollieren zu können (z. B. Ermüdungseffekte) (Weirich, Haag & Roppelt, 2012). Insgesamt gibt es in der

Linking-Studie 32 Blöcke aus dem Ländervergleich Mathematik Primar, von denen jeder Block jeweils einen bestimmten mathematischen Inhaltsbereich erfasst und aus mehreren Aufgaben besteht. Sechs Blöcke zählen zum Inhaltsbereich *Zahlen und Operationen*, sieben Blöcke zu *Größen und Messen*, fünf Blöcke mit Aufgaben zu *Muster und Strukturen*, sieben zu *Raum und Form* und ebenfalls sieben Blöcke zum Inhaltsbereich *Daten, Häufigkeiten und Wahrscheinlichkeiten* zählen.

Tabelle 3.4: Testheftdesign der Linking-Studie am zweiten Testtag

	Teil I				**Teil II**			
	20 Minuten		20 Minuten		20 Minuten		20 Minuten	
01	B 02	B 19	B 31	B 12	B 03	B 35	B 11	B 26
02	B 23	B 13	B 18	B 33	B 11	B 24	B 19	B 35
03	B 18	B 31	B 10	B 04	B 14	B 17	B 35	B 06
04	B 35	B 06	B 13	B 26	B 32	B 05	B 15	B 27
05	B 34	B 26	B 06	B 32	B 23	B 12	B 03	B 15
06	B 14	B 08	B 27	B 11	B 18	B 28	B 05	B 16
07	B 27	B 10	B 08	B 25	B 04	B 18	B 13	B 20
08	B 09	B 01	B 25	B 18	B 07	B 21	B 23	B 12
09	B 11	B 29	B 23	B 07	B 15	B 34	B 05	B 31
10	NEPS			B 12	B 03	B 35	B 11	B 26
11	B 23	NEPS			B 11	B 19	B 19	B 35
12	B 18	B 31	B 10	B 04	NEPS			B 06
13	B 35	B 06	B 13	B 26	B 32	NEPS		
14	NEPS			B 32	B 23	B 12	B 03	B 15
15	B 14	NEPS			B 18	B 28	B 05	B 16
16	B 27	B 10	B 08	B 25	NEPS			B 20
17	B 09	B 01	B 25	B 18	B 07	NEPS		

Es wurden weitere acht Testhefte erstellt (Testheft M10-M17). Dazu wurden die ersten acht Testhefte kopiert und hierein rotiert wurde – an unterschiedlichen Positionen – das NEPS-Mathematiktestheft (Beschreibung hierzu vgl. Kapitel: 3.5.1). So wird beispielsweise das Testheft 1 zu Testheft 10, wobei die ersten drei Blöcke durch das NEPS-Testheft ersetzt werden und die übrigen 5 Blöcke erhalten bleiben. Insgesamt umfasst der Ländervergleichs-Test in der Linking-Studie 32 Aufgabenblöcke mit insgesamt 277 Items.

3.5.3 TIMSS in der Linking-Studie

In der Linking-Studie wurde der gesamte Mathematik- und Naturwissenschaftstest aus TIMSS eingesetzt. Die TIMSS-Testhefte für die vierte Klassenstufe wurden nach einem Matrix Sampling Design gestaltet (Mullis, Martin, Ruddock, O'Sullivan & Preuschoff, 2009). Insgesamt gibt es 14 Testhefte, die in Tabelle 3.5 dargestellt werden. Jedes Testheft enthält jeweils zwei Blöcke mit Mathematikaufgaben und zwei Blöcke mit naturwissenschaftlichen Aufgaben. 7 Testhefte beginnen mit zwei Mathematikblöcken (Testheft 1, 3, 5, …), 7 Testhefte mit zwei Naturwissenschaftsblöcken (Testheft 2, 4, 6, …). Die Testhefte wurden miteinander verlinkt, indem jeder der 14 Mathematikblöcke (Blöcke beginnend mit der Kennzeichnung M, z. B. M01, M02, …) und 14 Naturwissenschaftsblöcke (Blöcke beginnend mit der Kennzeichnung S, z. B. S01, S02, …) in jeweils zwei Testheften vorkommen, jedoch nie an derselben Position, um Positionseffekte berücksichtigen zu können. Jeder der 28 Blöcke besteht aus 10 - 15 Aufgaben (Wendt, Tarelli, Bos, Frey & Vennemann, 2012). Soweit möglich wurde darauf geachtet, dass jeder Block Items aus den unterschiedlichen Inhalts- und Anforderungsbereichen enthält. Jeweils acht der insgesamt 14 Mathematikblöcke und acht der insgesamt 14 Naturwissenschaftsblöcke wurden bereits in TIMSS 2007 verwendet (Mullis et al., 2009). Dadurch ist es möglich, einen Trend darzustellen. Neu entwickelt für TIMSS 2011 wurden damit Testaufgaben für jeweils sechs Mathematik- und sechs Naturwissenschaftsblöcke.

Tabelle 3.5: TIMSS 2011 Testheftdesign (nach Mullis, Martin et al., 2009)

	Originaldesign TIMSS			
	18 Minuten	18 Minuten	18 Minuten	18 Minuten
Booklet 1	M01	M02	S01	S02
Booklet 2	S02	S03	M02	M03
Booklet 3	M03	M04	S03	S04
Booklet 4	S04	S05	M04	M05
Booklet 5	M05	M06	S05	S06
Booklet 6	S06	S07	M06	M07
Booklet 7	M07	M08	S07	S08
Booklet 8	S08	S09	M08	M09
Booklet 9	M09	M10	S09	S10
Booklet 10	S10	S11	M10	M11
Booklet 11	M11	M12	S11	S12
Booklet 12	S12	S13	M12	M13
Booklet 13	M13	M14	S13	S14
Booklet 14	S14	S01	M14	M01

3.6 Aufbereitung und Analyse der Daten

In diesem Kapitel werden die zentralen methodischen Auswertungsverfahren vorgestellt, die für das Vergleichen und das Verlinken der drei Tests genutzt werden. Diese Verfahren entsprechen im Wesentlichen dem bewährten Vorgehen von anderen großen Studien im bildungswissenschaftlichen Bereich. Zudem basieren die Auswertungen und die Analysen der drei Tests, die in der Linking-Studie verwendet wurden, auf den Verfahren, die in den Hauptuntersuchungen genutzt wurden. Die Analysemethoden werden im Folgenden getrennt

nach den Methoden für die inhaltliche Gegenüberstellung und für die Analyse der dimensionalen und skalenbezogenen Zusammenhänge dargestellt.

3.6.1 Analysen zur inhaltlichen Gegenüberstellung der Studien

Grundsätzliches Ziel der inhaltlichen Gegenüberstellung der Studien ist, die Gemeinsamkeiten und Unterschiede der Studien aufzuzeigen (Forschungsfrage 1a bis Forschungsfrage 1e; vgl. Kapitel 2). Dies geschieht anhand vorher definierter Kriterien in einem strukturierten Reviewverfahren. Der Vergleich wird anhand der zur Verfügung stehenden Arbeiten aus den jeweiligen Studien, also beispielsweise aus den Rahmenkonzeptionen der Studien und aus den Ergebnisberichterstattungen, vorgenommen. Die Informationen aus den jeweiligen Arbeiten werden berichtet und einander gegenübergestellt. Ergänzend hierzu werden zwei weitere Expertenreviews durchgeführt und ausgewertet, um detailliertere Informationen über die Vergleichbarkeit der mathematischen Inhalte von NEPS mit TIMSS und mit dem Ländervergleich (Expertenreview I) und über die Merkmale der in den Studien verwendeten Items (Expertenreview II) zu erhalten. Diese beiden Expertenreviews werden im Folgenden kurz vorgestellt. Die Auswertung der Expertenreviews erfolgt auf einer deskriptiven Ebene, jedoch wurden Maße für die Übereinstimmung der Raterinnen und Rater berechnet. Daher wird im Anschluss an die Vorstellung der beiden Expertenreviews kurz erläutert, wie die Übereinstimmungsmaße berechnet werden.

3.6.1.1 Expertenreview I – Unterschiede in den Aufgaben hinsichtlich mathematischer Aspekte und den Anforderungsbereichen

Ziel des Expertenreviews ist die Klassifikation der NEPS-Items in die Rahmenkonzeptionen von TIMSS für die vierte Jahrgangsstufe und dem Ländervergleich Mathematik Primar, um Aussagen darüber treffen zu können, inwieweit sich die Rahmenkonzeptionen der drei Studien ähneln (jingle fallacy vs. jangle fallacy). (Kreuz-) Klassifikationen wurden bereits in anderen Linking-Studien verwendet, wie beispielsweise von Cartwright (2012), Grønmo und Olsen (2007), Neidorf et al. (2006) und Nohara (2001). Die Studien haben die Relevanz eines detaillierten Vergleichs durch (Kreuz-)Klassifikationen aufgezeigt, weil sich oftmals erst dadurch Unterschiede aufdecken lassen, die bei der Gegenüberstellung der Rahmenkonzeptionen nicht sichtbar geworden sind. Die Klassifikation der 24 NEPS-Items wird

von drei Experten vorgenommen, die viel Erfahrung im Bereich von mathematischen Schulleistungsstudien mitbringen. In einem ersten Schritt haben die Experten die NEPS-Items in die Rahmenkonzeption von TIMSS eingeordnet. Die Einordnung erfolgte hinsichtlich der Inhaltsbereiche (Einfachauswahl) und der kognitiven Anforderungsbereiche (Mehrfachzuordnung innerhalb der Subkategorien). In einem zweiten Schritt erfolgte die Einordnung der 24 NEPS-Items in die Rahmenkonzeption des Ländervergleichs. Hier wurden die Aufgaben den Inhaltsbereichen (Mehrfachauswahl) und den prozessbezogenen Kompetenzen (Mehrfachauswahl) zugeordnet. In einem dritten Schritt wurden die Aufgaben hinsichtlich ihrer curricularen Validität beurteilt. Grund hierfür war, dass der NEPS-Test ursprünglich für die fünfte Jahrgangsstufe entwickelt wurde und untersucht werden sollte, inwiefern der NEPS-Test den curricularen Vorgaben entspricht. Dafür wurde für jede Aufgabe erfasst, in welcher Klassenstufe der Inhalt laut Lehrplan behandelt wurde (Klassenstufe 1, 2, 3, 4 oder nach Klassenstufe 4), inwiefern die Schülerinnen und Schüler mit der Aufgabenstellung, den grafischen Darstellungsformen und den mathematischen Begriffen vertraut sind (Skala von überhaupt nicht vertraut bis sehr vertraut) und für wie wichtig die Experten die Förderung der erforderlichen Kompetenz für die Aufgabenlösung erachten (Skala von sehr gering bis sehr hoch).

3.6.1.2 Expertenreview II – Unterschiede in den Aufgaben hinsichtlich formaler Aspekte und der sprachlichen Schwierigkeit

Ziel der Aufgabenklassifikation ist, die Aufgaben hinsichtlich vielfältiger formaler und sprachlicher Merkmale zu untersuchen, um einen Vergleich der drei Studien auf Itemebene vorzunehmen. Die Einordnung wurde wiederum von drei Experten vorgenommen, die für die Aufgabenklassifikation extra geschult wurden. Insgesamt wurden 24 NEPS-Aufgaben, 175 TIMSS-Aufgaben und 261 Ländervergleichs-Aufgaben von den Experten klassifiziert. Die Mathematikaufgaben wurden in einem ersten Schritt in Bezug auf einige formalen Eigenschaften untersucht (vgl. Tabelle 3.6), um die Gemeinsamkeiten und Unterschiede in den drei Studien aufzuzeigen. Verschiedene Untersuchungen konnten zeigen, dass sich Studien u. a. hinsichtlich dieser formalen Kriterien unterscheiden bzw. dass diese formalen Kriterien einen Einfluss auf die empirische Schwierigkeit der Aufgaben haben können, für Zweitsprachenlerner und Muttersprachler unterschiedlich schwierig sein können und damit

Tabelle 3.6: Formale Merkmale der Aufgabenklassifikation

Formale Eigenschaften	**Erläuterungen**
Antwortformat	Es wird unterschieden zwischen geschlossenem, halboffenem und offenem Antwortformat. Zusätzlich wird erfasst, ob die Lösung in ein grafisches Element integriert werden muss und ob Text in der Antwort vorhanden ist.
Term oder Formel	1 = Term/Formel vorhanden, Beispiel: Entfernung = (SEKUNDEN: 3) km
Tabelle	1 = Tabelle vorhanden, Tabellen mit Informationen, die zur Lösung der Aufgabe genutzt werden müssen
Graph, Grafik oder Diagramm	1 = mathematische Abbildung vorhanden
Bild oder Foto	1 = Abbild aus der Realität vorhanden, Beispiel: Bild eines Reiters, Bild von Streichhölzern
Kontextbezug	1 = Kontextbezug vorhanden, wenn die Aufgabe in ein (konkretes) Beispiel aus der Realität eingebettet ist.
Stimulus	1 = Stimulus vorhanden; Als Stimulus werden Aufgaben verstanden, die einen Einführungstext haben, auf den sich mehrere Aufgaben beziehen.
Anzahl der Sätze	Die Anzahl der Sätze wurde erfasst, um die Menge des Lesetextes zu erfassen. Die Anzahl wird hier sowohl für den Stimulus (falls vorhanden) als auch für die jeweilige Aufgabenstellung angegeben.
Anzahl der Wörter	Die Anzahl der Wörter wurde erfasst, um die Menge des Lesetextes zu erfassen. Die Anzahl wird hier für den Stimulus (falls vorhanden), den Antworttext (falls vorhanden) und für die jeweilige Aufgabenstellung angegeben.
Anzahl der Aufgaben pro Stimulus	Als Stimulus werden Aufgaben verstanden, die einen Einführungstext haben, auf den sich mehrere Aufgaben beziehen.
Anzahl der Antwortalternativen	Die Anzahl von Antwortalternativen wird nur für Multiple-Choice Aufgaben erfasst.
Anzahl mathematischer Begriffe	Als mathematische Begriffe werden Begriffe gezählt, deren inhaltliche Bedeutung in der Umgangssprache entweder nicht vorkommt oder von diesem abweicht (z.B. Volumen eines Körpers, Figur, Bruch oder Dividieren).

auch einen Einfluss auf den gesamten Test haben können (vgl. u. a. Grønmo & Olsen, 2007; Haag, Heppt, Stanat, Kuhl & Pant, 2013; National Center for Education Statistics, 2013; Neidorf

et al., 2006; Nohara, 2001; Prenzel, Häußler, Rost & Senkbeil, 2002; Wolf & Leon, 2009; Wu, 2010; Shaftel, Belton-Kocher, Glasnapp & Poggio, 2006). Die Aspekte, hinsichtlich derer die Mathematikaufgaben klassifiziert werden sollen (vgl. Tabelle 3.6), ist eine Auswahl der Aspekte, die in den oben genannten Studien untersucht wurden. Zusätzlich wurden einige Aspekte aus dem Klassifikationsschema für Mathematikaufgaben von Jordan und seiner Forschungsgruppe (Jordan et al., 2006) für das eigene Klassifikationsschema – zum Teil in abgewandelter Form – übernommen.

Der zweite Fokus der Aufgabenklassifikation liegt auf der sprachlichen Komplexität der Testaufgaben in den drei Tests. Wie bereits aufgezeigt wurde, sollen die Mathematikaufgaben in den Schulleistungsstudien prozedurale Kompetenzen, wie z. B. Kommunizieren oder Argumentieren, erfassen. Diese Kompetenzen sind immer mit dem Gebrauch von Sprache verbunden, ob verbal oder schriftlich. Zudem werden Mathematikaufgaben vermehrt in einen Kontext eingebunden, der sprachlich dargeboten wird (Mathematical Literacy). Dadurch bedingt, rückt die Sprache in Mathematikaufgaben immer weiter in den Vordergrund und das Verstehen des Textes einer Mathematikaufgabe wird zu einer grundlegenden Voraussetzung für das Lösen. In vielen Studien konnte bereits gezeigt werden, dass die sprachlichen Anforderungen einer Mathematikaufgabe einen Einfluss auf die (empirische) Schwierigkeit der Aufgaben haben können bzw. für Zweitsprachenlerner und Zweitsprachenlernerinnen sowie Muttersprachler und Muttersprachlerinnen unterschiedlich schwierig sein können (vgl. u. a. Reusser, 1997; Barbu, Otilia C. & Beal, 2010; Beal, Adams, Cohen & Paul R., 2010; Heinze, Herwartz-Emden & Reiss, 2007; Shaftel et al., 2006; Wolf & Leon, 2009). Eine nähere Betrachtung der sprachlichen Gestaltung von Mathematikaufgaben scheint daher auch im Rahmen eines Vergleichs als sehr sinnvoll. Ist die sprachliche Schwierigkeit der Mathematikaufgaben in den drei Studien unterschiedlich, so hat dies ebenfalls Auswirkungen auf die Ähnlichkeit der gemessenen Konstrukte.

Der Text einer Mathematikaufgabe kann unterschiedliche sprachliche Anforderungen an die Schülerinnen und Schüler stellen. Diese sprachlichen Anforderungen entstehen auf verschiedenen Ebenen, denn die Sprache ist ein komplexes semiotisches System (Edmondson & House, 1993). Für eine systembezogene Betrachtung von Sprache nennen Linke, Nussbaumer und Portmann-Tselikas (2001) vier Ebenen, die mindestens zu unterscheiden sind: Die Ebene (1) der Laute, (2) der Morpheme und Wörter, (3) der Sätze und (4) der Texte.

Jede dieser Ebenen kann unterschiedliche Schwierigkeiten für die Rezipientinnen und Rezipienten einer Mathematikaufgabe bergen. In dieser Studie wurde basierend auf drei der vier dargestellten Sprachebenen ein Schema zur Feststellung der sprachlichen Komplexität von Mathematikaufgaben entwickelt (die Ebene der Laute wird nicht berücksichtigt). Das Schema (vgl. Tabelle 3.7) beinhaltet unterschiedliche Schwierigkeitsstufen, die u. a. auf den Theorien des (Zweit-)Spracherwerbs (vgl. u. a. Grimm & Weinert, 1998; Klann-Delius, 2008; Ahrenholz, 2010), des Lesenlernens (vgl. u. a. Rosebrock & Nix, 2011; Junk-Deppenmeier & Schäfer, 2010; Westhoff, 1997) und der Literalitätsentwicklung (vgl. u. a. Apeltauer, 2003; Becker, 2005) basieren.

Tabelle 3.7: Aspekte der sprachlichen Komplexität in der Aufgabenklassifikation

Sprachliche Aspekte	**Erläuterungen**
Wortebene	Die sprachliche Schwierigkeit der Wörter wurde mit einer vierstufigen Skala erfasst. Aufgaben, die keine zusammengesetzten und/oder abgeleiteten Wörter und vorwiegend Konkreta verwenden werden als sprachlich einfacher eingestuft als Sätze, die Modalverben oder zusammengesetzte Verben aufweisen.
Satzebene	Die sprachliche Schwierigkeit auf Satzebene wurde mit einer vierstufigen Skala erfasst. Als leichter wird eine Aufgabe angesehen, in der keine Stilmittel Verwendung finden. Als schwieriger werden Aufgaben erachtet, die z. B. implizite Wiederaufnahmestrukturen oder Ellipsen aufweisen.
Textebene	Die sprachliche Schwierigkeit auf Textebene wurde mit einer vierstufigen Skala erfasst. Als einfacher werden Aufgaben eingeschätzt, die nur aus Hauptsätzen bestehen und als schwieriger gelten Aufgaben, die beispielsweise Konditionalsätze oder Schachtelsätze aufweisen.

3.6.1.3 Beurteilerübereinstimmung und –reliabilitäten

Für die beiden Expertenreviews mit jeweils drei Experten werden Übereinstimmungsmaße und Beurteilerreliabilitäten berichtet. Die Beurteilung durch die Rater gilt als reliabel, wenn alle Rater zu einem ähnlichen bzw. zu einem gleichen Urteil kommen. Für die Analyse der Übereinstimmung sowie für die Bestimmung der Beurteilerreliabilität existieren viele Methoden, die je nach Datenstruktur und je nach Skalenniveau auszuwählen sind. Maße zur Übereinstimmung geben an, ob die Rater das

jeweilige Item exakt gleich ein- bzw. zuordnen (Konkordanz). Die Maße der Beurteilerreliabilität geben das Ausmaß an, wie ähnlich sich die Rater in ihrer Beurteilung sind (relative Gleichheit).

Maße zur Beurteilerübereinstimmung

Die prozentuale Übereinstimmung (PÜ) ist das einfachste Maß der Übereinstimmung und erlaubt eine anschauliche Interpretation (Wirtz & Caspar, 2002). Hierbei wird der prozentuale Anteil an Fällen angegeben, bei denen die Rater exakt übereinstimmen. Ein Problem besteht darin, dass es auch ein gewisses Maß an Übereinstimmung gibt, wenn die Rater völlig zufällig urteilen würden, was zu einer Überschätzung der Konkordanz führt. Daher bietet es sich an, zusätzlich zufallskorrigierte Übereinstimmungsmaße mit anzugeben, wie z. B. das am häufigsten angewendete Maß Cohens Kappa (κ). Das Cohens Kappa basiert grundsätzlich auf einer prozentualen Übereinstimmung, bezieht aber zusätzlich noch den Faktor des Zufalls mit ein. Cohens Kappa ist ein standardisiertes Maß, welches Werte zwischen -1 und +1 annehmen kann, wobei +1 als perfekte Übereinstimmung zu interpretieren ist. Welche Werte als gut zu interpretieren sind, ist abhängig von den zu ratenden Objekten (Wirtz & Caspar, 2002). Oftmals werden Werte größer als .75 als sehr gut interpretiert, Werte zwischen .60 und .75 als gut und Werte zwischen .4 und .6 als akzeptabel. Die Maße zur Beurteilerübereinstimmung gelten im Regelfall für zwei Rater. Wenn mehr als zwei Rater hinzugezogen werden, gilt der Median der Raterpaare als Gütemaß der Übereinstimmung (Wirtz & Caspar, 2002).

Beurteilerreliabilität

Maße zur Bestimmung von Beurteilerreliabilitäten verlangen keine exakte Gleichheit der Raterurteile, sondern bestimmen die relative Lage der Ergebniswerte zum Mittelwert (Wirtz & Caspar, 2002). Reliabilitätsmaße können sowohl für ordinalskalierte als auch für intervallskalierte Daten berechnet werden. Wurden die Abstände der ordinalskalierten Daten empirisch auf Gleichheit überprüft, können die Maße für intervallskalierte Daten verwendet werden. Dies gilt auch, falls die Gleichabständigkeit moderat verletzt ist, hier sollten jedoch im Zweifel zusätzlich Maße für ordinalskalierte Daten angegeben werden (Wirtz & Caspar, 2002). Für intervallskalierte Daten gilt die Intra-Klassen-Korrelation als geeignetstes Maß, welches die Korrelation beliebiger Raterpaare (auch mehrerer) angibt. Die Werte liegen

zwischen 0 und 1 und ein Wert ab .7 gilt als gut, wobei es sich auch hier nur um einen Richtwert handelt.

3.6.2 Analysen zu dimensionalen und skalenbezogenen Zusammenhängen zwischen den Tests

Für den empirischen Vergleich der drei Studien sowohl auf dimensionaler (Forschungsfrage 2a) als auch auf skalenbezogener Ebene (Forschungsfrage 2b) werden die Daten der Linking-Studie zum Teil jeweils gemeinsam skaliert. Dafür muss jedoch die Entscheidung für eine Skalierungsmethode erfolgen, da die drei Studien ihre Daten in den Hauptuntersuchungen unterschiedlich skalieren (vgl. Kapitel 4.5). Das 3-PL-Modell, welches in TIMSS Verwendung findet, stellt größere Anforderungen an die Größe der Stichprobe pro Item als das 1-PL-Modell, dass beim Ländervergleich und in NEPS verwendet wird (vgl. zusammenfassend Pietsch et al., 2009). Daher soll im Rahmen der Analysen zur Gegenüberstellung der Studien das 1-PL-Modell als Grundlage dienen. Die Daten sollen zudem mehrdimensional ausgewertet werden, um ebenfalls einen Vergleich auf Ebene der Teildimensionen zu ermöglichen. Die Unterschiede in der Between- bzw. Within-Item-Dimensionalität werden in den Analysen beibehalten (vgl. Kapitel 4.5: Methodischer Vergleich), da der Ländervergleich Mathematik Primar keine Einfachzuordnung der Items vorgenommen hat und hingegen für die Mathematikitems aus NEPS K5 und TIMSS K4 keine Mehrfachzuordnung vorhanden ist. Die Items, die in den jeweiligen Studien in einem Partial Credit Modell modelliert werden, werden in den nachfolgenden Analysen ebenfalls unter Verwendung des Partial Credit Ansatzes geschätzt. Da die Hintergrundmodelle in den drei Studien ebenfalls unterschiedlich sind und nicht alle Informationen vorhanden sind, wird der Plausible Value Ansatz (PV) für die Schätzung der Personenparameter nicht verwendet. Stattdessen werden für die Gegenüberstellung der Studien Weighted Likelihood Estimates (WLEs) geschätzt.

3.6.2.1 Dimensionaler Vergleich

Der dimensionale Vergleich umfasst zwei Aspekte. Zum einen soll die faktorielle Struktur innerhalb der Tests untersucht und verglichen werden und zum anderen zwischen den Tests (Forschungsfrage 2a).

Für den Vergleich der faktoriellen Struktur innerhalb der Tests werden die Datensätze aus der Linking-Studie zunächst in jeweils mehrdimensionalen 1-PL-Modellen skaliert. Die Mehrdimensionalität bezieht sich hierbei auf die in den Studien definierten Inhaltsbereiche (vgl. Kapitel 4.4.1). Der Vergleich anderer Dimensionen wie beispielsweise der prozeduralen Fähigkeiten kann nicht erfolgen, weil keine Zuordnung der Items hierzu vorliegt. Es wird davon ausgegangen, dass sich die Inhaltsbereiche in den Studien voneinander abgrenzen lassen. Weitere Modellannahmen werden wie in den Originalstudien getroffen (vgl. Tabelle 3.8). Im Folgenden werden daher kurz die Skalierungsmethoden der drei Testinstrumente TIMSS, Ländervergleich und NEPS vorgestellt.

TIMSS: Die TIMSS-Daten aus der Linking-Studie werden unter Annahme einer Between-Item-Dimensionalität und unter Verwendung des Partial-Credit-Modells geschätzt.

Ländervergleich: Die Daten aus dem Ländervergleich der Linking-Studie werden unter Annahme einer Within-Item-Dimensionalität skaliert.

NEPS: Die NEPS Daten der Linking-Studie werden unter Annahme einer Between-Item-Dimensionalität sowie eines Partial-Credit-Modells geschätzt.

Die Skalierung der drei Tests für den dimensionalen Vergleich basieren auf freien Schätzungen der Itemparameter. Hierfür gab es zwei Gründe. Zum einen geben nicht alle Studien die Itemparameter für die mehrdimensionale Schätzung an und zum anderen ermöglicht die freie Schätzung der Itemparameter die bestmögliche Modellanpassung an die Daten. Ein Nachteil ist, dass die Ergebnisse so nicht auf der Metrik der Hauptuntersuchungen liegen, jedoch wird kein Vergleich mit den originalen Metriken angestrebt, so dass dieser Nachteil zu vernachlässigen ist.

Zudem wurden die Personenparameter bei allen Schätzungen auf den Mittelwert „0“ zentriert. Dies liegt daran, dass dies bei der Modellannahme der Within-Item-Dimensionalität in ConQuest vorgeschrieben ist.

Anschließend werden die latenten Korrelationen zwischen den Inhaltsbereichen in den jeweiligen Studien mit der Computersoftware ConQuest berechnet (Wu et al., 2007) und die Ergebnisse werden einander gegenübergestellt.

Tabelle 3.8: Modellannahmen für die Skalierung der TIMSS-, Ländervergleichs- und NEPS-Daten für den dimensionalen Vergleich

	TIMSS	**Ländervergleich**	**NEPS**
Computerprogramm	ACER ConQuest	ACER ConQuest	ACER ConQuest
Identifizierung durch Zentrierung des Mittelwertes auf "0"	Personenparameter	Personenparameter	Personenparameter
1-PL- vs- 3-PL-Modell	1-PL-Modell	1-PL-Modell	1-PL-Modell
Ein- vs Mehrdimensional	Mehrdimensional	Mehrdimensional	Mehrdimensional
Between- vs Within-Item-Dimensionalität	Between-Item-Dimensionalität	Within-Item-Dimensionalität	Between-Item-Dimensionalität
Partial-Credit-Modell	ja	nein	ja
Bestimmung der Personenparameter	WLEs	WLEs	WLEs
Fixieren der Itemparamter aus den HU	nein	nein	nein

Für den Vergleich der faktoriellen Struktur zwischen den Tests werden zwei Modelle geschätzt. Im ersten Modell wird der NEPS-Test gemeinsam mit dem TIMSS-Test skaliert und in einem zweiten Modell wird der NEPS-Test gemeinsam mit dem Ländervergleichs-Test skaliert.

Modell 1: Der NEPS- und der TIMSS-Test werden gemeinsam unter Berücksichtigung der mehrdimensionalen Struktur (bezüglich der Inhaltsbereiche) modelliert. Die Daten werden in einem 1-PL-Modell unter Annahme einer Between-Item-Dimensionalität und eines Partial Credit Modells skaliert. Von dem NEPS-Test gehen alle 24 Items in die Analyse mit ein und von dem TIMSS-Test alle 175 Items. Die Stichprobe umfasst 733 Schülerinnen und Schüler, da nicht alle 752 Schülerinnen und Schüler, die am NEPS-Test am zweiten Testtag teilgenommen haben auch am TIMSS-Test am ersten Testtag teilgenommen haben. Anschließend werden die latenten Korrelationen zwischen den Inhaltsbereichen des NEPS und den Inhaltsbereichen des TIMSS-Tests berechnet. Es werden also beispielsweise die Korrelationen zwischen dem Inhaltsbereiche Quantität (NEPS) und den Inhaltsbereichen aus dem TIMSS-Test Arithmetik, Geometrie und Messen sowie Daten berechnet (vgl. Abbildung 3.3).

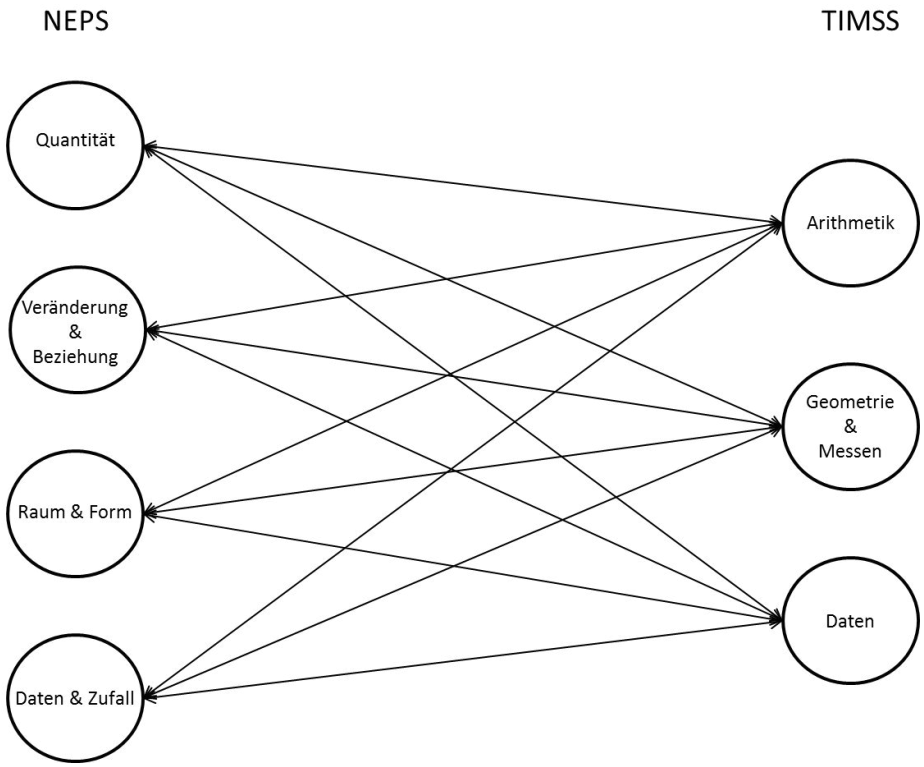


Abbildung 3.3: Berechnung der Korrelationen zwischen den Inhaltsbereichen von NEPS und TIMSS

Modell 2: Hierbei wird der NEPS-Test gemeinsam mit dem Ländervergleichs-Test aus der Linking-Studie skaliert. Die Daten werden unter Annahme eines 1-PL-Modells und einer mehrdimensionalen Struktur (bezüglich der Inhaltsbereiche) modelliert. Da der NEPS-Test ein Partial-Credit-Item aufweist, wurde auch dies bei der Skalierung berücksichtigt. In die Analyse gehen 24 NEPS-Items und 209 Ländervergleichs-Items mit ein. Da nur die Daten der 752 Schülerinnen und Schüler skaliert werden, die an beiden Tests teilgenommen haben, gehen auch nur die Testblöcke aus dem Ländervergleich mit ein, die in den Testheftnummern 10 - 17 vorkommen. Dies erklärt, warum statt der insgesamt 277 Ländervergleichs-Items nur 209 Items mit in die Analysen aufgenommen wurden. Anschließend werden die latenten Korrelationen zwischen den Inhaltsbereichen der Studien berechnet und einander gegenübergestellt. Es werden also beispielsweise die Korrelationen zwischen dem Inhaltsbereich Quantität im NEPS-Test mit den Inhaltsbereichen Zahlen und Operationen, Größen und Messen, Muster und Strukturen, Raum und Form sowie Daten, Häufigkeiten und Wahrscheinlichkeiten des Ländervergleichs-Tests berechnet.

3.6.2.2 Skalenbezogener Vergleich

Im Rahmen des skalenbezogenen Vergleichs soll untersucht werden, ob die Schülerinnen und Schüler in den drei Tests ähnlich abschneiden und ob es sich folglich um ein gemeinsames Konstrukt mathematischer Kompetenz handelt (Forschungsfrage 2b; vgl. Kapitel 2). Diesbezüglich soll zum einen die Vergleichbarkeit der Reliabilitäten und die Zusammenhänge zwischen den Tests überprüft werden.

Reliabilitäten

Eine adäquate Reliabilität ist eine der Grundvoraussetzungen, die bei einem Equating gegeben sein sollte (vgl. Kapitel 1.1). Genauer gesagt, sollte die Reliabilität der jeweiligen Tests, die gleichgestellt werden sollen (Equating), annähernd übereinstimmen (Dorans & Holland, 2000; Holland & Dorans, 2006; Dorans & Walker, 2007). Und auch wenn im Rahmen dieser Arbeit lediglich ein Linking erfolgen soll, so hat die Reliabilität der jeweiligen Tests dennoch einen Einfluss auf die Güte des Linking. Dorans und Holland (2000) zeigen, dass eine unterschiedliche Reliabilität in den Tests dazu führen kann, dass die Anforderung der Invarianz der Population verletzt wird. Jedoch führe eine Verletzung der Voraussetzung gleicher Reliabilität nicht immer zwangsläufig dazu, dass das Equating bzw. Linking nicht zufriedenstellend ist. Sie verweisen gleichsam darauf, dass der Fokus nicht nur auf der Ähnlichkeit der Reliabilität liegen sollte, sondern auch auf der Höhe der Reliabilität, denn desto höher die Reliabilität der Tests, desto besser sei dies für das Equating bzw. Linking. Die Reliabilitäten werden den Ergebnisberichten der jeweiligen Studien entnommen und einander gegenübergestellt.

Zusammenhänge zwischen den Tests

Es soll zunächst empirisch überprüft werden, ob es sich bei den Tests um unterschiedliche Konstrukte mathematischer Kompetenz handelt oder ob die Tests ein gemeinsames Konstrukt mathematischer Kompetenz abbilden. Wenn, wie in dem vorliegenden Fall, ein SG-Design verwendet wird und die Daten IRT skaliert werden, kann ein Korrelationskoeffizient dazu genutzt werden, Aussagen über den Zusammenhang von zwei Tests zu treffen (Kolen & Brennan, 2010; Dorans & Holland, 2000). Zudem ist es im IRT-Kontext möglich zu überprüfen, ob eher ein ein- oder ein zweidimensionales Modell die Daten der beiden Tests besser fittet.

In einem ersten Schritt wird dies für die Studien NEPS und TIMSS untersucht und in einem zweiten Schritt für die Studien NEPS und Ländervergleich.

Tabelle 3.9: Modellannahmen für die Skalierung der TIMSS-, Ländervergleichs- und NEPS-Daten für den skalenbezogenen Vergleich hinsichtlich der Zusammenhänge zwischen den Tests

	TIMSS	**Ländervergleich**	**NEPS**
Computerprogramm	ACER ConQuest	ACER ConQuest	ACER ConQuest
Identifizierung durch Zentrierung des Mittelwertes auf "0"	Personenparameter	Personenparameter	Personenparameter
1-PL- vs- 3-PL-Modell	1-PL-Modell	1-PL-Modell	1-PL-Modell
Ein- vs Mehrdimensional	Eindimensional	Eindimensional	Eindimensional
Between- vs Within-Item-Dimensionalität	-	-	-
Partial-Credit-Modell	ja	nein	ja
Bestimmung der Personenparameter	WLEs	WLEs	WLEs
Fixieren der Itemparameter aus den HU	nein	nein	nein

NEPS und TIMSS: Um der oben genannten Frage nachzugehen, wurden zwei Modelle geschätzt. In dem ersten Modell werden der NEPS-Test und der TIMSS-Test aus der Linking-Studie gemeinsam skaliert, wobei alle NEPS-Items auf einer Dimension laden und alle TMSS-Items auf einer anderen Dimension laden (vgl. Abbildung 3.4: Modell 1: zweidimensionale Skalierung des NEPS- und TIMSS-Tests). In die Analyse gehen 24 NEPS-Items und 175 TIMSS-Items mit ein, sowie die Daten von 733 Schülerinnen und Schülern. Die Daten werden unter Annahme eines 1-PL-Modells und eines Partial-Credit-Modells geschätzt (vgl. Tabelle 3.9). Anschließend wird die latente Korrelation zwischen den beiden Dimensionen berechnet. In dem zweiten Modell wird der NEPS-Test gemeinsam mit dem TIMSS-Test eindimensional unter gleichen Modellannahmen wie bei Modell 1 skaliert. Um zu überprüfen, welches Modell besser auf die Daten passt, werden Modellgeltungstests zum Vergleich herangezogen (vgl. Kapitel 3.6.2.3).

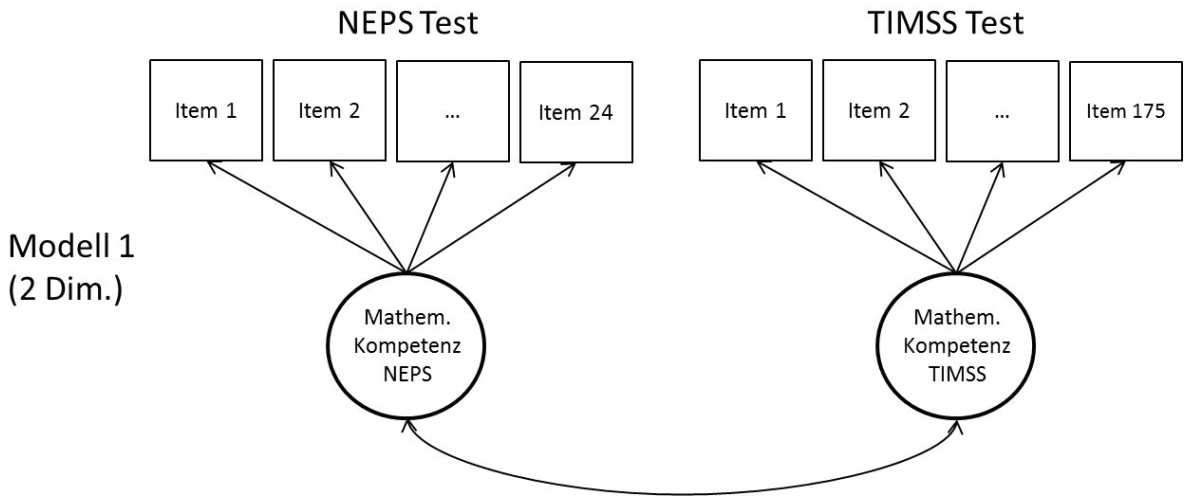


Abbildung 3.4: Modell 1: zweidimensionale Skalierung des NEPS- und TIMSS-Tests

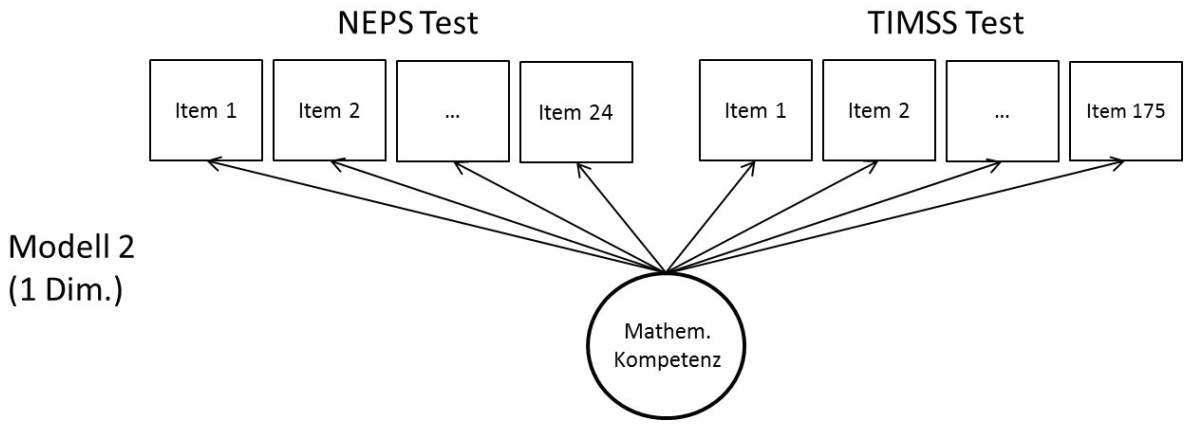


Abbildung 3.5: Modell 2: eindimensionale Skalierung des NEPS- und TIMSS-Tests

NEPS und Ländervergleich: In einem zweiten Schritt wird überprüft, ob der NEPS-Test und der Ländervergleichs-Test ein gemeinsames Konstrukt mathematischer Kompetenz erfassen. Hierzu werden wiederum zwei Modelle geschätzt. In einem ersten Modell werden die NEPS- und die Ländervergleichs-Items gemeinsam skaliert, wobei die Items auf zwei getrennten Dimensionen laden (vgl. Abbildung 3.6). In die Analyse gehen 24 NEPS-Items und 209 Items des Ländervergleich-Tests ein. Die Stichprobe besteht aus 752 Schülerinnen und Schülern. Die Annahmen, die der Skalierung zugrunde liegen sind der Tabelle 3.9 zu entnehmen. Anschließend wird die latente Korrelation zwischen den beiden Dimensionen berechnet. In dem zweiten Modell laden alle Items des NEPS- und des Ländervergleichs-Tests auf einer gemeinsamen Dimension mathematischer Kompetenz (vgl. Abbildung 3.7). Wiederum werden Modellgeltungstests zum Vergleich der beiden Modelle herangezogen (vgl. Kapitel 3.6.2.3).

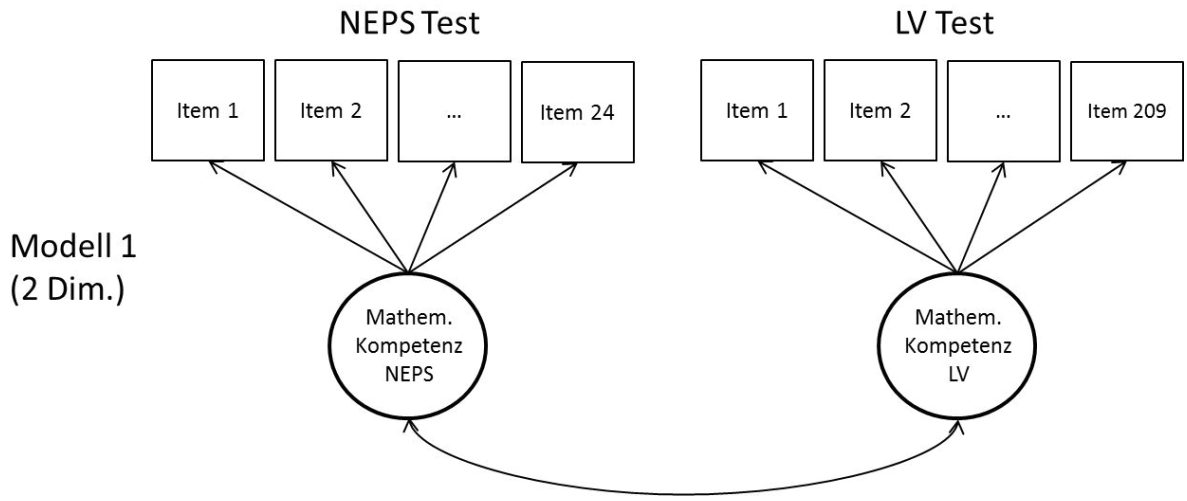


Abbildung 3.6: Modell 1: zweidimensionale Skalierung des NEPS- und Ländervergleich -Tests

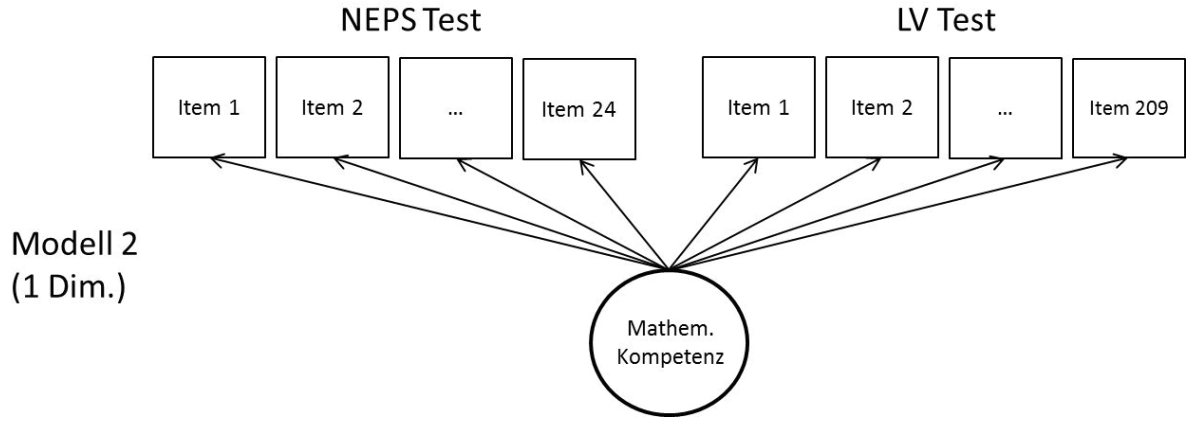


Abbildung 3.7: Modell 2: eindimensionale Skalierung des NEPS- und Ländervergleich -Tests

3.6.2.3 Modellgeltungstests und Modellvergleiche

Bei einer Auswertung der Daten stellt sich die Frage, ob das gewählte Modell tatsächlich zu den Daten passt. Dies kann mit Hilfe von Modellgeltungstests untersucht werden, wobei der Untersuchung immer ein Vergleich zu Grunde liegt. Ein einzelnes Modell kann nicht hinsichtlich seiner Gültigkeit untersucht werden, sondern es wird immer im Vergleich zu einem anderen Modell getestet (Rost, 1996). Bei der Entscheidung, welches Modell besser auf die Daten passt, sollte immer das Einfachheitskriterium berücksichtigt werden, welches aussagt: „Je einfacher eine Theorie ist, desto besser ist sie“ (Rost, 1996, S. 324). Die bedeutet, dass der Aufwand an Modellparametern möglichst gering gehalten werden sollte, zumindest solange mit der einfacheren Theorie auch dieselben Sachverhalte erklärt werden können (Parsimonitätsprinzip).

Für den Modellvergleich sollen die informationstheoretischen Maße herangezogen werden. Der erste Indize ist das Akaikes Information Criterion (AIC). Der AIC berücksichtigt die Devianz und die Anzahl der Modellparameter, überidentifiziert jedoch oftmals die Modellpassung (Dziak, Coffman, Lanza & Li, 2012). Der AIC ist dann den anderen Kriterien vorzuziehen, wenn wenig Items Verwendung fanden und große Patternhäufigkeiten vorliegen oder wenn die Stichprobe klein ist (Rost, 1996; Dziak et al., 2012). Das Best Information Criterion (BIC) berücksichtigt zusätzlich noch den Stichprobenumfang und bietet sich bei Tests mit vielen Items und kleinen Patternhäufigkeiten oder bei größeren Stichproben an (Rost, 1996; Dziak et al., 2012). Der BIC unteridentifiziert jedoch die Modellpassung. Das Consistent AIC (CAIC) ist das letzte informationstheoretische Maß und hierbei handelt es sich um den korrigierten AIC. Vorteil soll sein, dass der CAIC auch bei größeren Stichproben konsistent ist (Rost, 1996). Bei allen Indizes gilt, je kleiner der Wert, desto besser passt das Modell. Ob die Unterschiede zwischen den Indizes bedeutend sind, kann mit einem Chi-Quadrat Test berechnet werden.

3.6.3 Linking der Studien

Ziel der Verlinkung der drei Studien ist, die NEPS-Ergebnisse aus dem Mathematiktest für Fünftklässlerinnen und Fünftklässler in einem nationalen (Ländervergleich) und internationalen (TIMSS) Referenzmaßstab einordnen zu können und eine kriteriale Interpretation der NEPS-Ergebnisse zu erlauben. Bedingt durch das Design der Linking-Studie sind für die Übertragung der Skalenmetriken von TIMSS und dem Ländervergleich auf NEPS drei Schwellen zu überwinden: (1) Von den Hauptuntersuchungen $TIMSS_{HU}$ bzw. $Ländervergleich_{HU}$ zu der Untersuchung in der Linking-Studie $TIMSS_{Link}$ und $Ländervergleich_{HU}$, (2) innerhalb der Linking-Studie zwischen $TIMSS_{Link}$ bzw. $Ländervergleich_{Link}$ und $NEPS_{Link}$ sowie (3) von der $NEPS_{Link}$ auf die $NEPS_{HU}$ (vgl. Abbildung 3.8). Dies bedeutet, dass die Verlinkung in mehreren Schritten erfolgen muss. Im Folgenden werden diese Schritte und die verwendeten Methoden im Detail dargestellt, zuerst für die Verlinkung von NEPS auf die TIMSS-Metrik und anschließend von NEPS auf die Metrik des Ländervergleichs. Anschließend werden die Methoden beschrieben, die für die Analyse der Exaktheit und Stabilität der Linking berechnet werden.

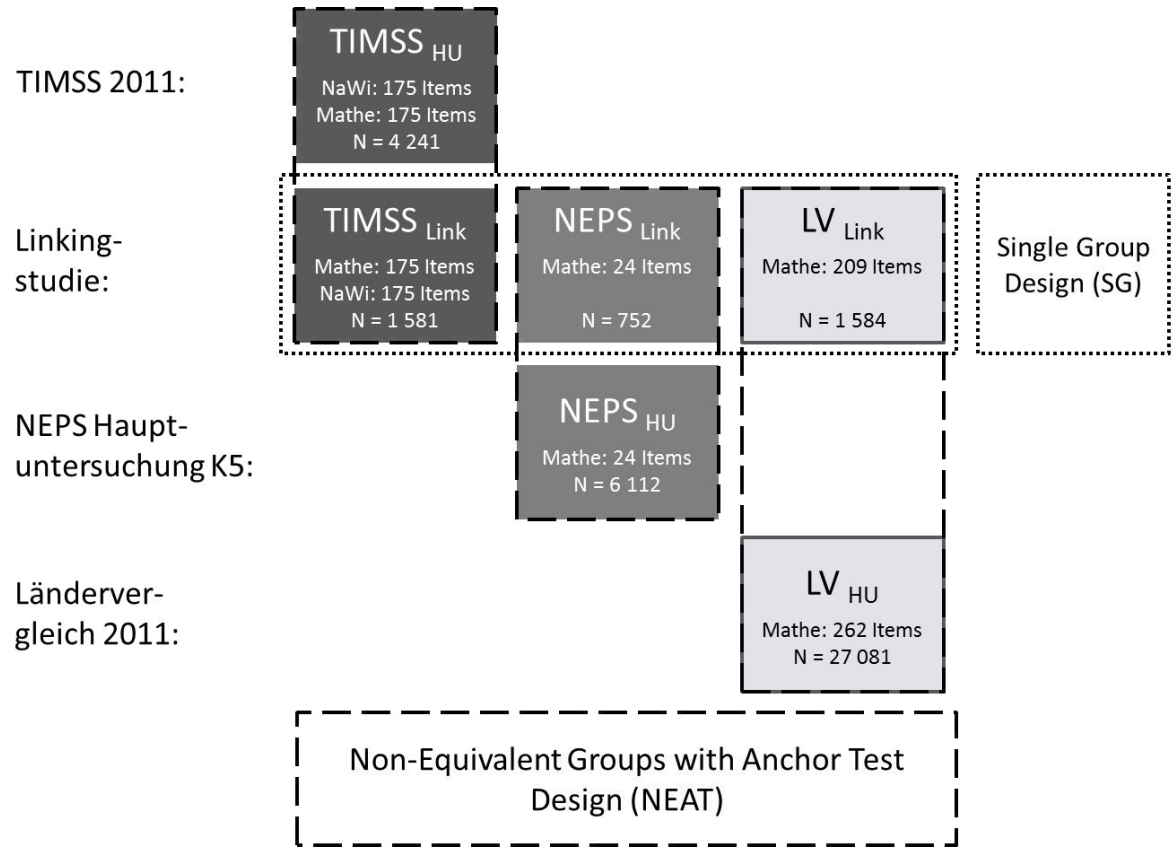


Abbildung 3.8: Datenerhebungsdesign der Linking-Studie

3.6.3.1 Equipercentiles Equating

Bei dem *equipercentile Equating* bzw. *Linking* handelt es sich um eine nicht lineare Funktion, durch die zwei unterschiedliche Skalen (also beispielsweise die der TIMSS- und der NEPS-Linking-Studie) miteinander verbunden werden, indem die Ergebniswerte beider Studien in Perzentilränge unterteilt werden. Anschließend werden die Ergebniswerte über die sich (so nahe wie möglich) entsprechenden Perzentilränge miteinander verlinkt (vgl. u. a. Dorans, Moses & Eignor, 2011; Kolen & Brennan, 2010; Muraki et al., 2000; Livingston, 2004). Wenn also beispielsweise der Ergebniswert -1.5 im NEPS-Mathematiktest (X) einem Perzentilrang von 15 entspricht, und im TIMSS-Mathematiktest (Y) der Ergebniswert 400 ebenfalls einem Perzentilrang von 15, dann ist der Ergebniswert -1.5 in NEPS äquivalent zum Ergebniswert 400 in NEPS. Anders ausgedrückt sind die Ergebniswerte x (ein Ergebniswert des Tests X) und y (ein Ergebniswert des Tests Y) vergleichbar für T (Zielpopulation), wenn $F_T(x) = G_T(y)$, wobei das F die kumulative Verteilungsfunktion der Ergebniswerte in Test X repräsentiert und G die kumulative Verteilungsfunktion der Ergebniswerte in Test Y.

Die Transformationsfunktion für X → Y ist dann wie folgt definiert:

$$y = Equi_{YT}(X) = G_T^{-1}[F_T(X)],$$
(Formel 1)

wobei G^{-1} die inverse Funktion von G bezeichnet.

Vorteil des equipercentile Linking ist, dass die Verteilung der äquivalenten Ergebniswerte (ey(X)) annähernd die gleiche ist wie beim Y-Test. Dies bedeutet, dass die Mittelwerte, die Standardabweichung, die Schiefe und Kurtosis sich ebenfalls approximativ entsprechen (Livingston, 2004; Kolen & Brennan, 2010; Yin et al., 2004). Das equipercentile Linking hat jedoch auch einen Nachteil, denn die Verteilungen der Ergebniswerte sind meist unregelmäßig und unterscheiden sich, je nachdem welche Substichprobe gezogen wird (Dorans et al., 2011; Livingston, 2004). Dies bedeutet, dass die prozentuale Anzahl der Schülerinnen und Schüler nicht sukzessive im Verhältnis zum Ergebniswert ansteigt bzw. wieder abfällt, sondern dass die prozentuale Anzahl schwankt (vgl. Abbildung 3.9). Darüber hinaus sind die äquivalenten Ergebniswerte, die durch eine Verlinkung geschätzt werden, nicht besonders stabil für Punktewerte, die nur von einer geringen Anzahl an Schülerinnen und Schüler erreicht wurden (Pommerich, Hanson, Harris & Sconing, 2004). Diese Irregularitäten in den Verteilungen der Ergebniswerte können zu Irregularitäten in der Equipercentile Linking Funktion führen.

Eine Möglichkeit, dem genannten Problem zu begegnen, ist das sogenannte *Smoothing*, also das Glätten der Daten entweder vor dem Linking (*Presmoothing*) oder nach dem Linking (*Postsmoothing;* vgl. Dorans et al., 2011; Livingston, 2004; Yin et al., 2004). Das Glätten der Daten soll die Irregularitäten der Verteilung eliminieren (vgl. Abbildung 3.9 und 3.10). Die meisten Smoothing-Methoden ermöglichen es, die Stärke der Glättung der Daten zu bestimmen. Desto stärker das Smoothing ist, desto glatter wird die Verteilung, desto eher besteht jedoch auch die Gefahr, dass sich die Form der Verteilung verändert. Wenn die Verteilung zu wenig geglättet wird, werden jedoch nicht alle Irregularitäten entfernt (Dorans et al., 2011; Livingston, 2004).

Eine weitere Einschränkung des equipercentile Linking besteht darin, dass keine äquivalenten Ergebniswerte bestimmt werden können, die über dem höchsten Ergebniswert bzw. unter dem niedrigsten Ergebniswert der beobachteten Ergebniswerte liegen (Livingston, 2004). Dafür werden hingegen auch keine äquivalenten Ergebniswerte geschätzt, die außerhalb des möglichen Ergebnisbereichs liegen (Yin et al., 2004). Jedoch entsprechen sich

die Perzentilränge der beiden Studien, die miteinander verlinkt werden sollen, selten genau. Wenn beispielsweise im NEPS der Testwert 1.67 dem Perzentilrang 64.7 entspricht, gibt es oftmals in Test B, also im TIMSS-Test, nicht exakt denselben Perzentilrang. Im TIMSS-Test entspricht etwa der Testwert 500 dem Perzentilrang 62.4 und der Testwert 501 dem Perzentilrang 65.9. Der Testwert 1.67 im NEPS-Test entspricht also eigentlich einem Testwert zwischen 500 und 501 im TIMSS-Test. Um den Wert zu bestimmen werden im equipercentile Linking die Daten interpoliert. Diese Methode ermöglicht es, eine möglichst gleiche Verteilung der äquivalenten Ergebniswerte zu erhalten (Livingston, 2004).

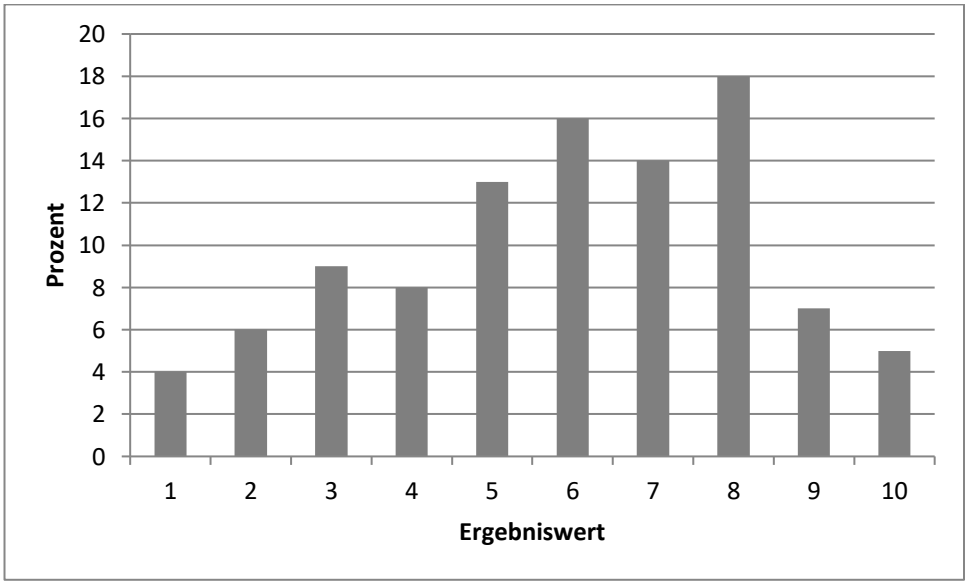


Abbildung 3.9: Beispiel für eine Verteilung von Ergebniswerten

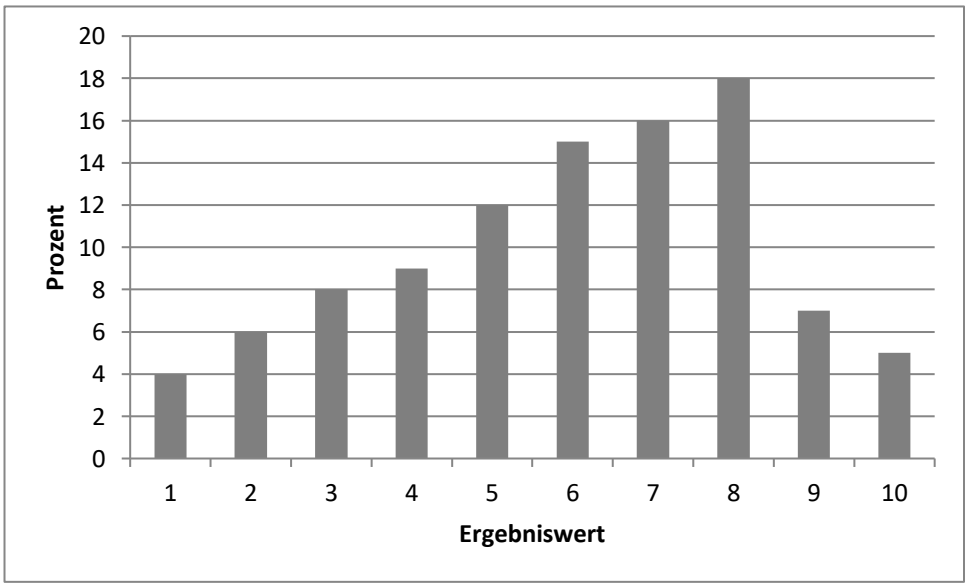


Abbildung 3.10: Beispiel für eine Verteilung von Ergebniswerten - geglättet

3.6.3.2 Linking von NEPS und TIMSS

Die Verlinkung von NEPS mit TIMSS erfolgt in vier Schritten (vgl. Abbildung 3.11). Ziel ist es, die Ergebnisse des NEPS-Mathematiktests K5 auf die Skalenmetrik von TIMSS zu übertragen, um Aussagen darüber treffen zu können, wie die Schülerinnen und Schüler, die an NEPS teilgenommen haben, im internationalen Vergleich abschneiden. An der TIMSS-Hauptuntersuchung der vierten Jahrgangsstufe in 2011 beteiligten sich 4 241 Schülerinnen und Schüler. Der Mathematiktest von TIMSS bestand aus 175 Items und wurde unter Annahme eines 3-PL-Modells modelliert. Martin und Mullis (2012a) berichten die Itemparameter dieser Skalierung.

(1) In einem ersten Schritt werden die 175 TIMSS-Items in der Linking-Studie 2011 ebenfalls unter Annahme eines 3-PL-Modells skaliert. Hierzu werden die Itemparameter (Schwierigkeits-, Trennschärfe- und Rateparameter) bei der Skalierung auf die Itemparameter aus der TIMSS-Hauptuntersuchung fixiert. Dadurch liegen die Daten der Linking-Studie auf der Metrik der TIMSS-Hauptuntersuchung. Anschließend erfolgt eine lineare Transformation der Personenparameter (WLEs) der Linking-Studie auf die internationale TIMSS-Metrik. Die Skalierung bezieht sich auf eine Stichprobe von 733 Schülerinnen und Schüler. Dies ist dem Umstand geschuldet, dass nur 733 von den insgesamt 1 581 Schülerinnen und Schüler in der Linking-Studie sowohl den TIMSS-Mathematiktest (erster Testtag) als auch den NEPS-Mathematiktest (zweiter Testtag) bearbeitet haben und die anschließende Verlinkung (siehe Schritt 3) eine gleiche Stichprobe voraussetzt.

(2) In einem zweiten Schritt wird die Übertragung der Metrik der NEPS-Hauptuntersuchung der mathematischen Kompetenzen in 2010 auf die NEPS-Daten der Linking-Studie 2011 vorgenommen. Hierzu werden die Itemparameter (Schwierigkeitsparameter) aus der NEPS-Hauptuntersuchung (24 Items; 5 193 Schülerinnen und Schüler), welche unter Annahme des Rasch-Modells geschätzt wurden, in der Linking-Studie fixiert. Die NEPS-Daten aus der Linking-Studie werden ebenfalls mit Hilfe des Rasch-Modells skaliert und es werden Personenparameter (WLEs) für die Schülerinnen und Schüler der Linking-Studie geschätzt. Der NEPS-Mathematiktest in der Linking-Studie, der am zweiten Testtag durchgeführt wurde, wurde insgesamt von 752 Schülerinnen und Schülern bearbeitet. In die Analyse gehen jedoch nur 733 der Schülerinnen und Schüler mit ein, da 19 Schülerinnen

und Schüler den TIMSS-Mathematiktest in der Linking-Studie, der am ersten Testtag durchgeführt wurde, nicht bearbeitet haben.

(3) Das Verlinken der NEPS-Linking-Studie auf die Metrik der TIMSS-Linking-Studie erfolgt unter Verwendung der Methode des equipercentilen Linking (vgl. Kapitel 3.6.3.1). Das equipercentile Linking wird mit der Computersoftware LEGS 2.0.1 (Brennan, 2004) durchgeführt. Das Ergebnis ist eine sogenannte Konkordanz-Tabelle, aus der die sich entsprechenden Ergebniswerte von der NEPS- und TIMSS-Linking-Studie abgelesen werden können.

(4) Diese Konkordanz-Tabelle erlaubt den Nutzerinnen und Nutzern des Scientific-Use-Files des NEPS-Mathematiktests für die fünfte Jahrgangsstufe eine Übertragung der NEPS-Mathematikskala auf die TIMSS-Skala vorzunehmen. Dieser vierte Schritt ist jedoch nicht Teil dieser Arbeit. Zudem soll an dieser Stelle auf die Relevanz aufmerksam gemacht werden, die daraus entstehenden Daten im Hinblick auf die in der Äquivalenzanalyse gefundenen Unterschiede vorsichtig zu interpretieren und die Einschränkungen, die durch die Unterschiede gemacht werden müssen, bei der Auswertung und Interpretation zu berücksichtigen. Zudem werden im Rahmen dieser Arbeit auch Ergebnisse zur Exaktheit und Stabilität des Linking berichtet. Auch diese Ergebnisse gilt es bei der Interpretation zu berücksichtigen.

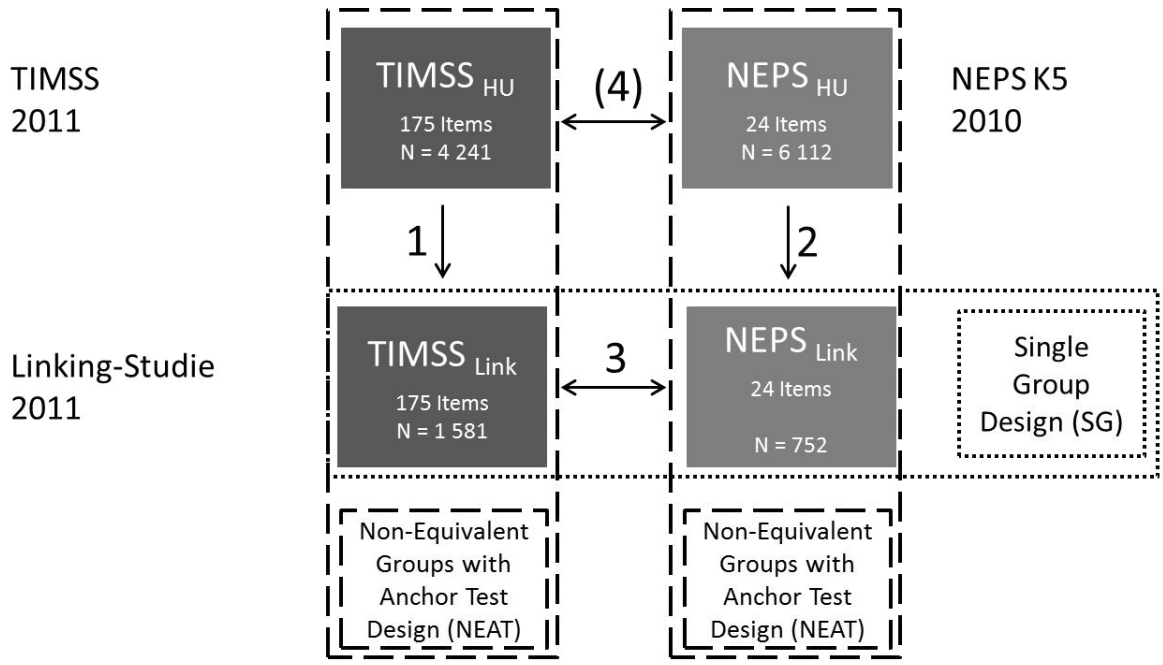


Abbildung 3.11: Verlinkung des NEPS-Tests mit dem TIMSS-Test

3.6.3.3 Linking von NEPS und dem Ländervergleich

Die Verlinkung des NEPS-Mathematiktests mit der nationalen Studie Ländervergleich Mathematik Primar erfolgt analog zu der Verlinkung von NEPS und TIMSS (vgl. vorheriges Kapitel) und umfasst ebenfalls vier Schritte (vgl. Abbildung 3.1.2). In der Linking-Studie wird neben dem TIMSS- und dem NEPS-Mathematiktest ebenfalls der Mathematiktest aus dem Ländervergleich Primar eingesetzt.

(1) Um diese Daten auf die Metrik der Hauptuntersuchung des Ländervergleichs zu transferieren, werden diese Daten unter Fixierung der Itemparameter der Hauptuntersuchung skaliert. Die Itemparameter der Hauptuntersuchung wurden mit Hilfe eines 1-PL-Modells gewonnen. Die Stichprobe bestand aus 27 081 Schülerinnen und Schüler. Der Mathematiktest beinhaltete in der Hauptuntersuchung insgesamt 330 Items. Diese Items werden jedoch nicht alle in der Linking-Studie verwendet (vgl. Kapitel 3.5.2). Wie bereits bei der Beschreibung des Testheftdesigns deutlich wurde, besteht der Ländervergleichs-Test in der Linking-Studie aus 277 der insgesamt 330 Items, die insgesamt von 1 584 Schülerinnen und Schüler bearbeitet wurden. Das NEPS-Testheft wurde jedoch nur in 8 von den 17 Testheften hinein rotiert. Aus diesem Grund besteht die gemeinsame Stichprobe aus N = 752 dieser Schülerinnen und Schüler, die sowohl den Ländervergleichs- als auch den NEPS-Mathematiktest bearbeitet haben. Zudem sind nicht alle 277 Items des Ländervergleichs, die in der Linking-Studie eingesetzt wurden, Teil dieser 8 Testhefte, so dass insgesamt nur 209 Items in die Analysen mit einfließen. Die Skalierung erfolgt analog zur Hautuntersuchung unter Annahme eines 1-PL-Modells, jedoch wird kein Hintergrundmodell mitaufgenommen und es werden WLEs als Personenparameter geschätzt.

(2) In einem zweiten Schritt wird wiederum der NEPS-Test, der in der Linking-Studie verwendet wurde, auf die Metrik der NEPS-Hauptuntersuchung gebracht. Die Skalierung aus dem vorherigen Linking kann nicht verwendet werden, da die Stichprobe leicht different ist. Der Ländervergleichs- und der NEPS-Test haben in der Linking-Studie eine gemeinsame Stichprobe von 752 Schülerinnen und Schülern. Diese Fälle gehen in die Skalierung der NEPS-Linking-Studie mit ein, unter Fixierung der Itemparameter auf die Itemparameter der NEPS-Hauptuntersuchung.

(3) In einem dritten Schritt erfolgt die Verlinkung der beiden Mathematiktests aus der Linking-Studie unter Verwendung der Methode des equipercentilen Linking (vgl. Kapitel 3.6.3.1). Das Ergebnis ist eine Konkordanz-Tabelle, mit Hilfe derer die Ergebnisse aus der NEPS-Linking-Studie auf die Metrik der Ländervergleichsuntersuchung transferiert werden können.

(4) Diese Konkordanz-Tabelle kann wiederum in einem vierten Schritt in der NEPS-Hauptuntersuchung genutzt werden, um die Ergebnisse in einem nationalen Referenzmaßstab interpretieren zu können. Wie bereits im vorherigen Kapitel angemerkt, sind bei der Interpretation jedoch die in dieser Arbeit dargestellten Unterschiede zwischen den Studien genauso zu berücksichtigen wie die Ergebnisse zur Exaktheit und Stabilität des Linking (vgl. das folgende Kapitel).

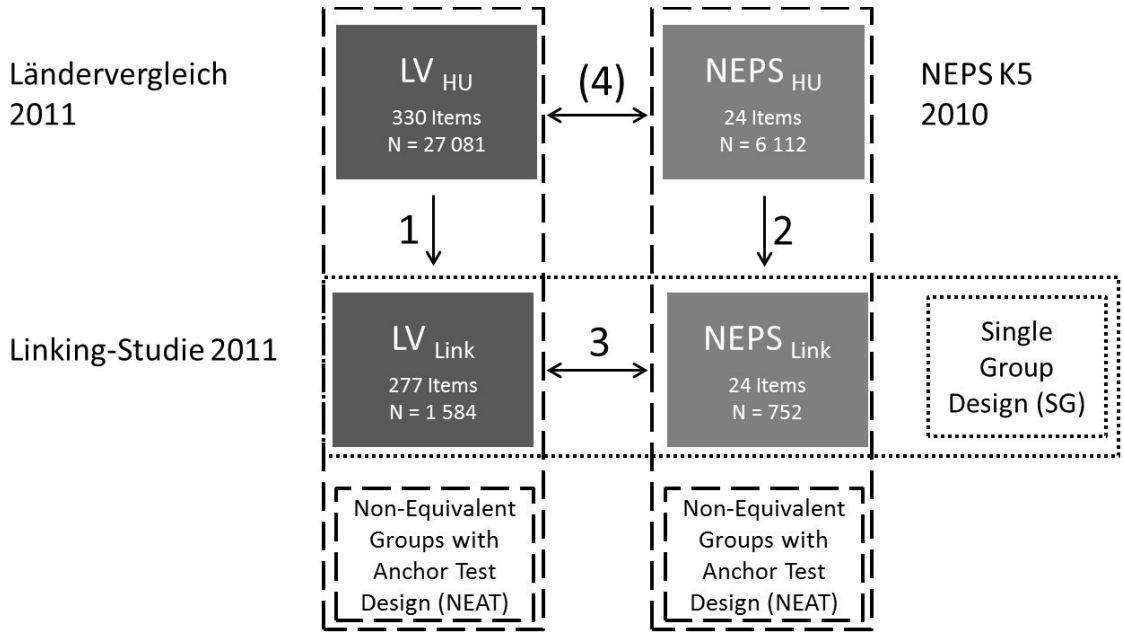


Abbildung 3.12: Verlinkung des NEPS-Test mit dem Ländervergleich -Test

3.6.3.4 Exaktheit und Stabilität des Linking

Im Folgenden wird das Vorgehen zur Analyse der Exaktheit und Stabilität des Linking beschrieben. Anschließend werden noch mögliche Richtwerte für die Interpretation der Ergebnisse vorgestellt.

Exaktheit des Linking

Um die Exaktheit des Linking zu analysieren, werden die Linkingergebnisse mit den Ergebnissen aus der Linking-Studie zum einen hinsichtlich deskriptiver Statistiken verglichen (Mittelwerte, Standardabweichung, Schiefe und Kurtosis) und zum anderen bezüglich der Zuordnungsgenauigkeit zu den Kompetenzstufen.

Sowohl TIMSS als auch der Ländervergleich bieten Kompetenzstufenmodelle an, damit eine kriteriale Interpretation der Ergebnisse erfolgen kann. Ziel dieser Studie ist, die Kompetenzstufenmodelle von TIMSS und dem Ländervergleich für die Interpretation der NEPS-Ergebnisse in einem nationalen und internationalen Rahmen zu nutzen. Nachdem die Daten der NEPS-Studie mit denen der Ländervergleichs- und TIMSS-Studie verlinkt wurden (vgl. Kapitel 3.6.32 und 3.6.3.3) befinden sich die äquivalenten Ergebniswerte auf den Metriken der Ländervergleich- und der TIMSS-Studie. Daher kann im Anschluss an das Linking eine Verteilung der NEPS-Ergebnisse auf die Kompetenzstufen des Ländervergleichs und von TIMSS erfolgen.

Die Zuordnungsgenauigkeit wird mit zwei Maßen angegeben, der prozentualen Übereinstimmung und dem Cohens Kappa (vgl. Kapitel 3.6.1.3).

Stabilität des Linking

Die Invarianz über relevante Subgruppen ist eine wesentliche Voraussetzung für ein Equating bzw. Linking (*Population Invariance Requirement*; vgl. Dorans & Holland, 2000; Holland & Dorans, 2006; Kolen & Brennan, 2010). Nach Dorans und Holland (2000) ist diese Voraussetzung nie exakt gegeben, die Subgruppen sollten jedoch zumindest annähernd invariant sein. Zudem könnten zwei Tests, die sich bezüglich des Konstrukts und der Reliabilität unterscheiden, bei einem equipercentilen Linking nicht invariant bestimmter Subgruppen sein. In diesem Fall sprechen Dorans und Holland von einer Konkordanz für eine bestimmte bzw. gegebene Population. Falls das Linking nicht invariant bezüglich der Subgruppen ist, empfehlen sie, unterschiedliche Linking-Funktionen für die Subgruppen zu schätzen.

Um die Stabilität über Subgruppen zu bestimmen wird die sogenannte *root expected mean square Difference* (REMSD) berechnet (vgl. Dorans & Holland, 2000; Holland & Dorans, 2006; Kolen & Brennan, 2010), die die Differenz zwischen den Ergebniswerten in den Subgruppen in den Fokus nimmt. Der REMSD ist doppelt gewichtet und bezieht die

Häufigkeitsverteilung für jeden Ergebniswert in der Subgruppe und die Subgruppenstichprobengröße mit ein. Zusätzlich wird der *ungewichtete (equally weighted) REMSD* (ewREMSD) mit angegeben. Dieser nimmt für die Differenz der einzelnen Ergebniswerte das gleiche Gewicht an – das Inverse der Gesamtanzahl der Ergebniswerte. Der REMSD basiert dabei auf den ungerundeten Ergebniswerten, da die Ergebnisse dann präziser sind (Yin et al., 2004). Im vorliegenden Fall wird die Invarianz über die Subgruppen getrennt nach Geschlecht berechnet.

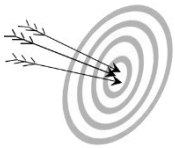
Richtwert für Interpretation der Ergebnisse

Um die Ergebnisse der Analysen hinsichtlich der Gruppen-Invarianzen und der Standardfehler interpretieren zu können, kann ein Richtwert errechnet werden, die sogenannte *score difference that matters* (DTM; vgl. hierzu u. a. Kolen und Brennan, 2010; Yin, Brennan und Kolen, 2004; Holland & Dorans, 2006; Dorans & Liu, 2009). Diese kann jedoch nur für die Interpretation der ungerundeten Werte verwendet werden. Die DTM entspricht einem halben transformierten Ergebniswert. Wenn die Ergebniswerte also in fünferschritten erfolgen, z. B. 5, 10, 15, 20 … Punkte erreicht werden können, liegt die DTM dementsprechend bei DTM = 2.5). Weiterhin kann eine standardisierte DTM (SDTM) berechnet werden, welche zusätzlich die Standardabweichung berücksichtigt und damit den Vorteil bietet, auf der gleichen Metrik wie der REMSD zu liegen. Die SDTM berechnet sich wie folgt: sDTM = DTM/SD.

4 Analysen auf Ebene der inhaltlichen Gegenüberstellung

In diesem Kapitel werden die Rahmen- und Testkonzeptionen der Mathematiktests der drei Schulleistungsstudien TIMSS, Ländervergleich und NEPS beschrieben und einander gegenübergestellt. Dabei ist zu untersuchen, ob die drei Studien auf der Ebene der inhaltlichen Gegenüberstellung vergleichbar sind. Der Schwerpunkt liegt hierbei auf der Vergleichbarkeit hinsichtlich der Anlagen und Ziele der drei Studien (Forschungsfrage 1a), der Stichproben bzw. der Zielpopulationen (Forschungsfrage 1b), der Messbedingungen (Forschungsfrage 1c) sowie der Konstrukte auf konzeptioneller (Forschungsfrage 1d) und methodischer (Forschungsfrage 1e) Ebene. Der Vergleich erfolgt hinsichtlich der in den Forschungsfragen definierten Kriterien. Um einen detaillierteren Vergleich zu ermöglichen, wurden zusätzlich zu den Analysen der Frameworks und der Testkonzeptionen der Studien zwei Expertenreviews ausgewertet (vgl. Kapitel 3:6.1). Diese ermöglichen eine ausführlichere Untersuchung der Unterschiede bzw. Übereinstimmungen der drei Studien hinsichtlich mathematischer Inhalte sowie formaler und sprachlicher Aspekte in den Aufgaben

4.1 Anlage und Ziele



Ziel dieses Kapitels ist die Darstellung der drei Studien TIMSS, Ländervergleich und Nationales Bildungspanel, um zunächst einen Überblick über die drei Studien zu geben. Zusätzlich sollen folgende Fragen beantwortet werden: Warum wurden die Studien ins Leben gerufen bzw. welche Ziele verfolgen sie? Welche Kompetenzbereiche decken die Studien ab? Welche Schlüsse lassen sich aus den Studienergebnissen ziehen? Welche Konzeption liegt den Studien zu Grunde? Was für ein Untersuchungsdesign liegt in den drei Studien vor? In einem anschließenden Zwischenfazit werden die drei Studien einander gegenübergestellt und hinsichtlich der genannten Aspekte verglichen.

TIMSS

Seit dem Jahr 1959 werden Schulleistungsstudien von der *International Association for the Evaluation of Educational Achievement* (kurz: IEA) durchgeführt. Sie erforschen den Ertrag von Bildungssystemen im internationalen Vergleich sowie potentielle Einflussfaktoren zur Erklärung der unterschiedlichen Leistungen der Schülerinnen und Schüler. Ziel ist, eine

Möglichkeit des gegenseitigen Lernens aus Erfahrungen zu bieten (Wendt et al., 2012; Baumert et al., 1997). Erfasst werden u. a. die mathematischen und naturwissenschaftlichen Kompetenzen von Schülerinnen und Schülern in Schlüsseljahrgängen. Zunächst wurden die mathematischen und naturwissenschaftlichen Leistungen in getrennt stattfindenden Untersuchungen erhoben, z. B. FISS (First International Science Study) und FIMS (First International Mathematics Study). *The Third International Mathematics and Science Study* (TIMSS) erfasst erstmals im Jahre 1995 die beiden Inhalte gemeinsam und umfasste sowohl qualitative als auch quantitative Untersuchungsinteressen bzw. Untersuchungsgebiete. An dieser Studie nahm, neben 44 weiteren Staaten, auch die Bundesrepublik Deutschland nach vielen Jahren des Fernbleibens von Schulleistungsstudien teil (Baumert et al., 1997).

Die grundlegende Rahmenkonzeption setzt sich aus drei Aspekten zusammen und wird in Abbildung 4.1 graphisch dargestellt (Bonsen, Lintorf, Bos & Frey, 2008; Wendt et al., 2012). Der erste Schwerpunkt von TIMSS lag auf der Analyse der Curricula und der Schulbücher, bezeichnet als intendiertes Curriculum. Die Absicht war zu untersuchen, welche Inhalte die Schülerinnen und Schüler der beteiligten Länder lernen sollten. Hierzu wurden Expertinnen und Experten der jeweiligen Länder befragt. Der zweite Aspekt betrifft die tatsächlichen Unterrichtsinhalte der verschiedenen Länder, folglich das implementierte Curriculum. Diese beiden Forschungsbereiche werden unter dem Begriff *Curriculum Study* zusammengefasst und verfolgen das gemeinsame Ziel, Testinstrumente zu entwickeln, die inhaltlich möglichst nah an den Curricula aller Länder orientiert sind. Der dritte Schritt war die darauffolgende Erfassung der tatsächlichen mathematischen und naturwissenschaftlichen Kompetenzen der Schülerinnen und Schüler (*Assessment Study*), also das erreichte Curriculum. Hierzu gehörte des Weiteren eine Fragebogenerhebung, u. a. zu der Organisation und Kultur der Schulen sowie zu demografischen und motivationalen Merkmalen der Schülerinnen und Schüler.

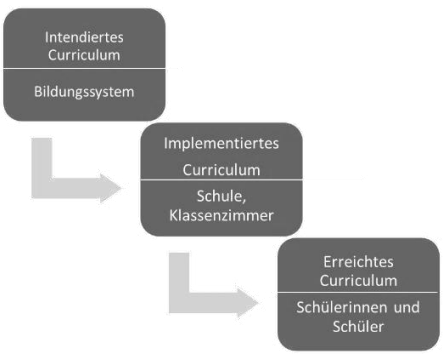


Abbildung 4.1: Das Curriculum Modell (in Anlehnung an Wendt et al., 2012)

TIMSS ist spezialisiert auf drei Gelenkstellen der schulischen Bildung, die meist gleichzeitig überprüft werden. In der Grundschule (Population I) werden die beiden Klassenstufen untersucht, in denen sich im jeweiligen Land die meisten neunjährigen Schülerinnen und Schüler befinden. Oft sind dies die 3. oder 4. Klassen. In der Sekundarstufe I (Population II) bezieht die Studie die beiden Klassenstufen ein, welche den größten Anteil 13-jähriger Schülerinnen und Schüler aufweisen (meist 7. oder 8. Klasse). In der Sekundarstufe II (Population III, TIMSS Advanced) werden die Schülerinnen und Schüler allgemeinbildender und beruflicher Einrichtungen im Abschlussjahrgang getestet (Baumert, 1998). Nicht alle Staaten haben jedoch auch an allen Untersuchungen teilgenommen. Deutschland beteiligte sich beispielsweise 1995 nur an der Population II und III.

Seit der ersten Durchführung im Jahr 1995, finden nun alle vier Jahre Erhebungen der IEA TIMSS Studie (das Akronym steht seit 2003 für *Trends in International Mathematics and Science Study)*, mit unterschiedlicher Länderbeteiligung statt. Aufgrund des Rhythmus ist bei Teilnahme an aufeinanderfolgenden Populationen (z. B. Klassenstufe 4 und 8), welche jeweils ebenfalls vier Jahre auseinander liegen, die Möglichkeit gegeben, die Entwicklung einer Kohorte aufzuzeigen, also Trendanalysen durchzuführen (Mullis, Martin, Foy & Arora, 2012).

Die aktuelle TIMSS-Studie, die 2011 durchgeführt wurde, untersuchte im internationalen Vergleich die Populationen I und II. Deutschland beteiligte sich jedoch – wie auch in 2007 – nur an der Grundschuluntersuchung (Stanat et al., 2012). Damit ist TIMSS in Deutschland eine reine Querschnittsuntersuchung. Insgesamt nahmen 63 Länder an dieser TIMS-Studie teil. Besonderheit in diesem Jahrgang war, dass TIMSS zeitgleich mit PIRLS (*Progress in International Reading Literacy Study*), in Deutschland eher unter dem Begriff IGLU (*Internationale Grundschul-Lese-Untersuchung*) bekannt, durchgeführt worden ist, da PIRLS seit 2001 in einem Fünfjahresrhythmus stattfindet (Wendt et al., 2012; Mullis et al., 2012). PIRLS wird ebenfalls von der IEA organisiert und durchgeführt. Insgesamt nahmen in etwa 40 Länder an beiden Studien gleichzeitig teil. Während sich TIMSS auf die mathematischen und naturwissenschaftlichen Fächer konzentriert, untersucht PIRLS die Lesekompetenz der Schülerinnen und Schüler am Ende der Grundschulzeit. Diese Vereinigung führt zu zusätzlichen Informationen nicht nur auf der Kompetenzebene, sondern auch zu weiteren Hintergrundinformationen, da z. B. in PIRLS eine Befragung der Erziehungsberechtigten zum Rahmenprogramm gehört, wohingegen TIMSS die Erziehungsberechtigten nicht befragt.

Neben den Testinstrumenten zur Erfassung der mathematischen und naturwissenschaftlichen Kompetenzen setzt TIMSS sowohl Schüler-, Lehrer- und Schulleiterfragebögen als auch einen Fragebogen für die Erhebung der curricularen Organisation, der Struktur und der Inhalte der Teilnahmeländer ein (Mullis et al., 2009). Für TIMSS 2011 wurden die Test- und Fragebogeninstrumente überarbeitet bzw. ergänzt, die bereits im Jahr 2007 Verwendung fanden, indem die Rückmeldungen zu den Erfahrungen der einzelnen Länder von einer internationalen Expertengruppe analysiert und diskutiert wurden (Mullis et al., 2009). Außerdem ist bei einer Studie, die schon lange Jahre Bestand hat, immer wieder zu kontrollieren, ob sie inhaltlich noch die aktuellsten Erkenntnisse der Forschung z. B. in Bezug auf die Lernziele und die gegenwärtigen politischen Interessen abbilden. Die Instrumente wurden von dem sogenannten Item Review Committee begutachtet und aktualisiert. Letztlich gingen sowohl in den Mathematik- als auch in den Naturwissenschaftstests jeweils acht unveränderte Blöcke aus TIMSS 2007 ein, um die Messung von Trends zu gewährleisten, sowie jeweils sechs neue Blöcke ein (Mullis et al., 2009).

Initiiert wurde die internationale Untersuchung durch das IEA. Die Beteiligung von Deutschland im Jahr 2011 wurde durch die Kultusministerkonferenz (KMK) beschlossen. Grundlage dieses Beschlusses war eine Vereinbarung zwischen der KMK und dem Bundesministerium für Bildung und Forschung (BMBF). In Deutschland wird TIMSS von dem Institut für Schulentwicklungsforschung (IFS) in Dortmund unter der wissenschaftlichen Leitung von Wilfried Bos und unter Mithilfe des IEA Data Processing and Research Center (DPC) durchgeführt (Bos et al., 2012).

Ländervergleich

Die TIMSS-Untersuchung im Jahre 1995, in der die Kompetenzen der deutschen Schülerinnen und Schüler in der Mathematik und den Naturwissenschaften nicht in der Leistungsspitze lagen und große Disparitäten innerhalb des deutschen Schulsystems im Vergleich zu anderen Ländern festgestellt wurden, führte dazu, dass die KMK im Jahr 2002 beschlossen hat, landesweit gültige Standards für die Kernfächer festzulegen. In der Expertise „Zur Entwicklung nationaler Bildungsstandards“ von Klieme et al. (2003), die auf Anregung des Bundesministeriums für Bildung und Forschung (BMBF) entstanden war, wurden die Grundlagen für die Einführung sowie die Bedingungen einer erfolgreichen Implementation der Bildungsstandards diskutiert und festgelegt. Diese Expertise bildete bei der Erstellung und

Festlegung der Bildungsstandards durch die KMK eine wichtige Grundlage (Böhme, Richter, Stanat, Pant & Köller, 2012).

Ziel der Einführung von Bildungsstandards ist, den Schulen – trotz der Kulturhoheit der Bundesländer – eine deutschlandweit verbindliche Orientierung zu bieten und damit eine Voraussetzung zur Qualitätsentwicklung und -sicherung zu schaffen (KMK, 2005). Konkret werden Leistungsstandards bzw. zentrale Kompetenzen[3] beschrieben, die die Schülerinnen und Schüler am Ende einer bestimmten Jahrgangsstufe – getrennt nach Unterrichtsfächern und Schulformen – erreicht haben sollen (Regelstandards). Weg von der bisherigen Vorgabe von Inhalten in Lehrplänen hin zu erwarteten Lernergebnissen respektive Kompetenzen – damit wird dem ursprünglichen Problem des sogenannten ‚klassischen Vertrauens'[4] Rechnung getragen.

Bis heute wurden die folgenden, für die Bundesrepublik Deutschland geltenden, Bildungsstandards von der KMK für die Kernfächer entwickelt (Rabe, 2012): Für den Primarbereich existieren Regelstandards für die Fächer Deutsch und Mathematik, welche die Schülerinnen und Schüler am Ende der Jahrgangsstufe 4 erreicht haben sollen. Für den Hauptschulabschluss (Ende Jahrgangsstufe 9) wurden Bildungsstandards für die Fächer Deutsch, Mathematik und die erste Fremdsprache (Englisch/Französisch) festgelegt. Des Weiteren wurden Leistungsanforderungen in ebendiesen Fächern für den Mittleren Schulabschluss (Ende Jahrgangsstufe 10) sowie für das Ende der gymnasialen Oberstufe (allgemeine Hochschulreife) beschlossen. Darüber hinaus gibt es für den Mittleren Schulabschluss Regelstandards für die Fächer Biologie, Physik und Chemie.

Schon vor der Erarbeitung der beschriebenen Bildungsstandards legte die KMK darüber hinaus fest, dass eine landesweite Überprüfung erfolgen soll. Die in diesem Rahmen stattfindenden Testerhebungen an Schulen sollen dazu beitragen, die Qualität der Schulen und des Unterrichts sowie die Professionalität der Lehrenden zu verbessern (vgl. Klieme et al.,

[3] Kompetenzen werden in den Bildungsstandards entsprechend der Definition von Weinert (2002, S. 27) verstanden. Kompetenzen sind demnach die „bei Individuen verfügbaren oder durch sie erlernbaren kognitiven Fähigkeiten und Fertigkeiten, um bestimmte Probleme zu lösen, sowie die damit verbundenen motivationalen, volitionalen und sozialen Bereitschaften und Fähigkeiten um die Problemlösungen in variablen Situationen erfolgreich und verantwortungsvoll nutzen zu können".

[4] Im klassischen Vertrauen wurde davon ausgegangen, dass die Inhalte, die im Unterricht durchgenommen werden, auch tatsächlich von den Schülerinnen und Schülern gelernt werden (vgl. Merkens (2006).

2003). Eigens für die Präzisierung der Bildungsstandards, der Aufgabenentwicklung, der Normierung der Testinstrumente, der Organisation dieser Testverfahren und die Begleitung der Implementation der Bildungsstandards in den Schulen gründete die KMK im Jahre 2004 das Institut für Qualitätsentwicklung im Bildungswesen (IQB) in Berlin. Bei der Durchführung der Studien wird das IQB von dem IEA Data Processing Center in Hamburg unterstützt.

Neben landesweiten Vergleichsstudien (VERA), die jährlich die Leistungsfähigkeiten auf Ebene einzelner Schülerinnen und Schüler untersuchen, führt das IQB Ländervergleiche durch. Ziel ist die im Querschnitt angelegte Überprüfung der national geltenden Bildungsstandards im Bundesländervergleich (Köller, 2008). Diese Ländervergleiche lösen die nationalen Zusatzerhebungen ab, die z. B. im Rahmen von PISA und IGLU erfolgten (PISA-E, IGLU-E). Die hierfür entwickelten Tests orientieren sich nun an den national gültigen Bildungsstandards. Um weiterhin eine internationale Verknüpfung zu gewährleisten, finden die Ländervergleiche in Anbindung an große internationale Schulleistungsstudien statt (vgl. Abbildung 4.2). Der Ländervergleich für die Sekundarstufe (Klassenstufe 9) erfolgt in zeitlicher Abstimmung mit der PISA-Studie, erstmals im Jahre 2009 für die Sprachen und 2012 für die Mathematik und die Naturwissenschaften und dann jeweils alle sechs Jahre. Die Untersuchung im Primarbereich für Deutsch und Mathematik findet zusammen mit IGLU – und damit im Jahre 2011 auch in Verbindung mit TIMSS – in einem Fünfjahresrhythmus statt (Pöhlmann, Neumann, Tesch & Köller, 2010).

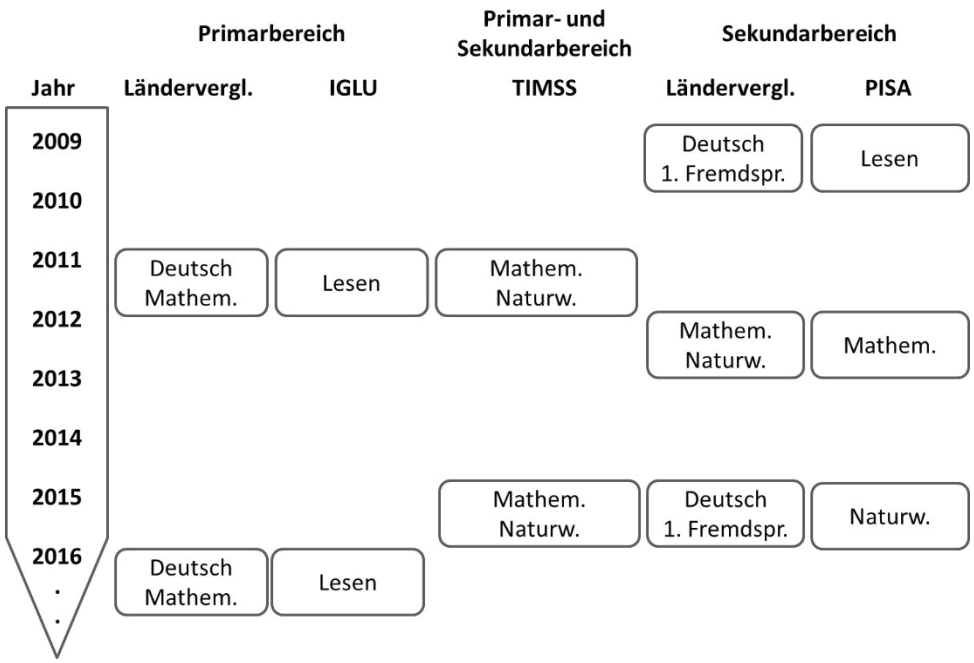


Abbildung 4.2: Überblick über die Testzeitpunkte ausgewählter Schulleistungsstudien

Für die Überprüfung der Bildungsstandards mussten zunächst Kompetenztests entwickelt werden. Das mehrstufige Verfahren der Aufgabenentwicklung für die Ländervergleichs-Tests wurde durch das IQB koordiniert. Vor der Gründung des IQB wurde z. B. das PISA-Konsortium mit der Aufgabenentwicklung für das Fach Mathematik in der Sekundarstufe betraut. An der Entwicklung der Aufgaben waren sowohl Lehrkräfte aus allen Bundesländern als auch nationale und internationale Experten aus den Fachdidaktiken und aus der Bildungswissenschaft beteiligt. Sieben Arbeitsschritte wurden bei der Testerstellung durchgeführt (vgl. u. a. Granzer, 2009; IQB - Institut zur Qualitätsentwicklung im Bildungswesen, 2007): (1) Konkretisierung und Operationalisierung der Kompetenzen, die in den Bildungsstandards definiert sind, (2) Erarbeitung von Richtlinien zur Konstruktion von Testaufgaben, (3) Entwicklung von Aufgaben durch Lehrkräfte, (4) Optimierung der Aufgaben durch einen Austausch mit Experten (u. a. in Bezug auf die Effektivität der nicht korrekten Antwortmöglichkeiten und die Passung in das zu messende Konstrukt), (5) Pilotierung der Aufgaben (Überprüfung in Bezug auf psychometrische Anforderungen), (6) Normierung des Tests (Ziel: national gültige Niveaustufen festlegen, die Auskunft darüber geben, ob die in den Bildungsstandards formulierten Kompetenzen von den Schülerinnen und Schülern erreicht wurden), (7) Bereitstellung dieser Kompetenzmodelle (erlaubt Aussagen darüber, wie viele Schülerinnen und Schüler die erwarteten Kompetenzen (Standards) erreicht haben). Nach diesen Schritten erfolgte eine Entwicklung von Sprach-, Mathematik- und Naturwissenschaftstests für den Primarbereich und für die Sekundarstufe I. Die Mathematiktests beispielsweise bilden – wie von den Bildungsstandards vorgegeben – sowohl die inhaltsbezogenen, die allgemeinen Kompetenzen als auch die drei Anforderungsbereiche ab (siehe Kapitel 4.4.1).

Im Jahr 2011 fand der Ländervergleich erstmals in der Primarstufe statt. Zielgruppe war die vierte Jahrgangsstufe deutscher Grundschulen, die in Bezug auf ihre mathematischen und sprachlichen Fähigkeiten untersucht wurden. Neben den Kompetenztests wurden Fragebögen eingesetzt, um Hintergrundinformationen zu erfassen. Die Schülerfragebögen erfassen u. a. die soziale Herkunft, den Migrationshintergrund sowie Informationen zu der Schule und dem Unterricht der Schülerinnen und Schüler. Die Lehrerinnen und Lehrer der teilnehmenden Klassen machen Angaben zu ihrem Unterricht (z. B. zu ihrem Umgang mit den Schülerinnen und Schülern), ihrer Schule (z. B. über die Schulausstattung und zum Kooperationsverhalten) sowie zu ihrer Person (z. B. ihr Geschlecht und zu ihrer Ausbildung) (Richter et al., 2012).

Die Verantwortung für die Ländervergleichsstudien trägt das IQB. Das DPC wurde mit der Organisation und Durchführung betraut und übernimmt die anschließende Kodierung der Daten und bereitet diese für die folgenden Analysen auf. Die Auswertung findet wiederum am IQB statt.

NEPS

Das Nationale Bildungspanel (*National Educational Panel Study*, kurz NEPS) ist eine Studie, die im Längsschnitt den Bildungsprozess sowie den Bildungsverlauf und die Kompetenzentwicklung über den gesamten Lebensverlauf in der Bundesrepublik Deutschland beschreiben soll. Im Gegensatz zu vielen anderen Schulleistungsstudien, wie z. B. der PISA-Studie, welche eine Momentaufnahme (Querschnittserhebungen) abbilden sollen im Rahmen des Nationalen Bildungspanels Daten im Längsschnitt erhoben werden, die es ermöglichen, die Veränderungen der Kompetenzen und mögliche Einflussfaktoren für deren Entwicklung im Bildungsverlauf zu beschreiben, indem dieselben Personen mehrmals befragt bzw. getestet werden (Leuze, 2008). Um dieses Ziel zu erreichen und vor allem um möglichst schnell auswertbare Daten zu erhalten, damit Rückschlüsse gezogen werden können, bedient sich das Nationale Bildungspanel der Methode des Multi-Kohorten-Sequenz-Designs (Blossfeld, Maurice et al., 2011). Dies bedeutet im Falle des NEPS, dass sechs getrennte Starthauptstichproben gezogen werden (vgl. Abbildung 4.3). Die Stichprobenziehung orientiert sich in erster Linie an dem Schuleintritt, den Übergängen im Bildungssystem und an dem Arbeitsmarkteintritt, umfasst jedoch noch weitere Bildungsprozesse im Lebensverlauf, so dass sich insgesamt acht Bildungsabschnitte bzw. sogenannte Etappen ergeben: Etappe 1 ‚Neugeborene und frühkindliche Betreuung', Etappe 2 ‚Kindergarten und Übergang in die Grundschule', Etappe 3 ‚Grundschule und Übergang in die Sekundarstufe I', Etappe 4 ‚Sekundarstufe I und Übergang in Sekundarstufe II bzw. Arbeitsmarkteintritt', Etappe 5 ‚Gymnasium und Übergang in Studium bzw. Berufsausbildung', Etappe 6 ‚Berufsbildung und Arbeitsmarkteintritt', Etappe 7 ‚(Fach-) Hochschulen und Arbeitsmarkteintritt' und Etappe 8 ‚Berufliche Weiterbildung und lebenslanges Lernen' (siehe Abbildung 4.4). Damit historische

Veränderungen aufgezeigt werden können, ist ergänzend eine Kohortensukzession[5] geplant (Blossfeld, Doll & Schneider, 2009).

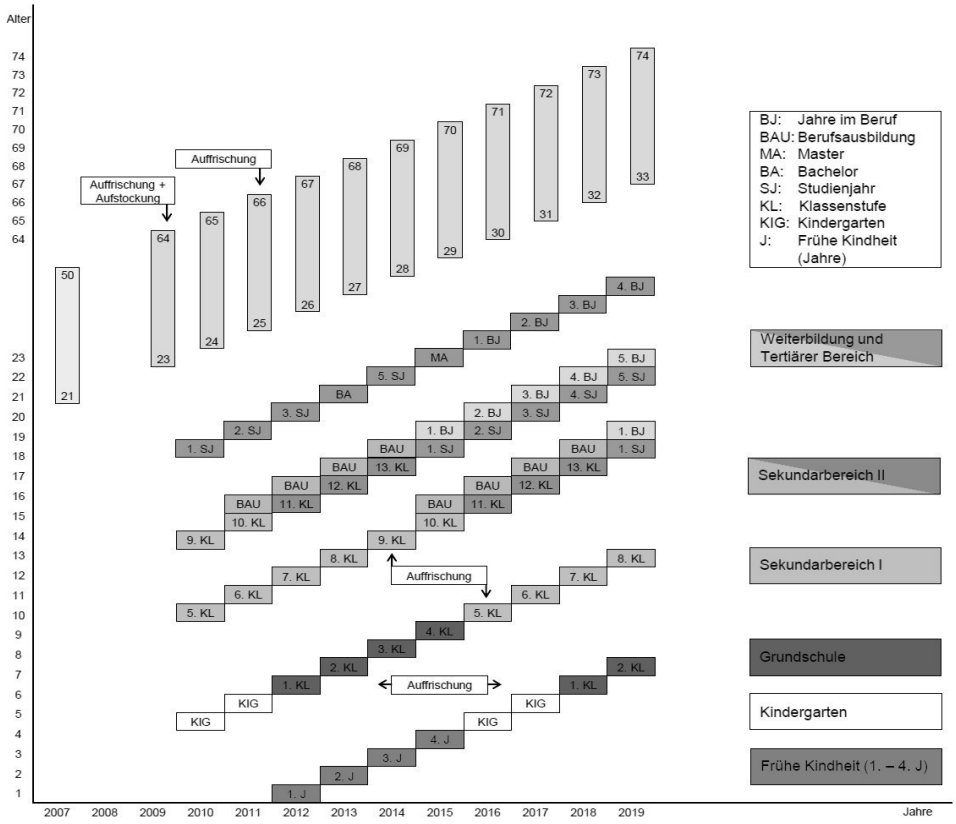


Abbildung 4.3: Sequenzdiagramm

Für die Messung der Kompetenzentwicklung wurden bzw. werden daher Kompetenztests für alle Klassenstufen, sowie für den vorschulischen Bereich (Neugeborene und Kindergarten) und den Tertiären- und Weiterbildungsbereich entwickelt. Da es nur für die institutionelle Ausbildung, und auch hier nicht für jede Klassen- bzw. Altersstufe separate Curricula gibt und eine Anbindung an internationale Schulleistungsstudien erfolgen soll, wird ein ganzheitlicheres Konzept angestrebt (vgl. u. a. Weinert et al., 2011). So orientieren sich z. B. die Mathematikleistungstests also nicht nur an den nationalen Curricula, sondern auch an dem, durch die PISA-Studie bekannt gewordenen, *Literacy Konzept*. Das mathematische Literacy-Konzept (mathematical literacy) von PISA distanziert sich bewusst von der Ausrichtung der Tests an Curricula und legt den Schwerpunkt auf die Erfassung der

[5] Unter Kohortensukzession wird verstanden, dass neue Startstichproben – in ähnlicher Struktur wie zu Beginn der Studie – gezogen werden.

Grundbildung, im Sinne von Kompetenzen, die die funktionale Anwendung der Mathematik betreffen:

> „Mathematical literacy is an individual's capacity to identify and understand the role that mathematics plays in the world, to make well-founded judgments and to use and engage with mathematics in ways that meet the needs of that individual's life as a constructive, concerned and reflective citizen." (OECD, 2004, S. 24). Unter mathematischer Kompetenz wird hier vor allem verstanden, dass die Schülerinnen und Schüler ihr Wissen in altersspezifischen authentischen Problemsituationen im Alltag erfolgreich anwenden können.

Das übergeordnete Ziel des Nationalen Bildungspanels, die Entwicklungsprozesse in der Bildung über den gesamten Lebensverlauf hinweg zu analysieren, wird in die folgenden fünf Forschungsinteressen bzw. Säulen gegliedert (vgl. Abbildung 4.4), die sich unterschiedlicher Methoden wie z. B. Fragebögen oder Kompetenztests bedienen (u. a. Blossfeld, Doll et al., 2009; Blossfeld, Schneider & Doll, 2009): *Säule 1 ‚Kompetenzentwicklung im Lebenslauf':* In dieser Säule wird das Ziel verfolgt, die Kompetenzen der Probandinnen und Probanden sowie ihre Entwicklung über den Lebensverlauf zu untersuchen sowie mögliche Voraussetzungen und Bedingungen zu analysieren, die die Entwicklung beeinflussen können. *Säule 2 ‚Bildungsprozesse in lebenslaufspezifischen Lernumwelten':* Ziel der zweiten Säule ist zu untersuchen, welche Bedingungen in den formalen, nicht-formalen und informellen Lernumwelten (in Anlehnung an Fend, 2009) vorliegen, wie relevant diese für den Kompetenzerwerb bzw. die Kompetenzentwicklung sind und ob die Bildungsentscheidungen sowie die Bildungsrenditen von den drei Lernumwelten beeinflusst werden. *Säule 3 ‚Soziale Ungleichheit und Bildungsentscheidungen im Lebenslauf':* Die dritte Säule befasst sich mit den Gründen und den Ursachen von sozialer Ungleichheit speziell in Bezug auf die hieraus resultierenden variierenden Bildungsentscheidungen, zu denen u. a. die Schul-, Studien- und Berufswahl sowie die Teilnahme an Weiterbildungsmaßnahmen gezählt werden. *Säule 4 ‚Bildungserwerb von Personen mit Migrationshintergrund im Lebenslauf':* Diese Säule befasst sich gesondert mit den Auswirkungen auf die Kompetenzen der Probandinnen und Probanden, die durch migrationsspezifische Merkmale zu erklären sind. Insbesondere werden hier die Spätaussiedler und die Familien mit türkischer Herkunft berücksichtigt. *Säule 5 ‚Bildungsrenditen im Lebenslauf':* Die 5. Säule erfasst die Bildungsrenditen, also die Erträge,

die aus einer Ausbildung bzw. einer Bildungsinvestition erzielt werden können (vgl. u. a. Anger, Plünnecke & Schmidt, 2010; Bauer & Jacob, 2010).

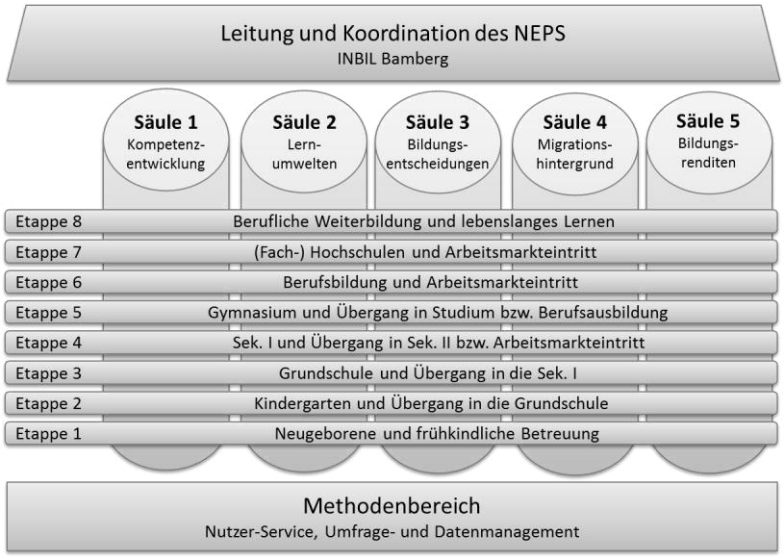


Abbildung 4.4: Rahmenkonzeption des NEPS (in Anlehnung an u.a. Blossfeld, Maurice et al., 2011, S. 12)

Das Nationale Bildungspanel wurde von dem Bildungsministerium für Bildung und Forschung (BMBF) initiiert und wird von diesem finanziert. Das zur Organisation der Studie gegründete Institut für bildungswissenschaftliche Längsschnittforschung Bamberg (INBIL) übernimmt die Koordination und Umsetzung des NEPS (Blossfeld, Roßbach et al., 2011). An der Durchführung sind ferner deutschlandweit mehrere Expertinnen und Experten verschiedener Institute und Forschungsbereiche beteiligt – u. a. aus der Erziehungswissenschaft, der Bildungsökonomie und der Familienforschung, um nur eine kleine Auswahl zu nennen – so dass der interdisziplinäre Ansatz viele Möglichkeiten zur Auswertung und Interpretation der Daten erlaubt. Darüber hinaus werden die Daten als anonymisierte Scientific Use Files der nationalen und internationalen Forschungsgemeinschaft zur Verfügung gestellt. Damit können an das Nationale Bildungspanel mit seinen Hauptzielen viele zusätzliche Studien angegliedert werden.

Der im Rahmen dieser Arbeit verwendete Mathematikleistungstest wurde vom Nationalen Bildungspanel für die fünfte Klassenstufe entwickelt. Die Haupterhebung fand 2010 statt und untersuchte die Leistungen der Schülerinnen und Schüler zu Beginn des fünften Jahrgangs. Zusätzlich wurden im Rahmen der dritten Startkohorte Kontextbedingungen der

Schülerinnen und Schüler erfasst, in dem sowohl die Eltern (via CATI[6]), die Lehrerinnen bzw. Lehrer, die in der getesteten Klasse Deutsch-, Mathematik- und Klassenlehrkräfte sind, als auch die Schulleitung befragt wurden. Ebenso fand eine Befragung der Schülerinnen und Schüler statt (Blossfeld, Maurice et al., 2011). Die Datenerhebung übernahm das DPC mit Unterstützung des infas (Institut für angewandte Sozialwissenschaft) in Bonn, die für die CATI der Eltern zuständig waren.

Curriculare Validität der NEPS-Mathematikaufgaben

In einem Expertenreview wurde die curriculare Validität der NEPS-Aufgaben für die vierte Jahrgangsstufe von drei Experten eingeschätzt (vgl. Kapitel 3.6.1.1). Grund hierfür war, dass der NEPS-Test ursprünglich für die fünfte Jahrgangsstufe konzipiert wurde und daher überprüft werden sollte, ob Viertklässlerinnen und Viertklässler laut Lehrplan die dort gestellten Aufgaben überhaupt schon bearbeiten können. Die curriculare Validität wurde mit fünf Fragen erfasst (vgl. Kapitel 3.6.1.1). Die Experten schätzen ein, dass der Inhalt von 20 Aufgaben laut Lehrplan in den ersten vier Jahrgangsstufen behandelt wird (PÜ = 75%, K = .123). Der Inhalt von vier Aufgaben hingegen wird laut Lehrplan erst nach der vierten Jahrgangsstufe behandelt, wobei drei Aufgaben denselben Aufgabenstamm haben. Bei diesen Aufgaben schätzen zwei von drei Experten die Aufgaben als curricular nicht valide ein. Dies entspricht einer prozentualen Anzahl von etwa 17 % des gesamten NEPS-Mathematiktests und ist damit in etwa vergleichbar mit dem TIMSS-Mathematiktest, bei dem etwa 21 % der Aufgaben als nicht valide eingestuft wurden. Es ist anzunehmen, dass der NEPS-Test und der TIMSS-Test daher für die Schülerinnen und Schüler etwas schwieriger ist als der LV-Test, da dieser nur curricular valide Aufgaben bezogen auf die Bildungsstandards nutzt.

Zudem wurde mit einer vierstufigen Skala von überhaupt nicht vertraut bis sehr vertraut erfasst, wie vertraut die Schülerinnen und Schüler nach Meinung der Experten mit der Aufgabenstellung, den graphischen Darstellungsformen und den mathematischen Begriffen in den NEPS-Mathematikaufgaben sind (PÜ = 45.8%, ICC = .69). Die Art der Aufgabenstellung ist den Schülerinnen und Schülern nach Expertenmeinung in acht Aufgaben eher nicht vertraut. Auch hier gehören wieder drei Aufgaben zu einem Aufgabenstamm. 16 Aufgaben sind den Schülerinnen und Schülern hingegen eher vertraut oder sehr vertraut. Graphische

[6] Computergestützte Telefoninterviews

Darstellungsformen kommen in insgesamt elf Aufgaben vor. In fünf dieser Aufgaben schätzen die Experten die Darstellungsformen als eher nicht vertraut ein. Hierzu zählt auch wieder der Aufgabenstamm mit drei Aufgaben. In zwölf Aufgaben sind mathematische Begriffe vorhanden. Die mathematischen Begriffe werden in allen Aufgaben als eher vertraut oder als sehr vertraut eingestuft. Weiterhin wurden die Experten gebeten, die Relevanz des Aufgabeninhalts einzuschätzen. Die Auswertung ergab, dass die Experten die Relevanz von nur zwei Aufgaben als eher gering einstuften (PÜ = 47.2, ICC = .59). Neben den Einschätzungen der Experten zur curricularen Validität des NEPS-Tests wurden die mathematischen Inhalte, die der NEPS-Test erfasst, eingeschätzt. Die Ergebnisse sind dem Kapitel 4.4.2 zu entnehmen.

Zwischenfazit

Ziel dieses Kapitel war, die Studien hinsichtlich ihrer Anlage und Ziele zu vergleichen und mögliche Unterschiede bzw. Überschneidungen darzustellen. Eine zusammenfassende Übersicht über die Anlage und die Ziele der drei Studien wird in Tabelle 4.1 gegeben. Den Studien gemein ist, dass sie u. a. die mathematischen Kompetenzen von Schülerinnen und Schülern überprüfen bzw. messen wollen. Weiterhin werden bei allen drei Studien Hintergrunddaten erhoben, die in die Analysen mit einfließen und beispielsweise Unterschiede zwischen Jungen und Mädchen aufzeigen können. Ziel der Studien ist damit u. a. das Bildungsmonitoring. Die drei Schulleistungsstudien erlauben eine objektive Leistungserfassung, da die Daten standardisiert beispielsweise extern durch das DPC erhoben und von geschultem Personal aufbereitet und codiert werden. Unterschiede zwischen den Studien bestehen hinsichtlich der Referenzrahmen. TIMSS und Ländervergleich bieten sowohl eine kriteriale Verortung der Ergebnisse über Kompetenzstufen an als auch einen sozialen Vergleich (z. B. Staaten- bzw. Ländervergleiche oder Vergleiche zwischen Jungen und Mädchen). NEPS hingegen bietet bislang keine kriteriale Verortung der Ergebnisse, jedoch aber soziale Vergleiche sowie ein Vergleich von Kompetenzen einer Gruppe bzw. Person mit den Kompetenzen der gleichen Probanden zu einem späteren Zeitpunkt (ipsativ).

TIMSS erlaubt eine Interpretation der Ergebnisse im internationalen Vergleich, d. h. es kann beispielsweise eine Verortung der mathematischen Kompetenz von Viertklässlern in Deutschland in einem internationalen Referenzmaßstab erfolgen. Für einen nationalen Vergleich auf Bundesländerebene reicht die Stichprobengröße jedoch nicht aus. Hier greift der Ländervergleich ein, der eine Gegenüberstellung bzw. einen Vergleich auf Bundeslandebene

Tabelle 4.1: Vergleich der Anlagen und der Ziele der drei Studien

	TIMSS 2011	**Ländervergleich 2011**	**NEPS 2010**
Messintention	- Bildungskontexte erfassen - Ertrag des Bildungssystems erfassen - Möglichkeit des gegenseitigen Lernens schaffen	- Überprüfung des Erreichens der nationalen Bildungsstandards	- Kompetenzentwicklung aufzeigen - mögliche Einflussfaktoren aufzeigen
mögliche Schlüsse			
-Referenzrahmen	sozial und kriterial	sozial und kriterial	sozial und ipsativ
-national/international	internationale Vergleichsstudie	nationale Vergleichsstudie	nationale Panelstudie
Untersuchungsdesign	Querschnitt	Querschnitt	Längsschnitt
Kompetenzbereiche	- Mathematik - Naturwissenschaften	- Mathematik - Deutsch - u.a.	- Mathematik - Deutsch - u.a.
zugrundeliegende Testkonzeption	Curriculum-Modell	Überprüfung der Bildungsstandards	- Literacy-Konzept - Orientierung an den Curricula

zulässt. Beide Studien sind jedoch im Querschnitt angelegt und erlauben keinen direkten Vergleich der Ergebnisse über die Zeit. Dieses Desiderat wird von NEPS aufgegriffen. NEPS ist als Panelstudie angelegt und kann damit beispielsweise intra-individuelle Veränderungen im Bildungsverlauf aber auch Veränderungen über die Zeit aufzeigen.

Wird die Nähe zum deutschen Curriculum betrachtet, zeigen sich deutliche Unterschiede zwischen den Studien. Durch das Curriculum Modell der TIMSS-Studie ist eine Nähe zum Curriculum gewährleistet. Jedoch wurden bei der Erstellung der Rahmenkonzeption, also bei der Festlegung welche Kompetenzen von den Schülerinnen und Schülern in Mathematik und in den Naturwissenschaften erwartet werden, die Curricula aller Teilnahmeländer berücksichtigt. Die Rahmenkonzeption fasst nun die unterschiedlichen Interessen und Inhalte zusammen, bei denen eine größtmögliche Übereinstimmung aller Länder gegeben ist. Selter, Walther, Wessel und Wendt (2012) konnten zeigen, dass 21 % der Mathematikitems in TIMSS 2011 bezogen auf die Bildungsstandards in Mathematik für die Grundschule als curricular nicht valide einzustufen sind. Die Einstufung erfolgte von einer Expertengruppe von Fachdidaktikerinnen und Fachdidaktikern. Hingegen ist bei nationalen Studien eine deutlich höhere Nähe zum Curriculum zu erwarten, sofern sich die Studie an nationalen Vorgaben orientiert. Dies ist beim Ländervergleich Mathematik Primar der Fall. Der Ländervergleich soll das Erreichen der landesweit gültigen Bildungsstandards untersuchen. Da der Ländervergleichs-Test mit diesen Zielvorstellungen entwickelt wurde, ist die Nähe zum Curriculum gegeben. Der NEPS-Test hingegen orientiert sich sowohl an den nationalen Bildungsstandards als auch an der Rahmenkonzeption von PISA. Hier ist demnach auch eine Nähe zum Curriculum gewährleistet, jedoch beruht die Rahmenkonzeption auf dem Mathematical Literacy Konzept und strebt dadurch ein ganzheitlicheres Konzept an, d. h. Ziel ist die Erfassung der mathematischen Grundbildung. Das eingesetzte Testinstrument vom Nationalen Bildungspanel – eigentlich für die fünfte Jahrgangsstufe entwickelt – weist 4 Aufgaben auf (17%), die nach Einschätzung der Experten Kompetenzen erfasst, die laut Curriculum erst nach der Grundschulzeit erworben werden sollen.

Zusammenfassend lässt sich festhalten, dass die drei Studien hinsichtlich ihrer Anlagen und Zielen grundlegend viele Gemeinsamkeiten aufweisen. NEPS hebt sich insofern von den beiden anderen Studien ab, dass Daten im Längsschnitt erfasst werden und keine kriteriale Interpretation – im Sinne von den Studien TIMSS und Ländervergleich – zulässt. Damit

verbunden liegt auch eine etwas andere Fokussierung der Messintention zugrunde. Die unterschiedlichen Zielsetzungen der Studien können dazu führen, dass nicht genau das gleiche Konstrukt mathematischer Kompetenz gemessen wird. Jedoch ist die Messung eines gleichen Konstrukts eine Voraussetzung für ein Equating und selbst wenn ein Linking angestrebt wird, so müssen sich hohe Ähnlichkeiten in den Konstrukten nachweisen lassen. Die Gemeinsamkeiten und Unterschiede hinsichtlich der gemessenen Konstrukte werden daher noch differenziert in Kapitel 4.4 analysiert. Zudem können sich auch die gefundenen Unterschiede in den zugrundeliegenden Konzeptionen auf die Definition der Konstrukte auswirken und hat darüber hinaus auch Einfluss auf die curriculare Validität der Tests. Die hieraus eventuell resultierenden Unterschiedlichkeiten in der Schwierigkeit der Tests sollten sich jedoch bei einem Linking ausgleichen lassen. Es handelt sich hierbei um die ursprüngliche Form des Equating, bei der unterschiedliche Schwierigkeiten in den Testheftvarianten ausgeglichen werden sollen. Insofern dürfte dies bei einem Linking keine Nachteile haben.

4.2 Stichprobe / Zielpopulation

Der zweite Aspekt, der hinsichtlich seines Maßes an Übereinstimmung untersucht werden soll, sind die Stichproben bzw. die Zielpopulationen der Studien (Forschungsfrage 1b). Dabei sollen die folgenden Fragen beantwortet werden: Nimmt die Studie Aussagen hinsichtlich der Population oder der Stichprobe vor? Wurde die Stichprobe alters- oder jahrgansbasiert gezogen? Sind die Stichproben in den Studien ähnlich oder gleich groß und wurden sie unter ähnlichen oder gleichen Bedingungen gezogen? Sind die Ausschlusskriterien der Studien vergleichbar? Im Folgenden werden zuerst die Studien getrennt voneinander hinsichtlich der oben genannten Fragestellungen dargestellt und anschließend in einem Zwischenfazit miteinander verglichen.

An **TIMSS** 2011 nahmen insgesamt 77 Bildungssysteme teil, 63 Länder und 14 Benchmarking-Teilnehmer (Mullis et al., 2012). In Deutschland haben sich alle 16 Bundesländer mit insgesamt 197 Grund- und Förderschulen an der Studie für Viertklässlerinnen und Viertklässler beteiligt. Das entspricht einer Teilnehmerzahl von 3.995 Schülerinnen und Schülern. Sowohl die Schulen als auch die Klassen wurden nach einem Zufallsverfahren – einem zweistufigen stratifizierten Clusterdesign – für eine Teilnahme

ausgewählt (Wendt et al., 2012). Zunächst wurden aus einer Liste mit allen Schulen in Deutschland 150 Schulen gezogen. Berücksichtigt wurde hierbei die Größe der Jahrgangsstufe 4 in den jeweiligen Schulen sowie weiterer Strata (z. B. Schultyp oder Bundesland). Als nächstes erfolgte die Auswahl der Klassen an den gezogenen Schulen. Hierbei hatte jede Klasse die gleiche Wahrscheinlichkeit in die Stichprobe mit aufgenommen zu werden. Es gab unterschiedliche Ausschlusskriterien, zum einen auf Schülerebene und zum anderen auf Schulebene. Auf Schülerebene wurden Schülerinnen und Schüler ausgeschlossen, die eine körperliche, geistige oder emotionale Beeinträchtigung nachweisen konnten oder die weniger als ein Jahr in Deutschland leben. Auf Schulebene wurden Schulen ausgeschlossen, die in einer schwer zugänglichen Region liegen oder deren Lehrplan oder Schulstruktur von dem nationalen Schulsystem abweicht (Wendt et al., 2012).

An dem **Ländervergleich** 2011 für den Primarbereich beteiligten sich insgesamt 1 349 Grund- und Förderschulen mit insgesamt 27 081 Schülerinnen und Schüler aus 16 Bundesländern. Die Stichprobenziehung erfolgte in mehreren Schritten und wird im Folgenden nach Richter et al. (2012) dargestellt. Grundlage bot die Schulstatistik für Drittklässlerinnen und Drittklässler des Schuljahrs 2009/2010. Zunächst wurde in jedem Bundesland eine Zufallsstichprobe gezogen. Die Anzahl der Schülerinnen und Schüler pro Bundesland orientierte sich dabei an der Leistungsstreuung vorheriger Vergleichsuntersuchungen. Anschließend wurden pro Flächenland 80 Schulen, in Berlin und Hamburg 120 Schulen und in Bremen 100 Schulen gezogen. Insgesamt wurde demnach eine Stichprobe von insgesamt 1380 Schulen angestrebt. In einem weiteren Schritt wurde pro Schule zufällig eine vierte Klasse für die Teilnahme gezogen. Die Teilnahme am Ländervergleich 2011 war für öffentliche Schulen verpflichtend, für Schulen in privater Trägerschaft freiwillig. Ausgeschlossen wurden insgesamt 171 Schülerinnen und Schüler, aufgrund vorher definierter Ausschlusskriterien, beispielsweise einer dauerhaften körperlichen oder geistigen Beeinträchtigung oder einer Aufenthaltsdauer von weniger als einem Jahr in Deutschland (Richter et al., 2012).

Das **Nationale Bildungspanel** führte die Untersuchung in der dritten Startkohorte, also der 5. Jahrgangsstufe, im Herbst und Winter des Jahres 2010 durch. Es wurde unterschieden zwischen Regelschulen und Förderschulen. Bei den Regelschulen handelte es sich um eine mehrstufig geschichtete Clusterstichprobe. Zunächst wurde auf die Stichprobe der neunten

Jahrgangsstufe zurückgegriffen und aus diesen wurden mittels größenproportionaler Zufallsauswahlen Schulen für die dritte Startkohorte ausgewählt. Zusätzlich wurden Schulen mit hinzugenommen, die nicht an der Erhebung der Startkohorte vier teilnehmen. In den Schulen wurden, wenn vorhanden, zwei fünfte Klassen zufällig ausgewählt. Die Teilnahme an der Studie war freigestellt. Insgesamt wurden für die Analysen der Daten die Ergebnisse von 5 193 Schülerinnen und Schüler an Regelschulen hinzugezogen (Duchhardt & Gerdes, 2012a).

Zwischenfazit

Ziel dieses Abschnitts war, die Stichproben bzw. den Stichprobenziehungsprozess in den drei Studien aufzuzeigen. Eine zusammenfassende Übersicht liefert die Tabelle 4.2. Alle Studien testen jahrgangsbasiert die Kompetenzen der Schülerinnen und Schüler. Der NEPS-Test konzentriert sich auf die fünfte Jahrgangsstufe, TIMSS und Ländervergleich testen hingegen die Kompetenzen von Viertklässlerinnen und Viertklässlern. Die Stichprobenziehung erfolgt in den Studien jeweils mehrschrittig. Dabei werden jeweils unterschiedliche Strata berücksichtigt (wie z. B. Bundesländer) und es werden Clusterstichproben gezogen, d. h. komplette Klassen, die an den Studien teilnehmen. Die Teilnahme an TIMSS und Ländervergleich ist verpflichtend, wohingegen die Teilnahme an NEPS freiwillig ist, was zu einer deutlichen Verzerrung innerhalb der Gegenüberstellung von Ergebnissen führen kann. Die Anzahl der teilnehmenden Schulen und Schülerinnen und Schülern unterscheidet sich ebenfalls zwischen den Studien, vor allem bedingt durch die unterschiedlichen Zielsetzungen der Studien. Der Ländervergleich hat mit 27 081 Schülerinnen und Schüler die größte Stichprobe. Die Anzahl der teilnehmenden Kinder ist hier höher, damit Bundesländervergleiche durchgeführt werden können. Dafür muss eine ausreichend große Anzahl an teilnehmenden Kindern pro Bundesland gewährleistet sein. TIMSS und NEPS nehmen Vergleiche solcher Art nicht vor, dennoch werden auch in diesen Studien alle Bundesländer mit einbezogen. Unterschiede bezüglich der Stichprobenmerkmale zwischen den Studien, auch bezüglich der Stichprobenmerkmale der Linking-Studie und den Hauptuntersuchungen, können sich insofern auf das Linking und auf einen möglichen Vergleich der Ergebnisse auswirken, als dass die Ergebnisse durch die Stichprobenunterschiede beeinflusst sein können. Dies bedeutet, dass bei einer anderen Stichprobenzusammensetzung eine andere Linkingfunktion entstehen könnte und sich die Linking-Ergebnisse nicht ohne weiteres auf andere Stichproben übertragen lässt. Daher ist es

wichtig zu überprüfen, ob hohe Korrelationen zwischen den Testwertergebnissen in den Studien gefunden werden können (vgl. Kapitel 5). Ist dies der Fall lassen sich trotz der Unterschiede vorsichtige Aussagen hinsichtlich deskriptiver Kennwerte treffen. Auf Individualebene sollten hingegen keine Aussagen getroffen werden. Weitere Hinweise hinsichtlich der Exaktheit und Stabilität der Linkingergebnisse liefern hier zusätzlich die Analysen auf Subgruppenebene (vgl. Kapitel 6).

Tabelle 4.2: Vergleich der Stichproben und der Zielpopulation der drei Studien

	TIMSS 2011	**Ländervergleich 2011**	**NEPS 2010**
Population/Stichprobe	Stichprobe	Stichprobe	Stichprobe
Alters-/Jahrgangsbasiert	jahrgangsbasiert	jahrgangsbasiert	jahrgangsbasiert
Stichprobenziehung	zweistufige stratifizierte Clusterstichprobe mit verpflichtender Teilnahme	mehrstufige stratifizierte Clusterstichprobe mit verpflichtender Teilnahme	mehrstufige stratifizierte Clusterstichprobe mit freiwilliger Teilnahme
Größe der Stichprobe	in Deutschland: 3 995 Schülerinnen und Schüler	27 081 Schülerinnen und Schüler	5 193 Schülerinnen und Schüler
Ausschlusskriterien	Schülerebene: - körperliche, geistige oder emotionale Beeinträchtigung - Aufenthalt in Deutschland seit weniger als einem Jahr Schulebene: - schwer zugängliche Regionen - Lehrplan oder Struktur der Schule weicht vom nationalen Schulsystem ab	- körperliche oder geistige Beeinträchtigung - Aufenthalt in Deutschland seit weniger als einem Jahr	- freiwillige Teilnahme - keine weiteren Ausschlusskriterien

4.3 Messbedingungen

Hinsichtlich der Messbedingungen (Forschungsfrage 1c) gilt es zunächst für die drei Studien aufzuzeigen, zu welchem Zeitpunkt die Messungen stattfanden, wie lange die Schülerinnen und Schüler getestet wurden, wie das Testheftdesign gestaltet ist und ob Hilfsmittel erlaubt waren oder nicht. In einem anschließenden Zwischenfazit werden die drei Studien hinsichtlich der genannten Aspekte miteinander verglichen.

TIMSS wurde im Frühsommer 2011 durchgeführt. Das Testheftdesign von TIMSS wurde bereits in Kapitel 3.5.3 vorgestellt. Zusammenfassend lässt sich festhalten, dass 14 Testhefte eingesetzt wurden, die sowohl die mathematische als auch die naturwissenschaftliche Kompetenz der Schülerinnen und Schüler messen sollten. Die Testhefte sind miteinander verlinkt, indem jeder der insgesamt 28 Blöcke in jeweils zwei Testheften an unterschiedlichen Stellen vorkommt. Neben sechs neuen Blöcken gingen jeweils acht unveränderte Blöcke aus TIMSS 2007 mit ein. Um Positionseffekte zu vermeiden, wurden die Mathematik- und Naturwissenschaftsblöcke abwechselnd an den Anfang gesetzt. Insgesamt besteht der TIMSS-Mathematiktest aus 177 Aufgaben.

Tabelle 4.3: TIMSS 2011, Ablauf der Testung

Ablauf	**Zeit**
Testheft, 1. Teil	36 Min.
Pause	
Testheft, 2. Teil	36 Min.
Pause	
Fragebogen für Schülerinnen und Schüler	30 Min.

Es wird davon ausgegangen, dass die Schülerinnen und Schüler im Durchschnitt 18 Minuten für die Lösung eines Blockes benötigen. Bei vier Blöcken pro Testheft bearbeiten die Schülerinnen und Schüler insgesamt 72 Minuten lang Aufgaben in Mathematik und den Naturwissenschaften (vgl. Tabelle 4.3). Die Testzeit wird nach 36 Minuten durch eine Pause unterbrochen. Nach dem zweiten Teil des Testhefts folgt eine weitere Pause, bevor die

Schülerinnen und Schüler einen Fragebogen ausfüllen. Für diesen sind 30 Minuten eingeplant (Mullis et al., 2009; Wendt et al., 2012).

Der **Ländervergleich 2011** hat von Mai bis Juli 2011 an zwei – meist aufeinanderfolgenden – Vormittagen in den Schulen stattgefunden. An einem Vormittag wurden die Kompetenztests in Deutsch durchgeführt, an dem anderen Vormittag in Mathematik. Der Ablauf der Testung in Regelschulen unterscheidet sich von dem Ablauf in Förderschulen. Die Regelschulen wurden jeden Tag 80 Minuten getestet mit einer 10-minütigen Pause nach 40 Minuten (vgl. Tabelle 4.4 und Richter et al., 2012). Die Testungen an den Förderschulen dauerten an beiden Tagen nur jeweils 40 Minuten, wobei hier kürzere und leichtere Testhefte eingesetzt wurden. Am ersten Testtag wurde nach den Kompetenztests noch eine Befragung der Schülerinnen und Schüler durchgeführt. Am zweiten Testtag bearbeiteten die Schülerinnen und Schüler überdies einen Test zu allgemeinen Grundfähigkeiten (KFT 4-12+K), ebenfalls im Anschluss an die Testung (Richter et al., 2012). Hierfür wurden die beiden Subtests Wortschatz und Figurenanalogien ausgewählt, die jeweils 10 Minuten in Anspruch nahmen. Zusätzlich wurde noch das Salzburger Lese-Screening (SLS 1-4) durchgeführt. Für die Bearbeitung hatten die Schülerinnen und Schüler insgesamt 3 Minuten Zeit.

Tabelle 4.4: Ländervergleich 2011, Ablauf der Testung

Ablauf erster Testtag an Regelschulen	**Zeit**
Testheft, 1. Teil	40 Min.
Pause	
Testheft, 2. Teil	40 Min.
Pause	
Fragebogen für Schülerinnen und Schüler	20 Min.

Für die Zusammenstellung der Testinstrumente wurde ein Multi-Matrix-Design verwendet, so dass nicht jede Schülerin und jeder Schüler alle Testaufgaben bearbeitete (Richter et al., 2012). Die Testaufgaben, die für den Ländervergleich 2011 ausgewählt wurden, wurden zunächst Blöcken zugeordnet. Diese umfassen Aufgaben, die jeweils einem Inhaltsbereich eines Faches entsprechen, beispielsweise dem Fach Mathematik und dem Inhaltsbereich *Raum und Form*. Die Blöcke – für deren Bearbeitung 10 Minuten festgelegt

wurden – wurden anschließend auf die Testhefte verteilt, wobei jeder Block mehrmals auftauchte, jedoch immer an unterschiedlichen Positionen, um mögliche Effekte (vgl. hierzu u. a. Robitzsch, 2009) kontrollieren zu können. Ein Testheft enthielt immer entweder Mathematik- oder Deutschaufgaben. In Mathematik wurde darauf geachtet, dass jedes Testheft drei bis fünf Inhaltsbereiche abdeckt. So entstanden insgesamt 35 Testhefte für Regelschulen und drei Testhefte für Förderschulen mit insgesamt 330 unterschiedlichen Mathematikaufgaben.

Die Haupterhebung der 3. Startkohorte im **NEPS** wurde im Herbst/Winter 2010 durchgeführt. Die Tests fanden in Gruppen an einem Tag statt. Es gab zwei unterschiedliche Testheftvarianten. Beide Testhefte begannen mit einem Test zur Lesegeschwindigkeit, der zwei Minuten in Anspruch nahm (vgl. Tabelle 4.5). Je nach Testheftvariante folgte der Test zur Lesekompetenz (28 Minuten) und den prozeduralen Metakognitionen (3 Minuten) oder zu der mathematischen Kompetenz (28 Minuten) mit dem dazugehörigen prozeduralen Metakognitionstest (3 Minuten). Im Anschluss daran erfolgte der jeweils andere Test. Danach folgen der Test zur kognitiven Grundfähigkeit (10.5 Minuten) und der Orthografie Test (25 Minuten) inklusive der prozeduralen Metakognition (2 Minuten). Insgesamt belief sich die reine Testzeit demnach auf 100.5 Minuten zuzüglich einer 15-minütigen Pause vor dem Test zur kognitiven Grundfähigkeit (Nationales Bildungspanel, 2011). Innerhalb der Mathematik- und Lesetests wurde kein Multi-Matrix-Design, sondern jeweils nur ein Testheft verwendet (Duchhardt & Gerdes, 2012a). Der Mathematiktest bestand aus insgesamt 25 Aufgaben.

Tabelle 4.5: NEPS Startkohorte 3 2010, Ablauf der Testung

Ablauf erster Testtag an Regelschulen	**Zeit**
Lesegeschwindigkeit	2 Min.
Lesekompetenz (inkl. prozedurale Metakognition)	31 Min
Mathematische Kompetenz (inkl. prozedurale Metakognition)	31 Min.
Pause	
Kognitive Grundfähigkeit	10.5 Min.
Orthografische Kompetenz (inkl. prozedurale Metakognition)	27 Min.

Zwischenfazit

In diesem Kapitel wurden die Durchführungsbedingungen in den drei Studien dargestellt. Ziel war zu untersuchen, in welchen Aspekten sich die Studien ähneln bzw. gleichen und wo Abweichungen bezüglich der Messbedingungen auftreten. Tabelle 4.6 gibt einen Überblick über die Ergebnisse. NEPS wird im Gegensatz zu den beiden anderen Studien in der fünften Jahrgangsstufe durchgeführt. Jedoch liegt kein vollständiges Schuljahr zwischen der NEPS-Studie und TIMSS bzw. dem Ländervergleich, da die NEPS-Erhebung im Herbst/Winter erfolgt und die Erhebungen für TIMSS und den Ländervergleich im Frühsommer. Die Schülerinnen und Schüler sind bei NEPS dementsprechend etwa ein halbes Schuljahr weiter und es ist anzunehmen, dass sie dadurch bedingt eine tendenziell höhere mathematische Kompetenz aufweisen. Für die Übertragung der Linking Ergebnisse aus der Linking-Studie auf die NEPS-Hauptuntersuchung ist daher zu berücksichtigen, dass die Ergebnisse nicht direkt miteinander vergleichbar sind, da die Schülerinnen und Schüler, die an der NEPS-Hauptuntersuchung teilgenommen haben ein halbes Schuljahr weiter sind als die Schülerinnen und Schüler, die an TIMSS bzw. am Ländervergleich teilgenommen haben. Es ist davon auszugehen, dass die Schülerinnen und Schüler, die an der NEPS-Hauptuntersuchung teilgenommen haben, also ein halbes Jahr weiter sind, etwa 15 Kompetenzpunkte mehr erreichen als die Schülerinnen und Schüler der TIMSS- und Ländervergleichsstudie. Die 15 Punkte entsprechen einem angenommenen Kompetenzzuwachs von 30 Punkten für ein Schuljahr (Klieme et al., 2010).

Es zeigt sich zudem, dass sowohl TIMSS als auch der Ländervergleich ein Multi-Matrix-Design verwenden. NEPS hingegen verwendet in Mathematik nur eine Testheft-Variante. Vorteil ist, dass von jeder Schülerin und jedem Schüler die gleichen Aufgaben aus allen Inhaltsbereichen bearbeitet werden. Damit kann nicht der Fall eintreten kann, dass eine Schülerin oder ein Schüler andere Aufgaben vorgelegt bekommt, die eventuell sogar unterschiedlichen Inhaltsbereichen zugeordnet werden. Dieses kann die Testleistungen der Schülerinnen und Schüler beeinflussen. Nehmen wir an, ein Schüler ist besonders gut in Geometrie, bekommt aber aus diesem Bereich keine Aufgaben, so kann er sein spezifisches Können nicht zeigen, was wiederum sein Testergebnis verfälschen kann. Insofern sollten keine Interpretationen auf Individualebene vorgenommen werden, deskriptive Kennwerte hinsichtlich Tendenzen können hingegen aufgezeigt werden.

Ein weiterer Unterschied zeigt sich in den aufgeführten Gesamt-Testzeiten sowie in den Zeiten, die für den Mathematiktest zur Verfügung stehen. Hierbei stellt sich die Frage, ob der NEPS-Test, der mit 31 Minuten die kürzeste Testzeit aufweist, die mathematische Kompetenz genauso facettenreich erfassen kann wie beispielsweise der Ländervergleichstest, der eine Testzeit von 80 Minuten für den Mathematiktest aufweist. Ergebnisse zu dieser und anderen Fragestellungen bezüglich des erfassten mathematischen Konstrukts werden noch detaillierter in Kapitel 4.4, in dem ein konzeptioneller Vergleich der drei Studien erfolgt, und in Kapitel 5, in dem empirische Zusammenhänge zwischen den Studien untersucht werden, aufgezeigt (vgl. auch Zwischenfazit in Kapitel 4.1).

Tabelle 4.6: Vergleich der Messbedingungen der drei Studien

	TIMSS 2011	**Ländervergleich 2011**	**NEPS 2010**
Messzeitpunkt	Frühsommer 2011	Frühsommer 2011	Herbst/Winter 2010
Testzeit	1 Testtag Gesamt: 72 Minuten Mathematik: 36 Minuten	2 Testtage Gesamt: 160 Minuten Mathematik: 80 Minuten (Ein Testtag)	1 Testtag Gesamt: 100.5 Minuten Mathematik: 31 Minuten
Testdesign	Multi-Matrix-Design, 14 Testheftvarianten verlinkt durch 28 Blöcke (14 Mathematik, 14 Naturwissenschaften), 4 Blöcke/Testheft	Multi-Matrix-Design, 35 Mathematik Testheftvarianten (Regelschulen) verlinkt durch 35 Aufgabenblöcke 8 Blöcke/Testheft	2 Testheftvarianten (Testheftvariante 1: Mathematik an erster Stelle und Lesen an zweiter Stelle und Testheftvariante 2: umgekehrte Reihenfolge), 1 Block für Mathematik und 1 Block für Lesen
Erlaubte Hilfsmittel	keine	keine	keine

4.4 Konzeptioneller Vergleich

Die drei Studien messen alle ein Konstrukt mathematischer Kompetenz. Fraglich ist jedoch, ob die drei Studien tatsächlich das Konstrukt auch in gleicher oder zumindest ähnlicher Weise definieren. Daher soll im Folgenden der Frage nachgegangen werden, inwieweit die Definition des Konstrukts mathematischer Kompetenz im NEPS vergleichbar ist mit den Messungen in TIMSS und im Ländervergleich (Forschungsfrage 1d). Zunächst soll ein Vergleich auf Basis der Rahmenkonzeptionen der drei Studien erfolgen. Die drei Studien unterscheiden jeweils unterschiedliche mathematische Teilbereiche, wie beispielsweise mathematische Inhaltsbereiche, kognitive Anforderungsbereiche und prozedurale Fähigkeiten. Zusätzlich wird das erste Expertenreview (vgl. Kapitel 3.6.1.1), in dem eine Klassifikation der NEPS-Items in die Rahmenkonzeptionen von TIMSS und dem Ländervergleich vorgenommen wurde, für die Beantwortung der Frage herangezogen, ob es sich bei dem Konstrukt mathematischer Kompetenz und den einzelnen Teilbereichen dieses Konstrukts um eine sogenannte jingle oder jangle fallacy handelt (vgl. Kapitel 2).

Neben Unterschieden bzw. Gemeinsamkeiten hinsichtlich mathematischer Aspekte soll in dem vorliegenden Kapitel ebenfalls ein Vergleich auf Aufgabenebene hinsichtlich unterschiedlicher Merkmale stattfinden. Hierzu werden die Ergebnisse des zweiten Expertenreviews herangezogen (vgl. Kapitel 3.6.1.2). Es soll der Frage nachgegangen werden, ob die NEPS-Mathematikaufgaben hinsichtlich formaler und sprachlicher Merkmale hinreichend ähnlich sind im Vergleich zu den Mathematikaufgaben aus TIMSS und dem Ländervergleich (Forschungsfrage 1d).

4.4.1 Vergleich der Rahmenkonzeption

Schulleistungsstudien definieren meist in ihren Rahmenkonzeptionen verschiedene Aspekte mathematischer Kompetenz. Häufig werden zwei bis drei Aspekte unterschieden: die inhaltsbezogenen Kompetenzen, die kognitiven Prozesse zum Lösen einer Mathematikaufgabe und die unterschiedlichen Anforderungsbereiche, also die Komplexität der zu erbringenden kognitiven Leistung. Jede Mathematikaufgabe wird in jedem dieser Bereiche einem oder – je nach Rahmenkonzeption – auch mehreren Kompetenzbereichen

zugeordnet. Die Bereiche sind als untrennbar miteinander verschmolzen zu betrachten, denn es können z. B. keine kognitiven Prozesse ohne Inhalte erworben werden (Roppelt & Reiss, 2012). Im Folgenden werden die Dimensionen, die in den Rahmenkonzeptionen der drei Studien unterschieden werden, beschrieben und einander gegenübergestellt.

4.4.1.1 Überblick

Die **TIMSS-**Rahmenkonzeption der Mathematiktests basiert auf einem Curriculum-Modell (siehe Kapitel 4.1) und wurde seit der ersten Durchführung im Jahr 1995 weiter überarbeitet. Damit konnten Erfahrungen von Expertinnen und Experten der letzten Jahre aufgegriffen und integriert werden (Mullis et al., 2009). Nach dem Curriculum-Modell stützt sich die TIMSS-Rahmenkonzeption auf das intendierte, das implementierte und das erreichte Curriculum, wobei alle Länder in die Erstellung mit einbezogen werden. Es wird demnach überprüft, ob die Schülerinnen und Schüler die Kompetenzen erreicht haben, die in den meisten Teilnahmeländern in den Lehrplänen bzw. Prüfungsvorschriften und in den Schulbüchern vorhanden sind und von Expertinnen und Experten als relevant angesehen werden (Mullis et al., 2009). In den Leistungstests wird unterschieden zwischen *content domains* (Inhaltsbereichen) und *cognitive domains* (kognitiven Anforderungsbereichen). Die Inhaltsbereiche werden im Framework getrennt nach der Jahrgangsstufe (4. oder 8. Jahrgangsstufe) und nach Fach (Mathematik oder Naturwissenschaften) aufgeführt und hier wiederum in mehrere Teilgebiete mit verschiedenen Themenbereichen aufgegliedert.

Tabelle 4.7: Mathematische Inhalts- und Anforderungsbereiche von TIMSS

Inhaltsbereiche	**Anforderungsbereiche**
Arithmetik	Reproduzieren
Geometrie/Messen	Anwenden
Daten	Probleme lösen

Der mathematische Leistungstest differenziert – wie in Tabelle 4.7 ersichtlich – in der vierten Jahrgangsstufe die drei grundlegenden Inhaltsbereiche *Number* (Arithmetik), *Geometric Shapes and Measures* (Geometrie/Messen) und *Data Display* (Daten). Die drei kognitiven Anforderungsbereiche *Knowing* (Reproduzieren)*, Applying* (Anwenden) und *Reasoning* (Probleme lösen) sind fächerübergreifend, werden jedoch für die Jahrgangsstufen und die

Fächer getrennt durch konkrete Verhaltensweisen in unterschiedlichen Niveaus beschrieben (Mullis et al., 2009; Wendt et al., 2012).

Der **Ländervergleich** Mathematik Primar orientiert sich an den nationalen Bildungsstandards, welche die gewünschten Lernergebnisse der Schülerinnen und Schüler am Ende der Grundschulzeit (vierte Klasse) festlegen. Die Grundschule verfolgt prinzipiell das Ziel der Entfaltung grundlegender Bildung, um damit die Basis für weiterführendes Lernen zu legen sowie die Fähigkeit zur selbständigen Kulturaneignung zu vertiefen (KMK, 2005). Die Bildungsstandards in Mathematik zielen darauf ab, dass die Schülerinnen und Schüler in der Primarstufe ein gesichertes Verständnis mathematischer Inhalte entwickeln, um für die weitere Schullaufbahn sowie das gesellschaftliche und berufliche Leben entsprechend gerüstet zu sein (KMK, 2005). Zur Erreichung dieses Ziels nehmen die Bildungsstandards Abstand zu der ursprünglichen, traditionellen Einteilung in die Sachgebiete Arithmetik, Geometrie, Größen und Sachrechnen und streben nun vielmehr – in Anlehnung an Winter (1995) - ein ganzheitlicheres Konzept der mathematischen Grundbildung an (vgl. Abbildung 4.5). Sie unterscheiden nun zwischen allgemeinen und inhaltsbezogenen mathematischen Kompetenzen sowie drei Anforderungsbereichen KMK, 2005.

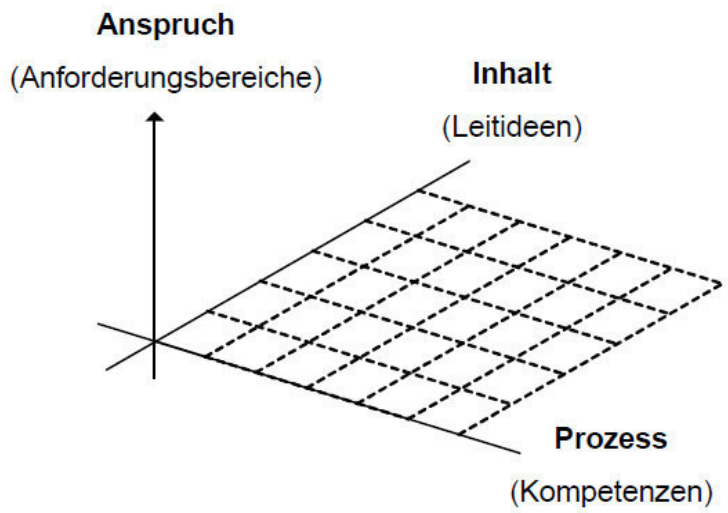


Abbildung 4.5: Kompetenzmodell der Bildungsstandards (KMK & IQB - Institut zur Qualitätsentwicklung im Bildungswesen, 2012, S. 3)

Bei den allgemeinen Kompetenzen wird zwischen *Darstellen*, *Modellieren*, *Argumentieren*, *Kommunizieren*, *Problemlösen* und *technischen Grundfertigkeiten* unterschieden. Bei den inhaltsbezogenen mathematischen Kompetenzen handelt es sich um

Zahlen und Operationen, *Größen und Messen*, *Muster und Strukturen*, *Raum und Form* sowie *Daten, Häufigkeit und Wahrscheinlichkeit*. Weiterhin werden drei Anforderungsbereiche differenziert: *Reproduzieren*, *Zusammenhänge herstellen* und *Verallgemeinern und reflektieren* (vgl. Tabelle 4.8).

Tabelle 4.8: Mathematische Inhaltsbereiche, Prozesse und Anforderungsbereiche im Ländervergleich

Inhaltsbereiche	**Prozesse**	**Anforderungsbereiche**
Zahlen und Operationen	Darstellen	Reproduzieren
Größen und Messen	Modellieren	Zusammenhänge herstellen
Muster und Strukturen	Argumentieren	Verallgemeinern und reflektieren
Raum und Form	Kommunizieren	
Daten, Häufigkeit und Wahrscheinlichkeit	Problemlösen	
	Technische Grundfertigkeiten	

NEPS untersucht sowohl kognitive als auch nicht-kognitive Kompetenzen über die Lebensspanne hinweg. Dabei wird zwischen vier Bereichen unterschieden: (A) die allgemeine kognitive Leistungsfähigkeit, (B) die domänenspezifischen Kompetenzen, (C) die Meta- sowie sozialen Kompetenzen sowie (D) Fähigkeiten, die je nach Lebensphase relevant sind (z. B. für den Beruf) (Weinert et al., 2011). Die mathematischen Kompetenzen zählen zu dem zweiten Aspekt, den domänenspezifischen Kompetenzen. In den jüngeren Alterskohorten wird die mathematische zunächst jährlich und dann alle zwei Jahre erhoben. Im Erwachsenenalter hingegen liegen sechs Jahre zwischen den Erhebungen (Ehmke et al., 2009).

Wie bereits beschrieben verfolgt NEPS ein ganzheitliches Konzept, das Literacy Konzept (vgl. Kapitel 4.1). Dies wird damit begründet, dass (a) NEPS die Kompetenzen im Längsschnitt erfasst, um die Notwendigkeit dieser Kompetenzen für die Zukunftsaussichten zu untersuchen, (b) dass das Literacy Konzept in der Politik, Wissenschaft und auch in der Öffentlichkeit immer mehr Zuspruch findet und (c) dass durch die Anpassung an die Konzeptionen von internationalen Schulleistungsstudien ein Linking ermöglicht wird (Weinert et al., 2011). Ausgehend von den Vorgaben des in der PISA-Studie definierten Literacy Konzept und den Vorgaben in der *National Council of Teachers of Mathematics framework conception*

(NCTM) wurde eine theoretische Rahmenkonzeption für die Erfassung der mathematischen Kompetenz entwickelt, der für alle Altersgruppen grundlegend ist. Ausgehend von diesen Verständnissen mathematischer Kompetenz wurde für NEPS eine Rahmenkonzeption strukturiert, die zwischen vier inhaltlichen (*Quantität, Veränderung und Beziehungen, Raum und Form und Daten und Zufall*) sowie sechs prozessbezogenen Kompetenzen (*Repräsentieren (Darstellungen verwenden), Modellieren, Mathematisch Argumentieren, Mathematisch Kommunizieren, Mathematische Probleme lösen* sowie *Technische Fertigkeiten einsetzen*) differenziert (vgl. Tabelle 4.9), die eng miteinander verzahnt sind. Jede Aufgabe des mathematischen Kompetenztests kann einem Inhaltsbereich zugeordnet werden und für die Lösung der Aufgabe benötigen die Schülerinnen und Schüler meist mehrere kognitive Prozesse. Die Ähnlichkeit zu anderen internationalen Schulleistungsstudien wie PISA und TIMSS wurde bewusst angestrebt, um zum einen Aufgaben aus bereits bestehenden Aufgabenpools nutzen zu können und zum anderen eine Voraussetzung für eine Verbindung zwischen den Kompetenzskalen von NEPS mit anderen Schulleistungsstudien zu schaffen (Ehmke et al., 2009).

Tabelle 4.9: Mathematische Inhaltsbereiche und Prozesse in NEPS

Inhaltsbereiche	**Prozesse**
Quantität	Repräsentieren (Darstellungen verwenden)
Veränderung und Beziehungen	Modellieren
Raum und Form	Mathematisch Argumentieren
Daten und Zufall	Mathematisch Kommunizieren
	Mathematische Probleme lösen
	Technische Fertigkeiten einsetzen

4.4.1.2 Inhaltsbezogene Kompetenzen

Die drei Schulleistungsstudien verfolgen das Ziel, die mathematische Kompetenz von Schülerinnen und Schülern in der Primarstufe in möglichst vielen Facetten zu erfassen. Somit decken sie auch ähnliche Inhaltsbereiche ab. Abweichungen können z. B. durch Anlehnung an unterschiedliche Curricula entstehen. Diese Unterschiede und Gemeinsamkeiten sollen im Verlauf dieser Analyse aufgezeigt werden.

Ein grundlegender Unterschied zwischen den Studien besteht zunächst in den unterschiedlichen Zusammensetzungen bzw. Bezeichnungen der inhaltsbezogenen Kategorien. TIMSS (Selter et al. 2012) unterscheidet drei Inhaltsbereiche: *Arithmetik*, *Geometrie/Messen* und *Daten*. Versucht man nun die Inhaltsbereiche des Ländervergleichs (Roppelt und Reiss 2012) und NEPS (Ehmke et al., 2009) diesen Kategorien zuzuordnen sowie auch diese beiden Studien miteinander in Beziehung zu setzen ergibt sich folgendes Bild[7]:

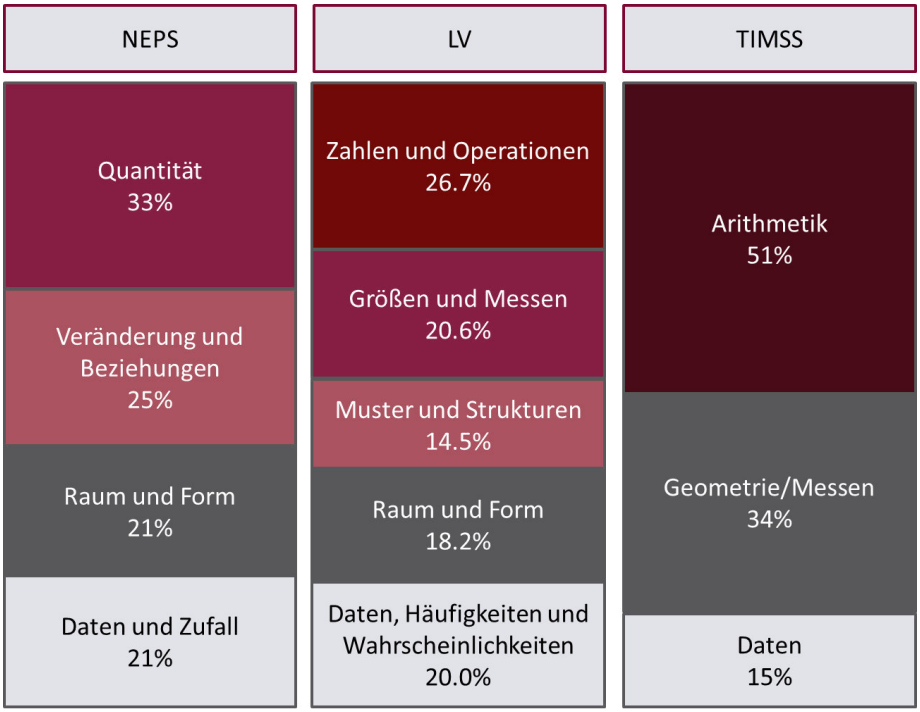


Abbildung 4.6: Inhaltsbezogene Komponenten von NEPS, dem Ländervergleich und TIMSS

Der Inhaltsbereich *Arithmetik* (TIMSS) umfasst mit wenigen Ausnahmen die inhaltsbezogenen Komponenten *Zahlen und Operationen*, *Größen und Messen* sowie *Muster und Strukturen*, welche im Ländervergleich formuliert werden. NEPS differenziert den Aspekt der *Arithmetik* (TIMSS) ebenfalls weiter aus und definiert die zwei Inhaltsbereiche *Quantität* und *Veränderung und Beziehungen*. Wird weiterhin NEPS dem Ländervergleich gegenübergestellt ergibt sich ein Zusammenhang zwischen dem Bereich *Quantität* (NEPS) und den beiden inhaltsbezogenen Komponenten *Zahlen und Operationen* sowie *Größen und Messen* im Ländervergleich. Beispielsweise beschreiben alle Studien die Kompetenz, die Grundrechenart zu beherrschen als relevant. Bei TIMSS zählt dieser Aspekt zu *Arithmetik*, beim Ländervergleich zu *Zahlen und Operationen* und bei NEPS zu *Quantität*. Die Inhalte von

[7] Ähnliche Inhaltsbereiche werden mit gleichen Farben dargestellt

Muster und Strukturen im Ländervergleich und *Veränderung und Beziehungen* bei NEPS stimmen ebenfalls annähernd überein. Als Beispiel lässt sich hier die Kompetenz ‚Muster zu erkennen und fortzusetzen' nennen. Diese Kompetenz fällt bei TIMSS unter *Arithmetik*, beim Ländervergleich unter *Muster und Strukturen* und bei NEPS unter *Veränderung und Beziehungen*.

Der zweite Inhaltsbereich, der bei TIMSS definiert wird, ist *Geometrie/Messen*. Dieser entspricht weitestgehend den Komponenten *Raum und Form* beim Ländervergleich wie auch bei NEPS. Beispielsweise lässt sich die Kompetenz geometrische Figuren zu erkennen und zu klassifizieren bei TIMSS dem Inhaltsbereich *Geometrie/Messen* zuordnen, bei dem Ländervergleich und NEPS würde diese Kompetenz zum Inhaltsbereich *Raum und Form* zählen.

Die dritte inhaltliche Komponente bei TIMSS ist *Daten*. Diesem gegenüber stehen die Inhaltsbereiche *Daten, Häufigkeit und Wahrscheinlichkeit* (Ländervergleich) und *Daten und Zufall* (NEPS). Unschwer lässt sich auch hier ein Zusammenhang erkennen. Alle drei Studien nennen z. B. die Kompetenz Daten zu ordnen und darzustellen als relevant. Diese wird in den Studien den oben genannten Inhaltsbereichen zugeordnet.

Neben den vielen begrifflichen und inhaltlichen Übereinstimmungen zwischen den drei Studien bezüglich der Inhaltsbereiche bleiben aber auch einige Unterschiede festzuhalten. Der Aspekt der Wahrscheinlichkeit bzw. des Zufalls wird bei den beiden nationalen Studien berücksichtigt, ist jedoch im Rahmen von TIMSS nicht definiert (Mullis et al., 2009). Grund hierfür könnte das Curriculum-Modell der TIMSS-Studie sein, welches u. a. die Lehrpläne mehrerer Länder berücksichtigt (siehe Kapitel 4.1). Wenn also das Thema Wahrscheinlichkeit bzw. Zufall in den meisten Teilnahmeländern von TIMSS bis zum Ende der Primarstufe nicht behandelt wird, erscheint es nicht sinnvoll diesen Inhalt in einem internationalen Leistungstest zu berücksichtigen. Da der Aspekt der Wahrscheinlichkeit bzw. des Zufalls jedoch in den meisten Lehrplänen, Curricula oder den Bildungsstandards in Deutschland definiert bzw. in den Schulbüchern oder Schulen behandelt werden, wird dieser in den nationalen Studien auch in den mathematischen Leistungstests erhoben. So heißt es z. B. in den Bildungsstandards, dass die Schülerinnen und Schüler die Kompetenz erlangen sollen „Wahrscheinlichkeiten von Ereignissen in Zufallsexperimenten vergleichen" zu können (Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik

Deutschland, 2005). Dies bezieht das Wissen von Grundbegriffen wie auch das Einschätzen von Gewinnchancen in einfachen Zufallsexperimenten mit ein.

Hingegen beinhaltet das Framework von TIMSS auch inhaltliche Aspekte, welche in den deutschen Rahmenlehrplänen nicht für den Primarbereich vorgesehen sind. Damit sind Aufgaben, die solche Inhaltsbereiche abdecken, als curricular nicht valide einzuordnen. In TIMSS wird z. B. vorausgesetzt, dass die Schülerinnen und Schüler eine elementare Vorstellung davon haben, was Brüche sind und dass sie mit ihnen rechnen können sowie dass sie mit Winkeln umgehen und diese vergleichen können. Die TIMSS-Aufgaben wurden von einer Expertengruppe (vgl. hierzu Selter et al., 2012) auf ihre curriculare Validität hinsichtlich der Bildungsstandards im Fach Mathematik für die Grundschule hin untersucht. Insgesamt wurden 21 % der TIMSS-Aufgaben von der Expertengruppe als nicht curricular valide eingeschätzt (vgl. hierzu auch Kapitel 4.1).

Eine weitere Unterscheidung ergibt sich bei der Zuordnung von Maßeinheiten zu den inhaltsbezogenen Komponenten der drei Studien. Im Ländervergleich zählen alle Maßeinheiten zu dem Inhaltsbereich *Größen und Messen*, bei NEPS zu dem Inhaltsbereich *Quantität*. Im TIMSS-Test hingegen werden die Maßeinheiten nicht nur dem Bereich *Arithmetik* zugeordnet. Zu *Arithmetik* zählen nur Größen wie Geldwerte und Zeitspannen. Größen wie Längen, Gewichte oder Raumeinheiten gehören beim TIMSS-Test zu dem Inhaltsbereich *Geometrie/Messen*.

Der Abbildung 4.6 kann weiterhin entnommen werden, dass die prozentuale Verteilung der Aufgaben auf die Inhaltsbereiche in den Studien unterschiedlich ist. Dies bedeutet, dass den Inhaltsbereichen in den drei Studien jeweils eine unterschiedliche Bedeutung beigemessen wird. Der TIMSS-Mathematiktest legt einen klaren Fokus auf den Inhaltsbereich *Arithmetik* (51 %). *Geometrie/Messen* sind nur 34 % der Aufgaben zuzuordnen. Am wenigsten Raum nimmt der Inhaltsbereich *Daten* ein (15 %). Der Ländervergleich hingegen misst dem Inhaltsbereich *Daten, Häufigkeit und Wahrscheinlichkeit* mit 20 % der Aufgaben eine höhere Bedeutung bei. NEPS ordnet sogar 21 % seiner Aufgaben diesem Bereich zu. Hingegen nehmen die Inhaltsbereiche *Raum und Form* in den Studien Ländervergleich (18.2 %) und NEPS (21 %) weniger Raum ein, als in der TIMSS-Studie. Ein weiterer Unterschied bei der prozentualen Verteilung zeigt sich beim Ländervergleich und beim NEPS beim Inhaltsbereich *Veränderung und Beziehungen* (NEPS) bzw. *Muster und Strukturen* (Ländervergleich). Der

NEPS-Test misst diesem Bereich mit 25 % eine deutlich höhere Bedeutung bei als der Ländervergleichs-Test (14.5 %). Hingegen werden im NEPS-Test weniger Aufgaben dem Inhaltsbereich *Quantität* zugeordnet (33 %) als im Ländervergleichs-Test den Inhaltsbereichen *Zahlen und Operationen* (26.7 %) und *Größen und Messen* (20.6 %).

4.4.1.3 Prozessbezogene Kompetenzen

Die prozessbezogenen Komponenten beim Ländervergleich (Roppelt und Reiss, 2012) und NEPS (Ehmke et al., 2009) werden nicht nur ähnlich bezeichnet, sie bilden auch die gleichen Inhalte ab. Dies liegt daran, dass sich NEPS bei der Definition der Prozess u. a. an dem Ländervergleich bzw. den Bildungsstandards orientiert hat (Ehmke et al., 2009). Damit ergibt sich das in Tabelle 4.10 dargestellte Bild, in dem die Prozesse des Ländervergleichs ihrem entsprechenden Pendant in NEPS gegenübergestellt werden.

Tabelle 4.10: Prozesse im Ländervergleich und NEPS

Ländervergleich Mathematik Primar	**NEPS Mathematik**
Darstellen	Repräsentieren (Darstellungen verwenden)
Modellieren	Modellieren
Argumentieren	Mathematisch Argumentieren
Kommunizieren	Mathematisch Kommunizieren
Problemlösen	Mathematische Probleme lösen
Technische Grundfertigkeiten	Technische Fertigkeiten einsetzen

Wie bereits kurz angerissen, definiert TIMSS (Selter et al., 2012) keine eigene Dimension für die Prozesse, jedoch enthalten die TIMSS-Anforderungsbereiche einige Aspekte der in Tabelle 4.10 aufgezeigten Prozesse des Ländervergleichs bzw. NEPS. So wird z. B. im Anforderungsbereich *Reproduzieren* definiert, dass die Schülerinnen und Schüler graphischen Darstellungsformen Daten entnehmen können. Dies zählt beim Ländervergleich zu dem Prozess *Darstellen* und bei NEPS zu *Repräsentieren (Darstellungen verwenden)*. Im Anforderungsbereich *Anwenden* wird bei TIMSS vorausgesetzt, dass die Schülerinnen und Schüler gezielt Rechenwege und Lösungsstrategien auswählen können. Dieser Aspekt würde beim Ländervergleich und bei NEPS in den Prozess *Problemlösen* bzw. *mathematische Probleme lösen* eingeordnet werden.

Diese Überschneidungen machen die eingangs beschriebene enge Verzahnung der unterschiedlichen Bereiche umso deutlicher.

4.4.1.4 Kognitive Anforderungsbereiche

Die beiden Studien TIMSS (Selter et al., 2012) und der Ländervergleich (Roppelt und Reiss, 2012) definieren jeweils drei Anforderungsbereiche (vgl. Tabelle 4.11), die sich auf die Bearbeitung der Aufgabe bzw. von Problemstellungen im Allgemeinen beziehen. Gemein haben die Studien, dass sich im ersten Anforderungsbereich routinierte Aufgabenstellungen oder Verfahren finden lassen. Die Schülerinnen und Schüler müssen hierzu ihr Grundwissen z. B. bezüglich Definitionen und Bezeichnungen wiedergeben.

Tabelle 4.11: Anforderungsbereich von TIMSS und dem Ländervergleich

TIMSS	**Ländervergleich**
Reproduzieren	Reproduzieren
Anwenden	Zusammenhänge herstellen
Komplexe Probleme lösen	Verallgemeinern und Reflektieren

Der zweite Anforderungsbereich umfasst bei beiden Studien Problemstellungen bzw. Aufgaben, die den Schülerinnen und Schülern noch eher vertraut sind. Während beim TIMSS-Mathematiktest das Anwenden des mathematischen Wissens der Schülerinnen und Schüler im Vordergrund steht, liegt im Ländervergleich der Fokus auf der Herstellung von Zusammenhängen und die Interpretationsfähigkeit der Schülerinnen und Schüler. Aufgaben, die dem dritten Anforderungsbereich zugeordnet werden können, beinhalten in den beiden Studien unbekannte Problemstellungen. Um Aufgaben des dritten Anforderungsbereichs lösen zu können, müssen die Schülerinnen und Schüler vielfältiges logisches und systematisches Denken beweisen sowie ihre Antworten bzw. Lösungen begründen können. Während es beim TIMSS-Test eher um das Lösen mehrschrittiger Probleme bzw. die Kombination von mehreren Lösungsschritten geht sowie um das Herstellen von Verknüpfungen, das Generalisieren und Spezifizieren, wird bei dem Ländervergleich von den Schülerinnen und Schülern erwartet, dass sie eigene Strategien entwickeln können und Reflexionsvermögen haben. Weiterhin bleibt ebenfalls an dieser Stelle festzuhalten, dass die Anforderungsbereiche von TIMSS auch zum Teil Prozesse beschreiben (vgl. Kapitel 4.4.1.3).

4.4.1.5 Zwischenfazit

Grundlegend bleibt zunächst festzuhalten, dass nicht alle drei Schulleistungsstudien zwischen den Bereichen Inhalt, Prozess und Anforderung unterscheiden. Zusammenfassend wird dies in der Abbildung 4.7 dargestellt. Die Häkchen symbolisieren, dass diese Bereiche in der jeweiligen Studie abgedeckt werden. Es hat sich gezeigt, dass alle drei Studien verschiedene inhaltsbezogene Kompetenzen unterscheiden, die eine hohe Übereinstimmung aufweisen. Vor allem zwischen den nationalen Studien besteht ein hoher Zusammenhang. Die hohen Überschneidungen liegen u. a. daran, dass sich die neueren Studien meist auf die Vorarbeiten von bereits bestehenden Studien stützen bzw. dass eine Anlehnung an nationale und internationale Schulleistungsstudien gezielt angestrebt wird, um z. B. bestehende Aufgaben nutzen zu können und Vergleichbarkeit zu ermöglichen (vgl. dazu auch Ehmke et al., 2009). Dies ist bei den prozessbezogenen Kompetenzen ebenfalls der Fall, wobei die prozessbezogenen Kompetenzen unter dieser Bezeichnung nur beim Ländervergleich und NEPS formuliert werden. NEPS orientierte sich bei den prozessbezogenen Kompetenzen an den KMK-Bildungsstandards und entspricht damit dem Ländervergleich approximativ. Jedoch werden in den Anforderungsbereichen von TIMSS auch prozessbezogene Kompetenzen formuliert. Die in TIMSS beschriebenen Anforderungsbereiche entsprechen wiederum weitestgehend den Anforderungsbereichen, die auch im Ländervergleich differenziert werden.

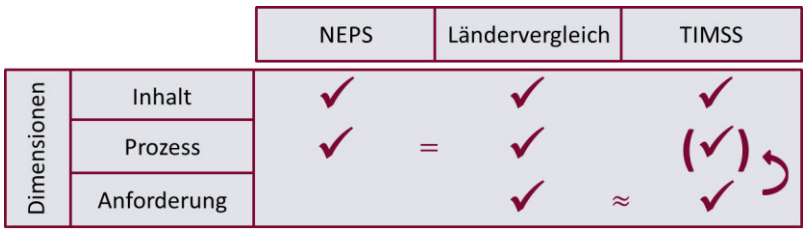


Abbildung 4.7: Untersuchte Dimensionen in den drei Studien

Die inhaltliche Gegenüberstellung der mathematischen Rahmenkonzeptionen weist darauf hin, dass sich die Konzeptionen der drei Studien gut miteinander vereinbaren lassen und in den Studien scheinbar ein sehr ähnliches Konstrukt mathematischer Kompetenz gemessen wird. Dies soll jedoch in den folgenden Kapiteln noch detaillierter untersucht werden, da die inhaltliche Analyse der Konzeptionen der Studien nur ein unvollständiges Bild liefern kann. Die Definition der unterschiedlichen Bereiche ist meist sehr knapp gehalten und

umfasst nicht alle möglichen Aspekte, sodass ein Vergleich zunächst auf einer oberflächlichen Ebene bleibt. Offen bleibt daher die Frage, ob die gleichen Bezeichnungen auch tatsächlich bedeuteten, dass sich die gleichen Inhalte dahinter verbergen oder ob es sich um eine sogenannte jingle fallacy handelt. Demgegenüber bleibt auch die Frage offen, ob sich hinter den unterschiedlichen Bezeichnungen tatsächlich unterschiedliche Inhalte verbergen oder aber gleiche (jangle fallacy). Um dies näher zu untersuchen, wurde in einem nächsten Schritt ein Expertenreview durchgeführt (vgl. Kapitel 3.6.1). Im folgenden Kapitel werden die Ergebnisse dieses Expertenreviews präsentiert.

Tabelle 4.12: Vergleich der Rahmenkonzeptionen der drei Studien

	TIMSS 2011	**Ländervergleich 2011**	**NEPS 2010**
Inhaltsbereiche	- Arithmetik	- Zahlen und Operationen	- Quantität
	- Geometrie/Messen	- Größen und Messen	- Veränderung und Beziehungen
	- Daten	- Muster und Strukturen	- Raum und Form
		- Raum und Form	- Daten und Zufall
		- Daten, Häufigkeit und Wahrscheinlichkeit	
Prozedurale Fähigkeiten		- Darstellen	- Repräsentieren
		- Modellieren	- Modellieren
		- Argumentieren	- Mathematisch Argumentieren
		- Kommunizieren	- Mathematisch Kommunizieren
		- Problemlösen	- Mathematische Probleme lösen
		- Technische Grundfertigkeiten	- Technische Fertigkeiten einsetzen
Kognitive Anforderungsbereiche	- Reproduzieren	- Reproduzieren	
	- Anwenden	- Zusammenhänge herstellen	
	- Probleme lösen	- Verallgemeinern und reflektieren	

4.4.2 Vergleich der mathematischen Inhalte

Im Rahmen eines Expertenreviews wurden die NEPS-Items in die Rahmenkonzeptionen von TIMSS für die vierte Jahrgangsstufe und dem Ländervergleich Mathematik Primar eingeordnet, um Aussagen darüber treffen zu können, inwieweit sich die NEPS-Mathematikaufgaben in die Frameworks der anderen Studien einordnen lassen. Die Übereinstimmung zwischen den drei Ratern wird in den Ergebnistabellen als prozentuale Übereinstimmung sowie als Cohens Kappa angegeben (vgl. Kapitel 3.6.1).

Ziel des Expertenreviews ist zu analysieren, ob die drei Studien das Konstrukt mathematischer Kompetenz gleich oder ähnlich messen (Forschungsfrage 1d). Im vorangegangen Artikel wurden die Übereinstimmungen und Unterschiede bezüglich der Konstrukte bereits anhand der in den Rahmenkonzeptionen beschriebenen Definitionen verglichen. Das Expertenreview erlaubt darüber hinaus Aussagen dazu, wie vergleichbar die Inhaltsbereiche der drei Studien auf Aufgabenebene sind. Zusätzlich können Analysen zu der Vergleichbarkeit der Gewichtung der Inhaltsbereiche in den drei Studien Auskunft darüber geben, wie ähnlich sich die Studien hinsichtlich ihrer definierten Konstrukte sind.

NEPS-Aufgaben in der Rahmenkonzeption von TIMSS

Inhaltsbereiche

Bezüglich der Einordnung der NEPS-Items in die Rahmenkonzeption von TIMSS lässt sich zunächst festhalten, dass sich von den 24 NEPS-Aufgaben 23 Aufgaben in die TIMSS-Rahmenkonzeption einordnen lassen. Lediglich eine Aufgabe lässt sich nicht zuordnen, weil die TIMSS-Rahmenkonzeption keine Einordnung von Aufgaben zur Wahrscheinlichkeitsrechnung berücksichtigt (vgl. Kapitel 4.4.1.2), diese Aufgabe jedoch vornehmlich Kompetenzen dieses Inhaltsbereiches erfordert. Die Tabelle 4.13 zeigt eine Übersicht über die Verteilung der Aufgaben zu den Inhaltsbereichen. Auf der linken Seite wird die Verteilung der NEPS-Mathematikaufgaben zugeordnet zu den TIMSS-Inhaltsbereichen dargestellt. Im Vergleich dazu, wird auf der rechten Seite die Verteilung der Mathematikitems aus der TIMSS-Hauptuntersuchung angegeben. Es zeigt sich, dass sich sowohl in der TIMSS-Hauptuntersuchung wie auch bei der Einordnung der NEPS-Items in die Inhaltsbereiche der TIMSS-Studie die meisten Aufgaben dem Inhaltsbereich *Arithmetik* zuordnen lassen (TIMSS 51%; NEPS in TIMSS 46%). Weiterhin werden sieben der NEPS-Aufgaben dem TIMSS-

Inhaltsbereich *Geometrie/Messen* sowie fünf der NEPS-Aufgaben dem Bereich *Daten* zugeordnet.

Weiterhin ist der Tabelle 4.13 zu entnehmen, dass die Reihenfolge, gemessen an der prozentualen Häufigkeit der eingeordneten Aufgaben, für beide Zuordnungen übereinstimmt. Die Unterschiede in den prozentualen Verteilungen sind nicht signifikant ($\chi^2 = 0.79,\ df = 2$). Ein detaillierterer Vergleich der prozentualen Verteilung zeigt, dass der NEPS-Test tendenziell über weniger Aufgaben verfügt, die den TIMSS-Inhaltsbereichen *Arithmetik* und *Geometrie/Messen* zugeordnet wurden, als in der TIMSS-Hauptuntersuchung. Für den Inhaltsbereich *Daten* verhält es sich genau andersherum. NEPS beinhaltet einen höheren Anteil von Aufgaben, die dem TIMSS-Inhaltsbereich *Daten* zugeordnet werden als der TIMSS-Test.

Tabelle 4.13: Anzahl und prozentuale Verteilung der Aufgaben auf die Inhaltsbereiche der TIMSS-Rahmenkonzeption in dem TIMSS-Test und dem NEPS-Test

	NEPS im TIMSS 2011 Framework*		**TIMSS 2011**	
	N	**%**	**N**	**%**
Arithmetik	11	46	90	51
Geometrie/Messen	7	29	61	34
Daten	5	21	26	15
Nicht zuordbar	1	4		
Gesamt	24	100	177	100

* Prozentuale Übereinstimmung (PÜ) = 91.7; Mittelwert (MW) des Cohens Kappa (κ) = .78

In einem zweiten Schritt wird die Einordnung der NEPS-Items zu den NEPS-Inhaltsbereichen mit der Zuordnung der NEPS-Items zu den TIMSS-Inhaltsbereichen verglichen. Die Ergebnisse sind der Tabelle 4.14 zu entnehmen. Es zeigt sich, dass die NEPS-Aufgaben, die dem NEPS-Inhaltsbereich *Quantität* zugeordnet wurden, in der TIMSS-Rahmenkonzeption schwerpunktmäßig dem Inhaltsbereich *Arithmetik* (6 Aufgaben) aber auch *Geometrie/Messen* (2 Aufgaben) zugeordnet werden. Bei dem NEPS-Inhaltsbereich *Veränderung und Beziehungen* findet wieder eine Aufteilung in zwei TIMSS-Inhaltsbereiche statt. Fünf Aufgaben fallen nach der TIMSS-Rahmenkonzeption dem Inhaltsbereich *Arithmetik*

und eine Aufgabe *Daten* zu. NEPS-Aufgaben, die dem Inhaltsbereich *Raum und Form* in der NEPS-Rahmenkonzeption entsprechen, fallen bei TIMSS alle dem Inhaltsbereich *Geometrie/Messen* zu. Die NEPS-Aufgaben zum Inhaltsbereich *Daten und Zufall* entsprechen alle dem Inhaltsbereich *Daten* in der TIMSS-Rahmenkonzeption – ausgenommen der einen Aufgabe zur Wahrscheinlichkeitsrechnung.

Tabelle 4.14: Anzahl und prozentuale Verteilung der NEPS-Aufgaben in der NEPS-Rahmenkonzeption und in der TIMSS-Rahmenkonzeption

		NEPS							
		Quantität		Veränderung und Beziehungen		Raum und Form		Daten und Zufall	
		N	**%**	**N**	**%**	**N**	**%**	**N**	**%**
TIMSS	Arithmetik	6	25	5	21				
	Geometrie/Messen	2	8			5	21		
	Daten			1	4			4	17
	Nicht zuordbar							1	4

* Prozentuale Übereinstimmung (PÜ) = 91.7; Mittelwert (MW) des Cohens Kappa (κ) = .78

Zusammenfassend lässt sich konkludieren, dass die Ergebnisse zur Vergleichbarkeit der Gewichtung auf die Inhaltsbereiche *Arithmetik*, *Geometrie/Messen* und *Daten* und zur Vergleichbarkeit der Inhaltsbereiche auf der Aufgabenebene eine hohe Übereinstimmung in unterschiedlichen Aspekten aufzeigen (vgl. Abbildung 4.8). Für die Zuordnung der Inhaltsbereiche von NEPS in TIMSS konnte festgestellt werden, dass die Reihenfolge der prozentualen Übereinstimmung der Aufgabenhäufigkeit innerhalb der Inhaltsbereiche übereinstimmt. Auch wenn die prozentualen Übereinstimmungen dabei nicht identisch sind, so sind die Unterschiede (z. B. zwischen *Arithmetik* bei NEPS in TIMSS 46% und bei TIMSS 51%) an keiner Stelle signifikant. Die hohen Übereinstimmungen bezüglich der Zuordnung der NEPS Aufgaben zu der NEPS- und der TIMSS-Rahmenkonzeption sind daher sehr zufriedenstellend. Erwartet wurde nach dem inhaltlichen Vergleich der Rahmenkonzeptionen bezüglich der Inhaltsbereiche (vgl. Kapitel 4.4.1.2), dass der Inhaltsbereich *Arithmetik* (TIMSS) höhe Überschneidungen mit den Inhaltsbereichen *Quantität* (NEPS) und *Veränderung und Beziehung*en (NEPS) aufweist. Dies konnte durch das Expertenreview bestätigt werden. Bei elf

NEPS-Aufgaben aus den Inhaltsbereichen *Quantität* und *Veränderung und Beziehungen* passt die Zuordnung zu dem TIMSS-Inhaltsbereich *Arithmetik*. Lediglich zwei Aufgaben aus dem Bereich *Quantität* (NEPS) wurden dem TIMSS-Inhaltsbereich *Geometrie und Messen* zugeordnet. Dies liegt daran, dass in der Aufgabe die Berechnung von Längen und Gewichten erforderlich ist und TIMSS die Berechnung von Längen, Gewichten und Rauminhalten dem Inhaltsbereich *Geometrie/Messen* zuordnet, und NEPS alle Maßeinheiten unter dem Inhaltsbereich *Quantität* fasst. Eine Aufgabe aus dem NEPS-Inhaltsbereich *Veränderung und Beziehungen* wurde dem TIMSS-Inhaltsbereich *Daten* zugeordnet. In dieser Aufgabe sind so

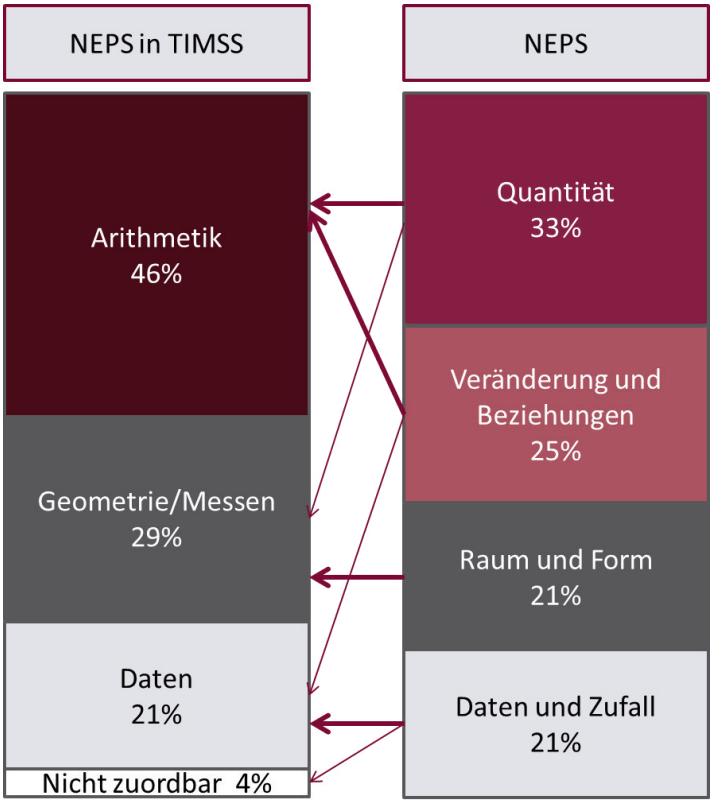


Abbildung 4.8: Verteilung der NEPS Mathematikaufgaben auf die Inhaltsbereiche der TIMSS-Rahmenkonzeption

wohl Kompetenzen in den Bereichen *Daten und Zufall* sowie *Veränderung und Beziehung*en erforderlich. Die Entwickler des NEPS-Tests haben sich dazu entschieden, dass vornehmlich die Kompetenz im Bereich *Veränderung und Beziehungen* nötig ist, um die Aufgabe zu lösen. In der TIMSS-Rahmenkonzeption heißt es unter dem Inhaltsbereich *Daten* jedoch explizit, dass hierzu auch Aufgaben zählen, bei denen Daten verwendet werden müssen und in denen das Kombinieren von Daten und/oder das Durchführen von Berechnungen auf Grundlage der Daten für die Beantwortung der Frage nötig sein kann. Die Inhaltsbereiche *Daten und Zufall*

sowie *Raum und Form* im NEPS entsprechen genau ihren Pendants in TIMSS (*Daten* und *Geometrie/Messen*). Insofern kann festgehalten werden, dass sich die Inhaltsbereiche der NEPS- und der TIMSS-Tests nicht eins zu eins entsprechen, sich jedoch die NEPS-Aufgaben bis auf eine Ausnahme der TIMSS-Rahmenkonzeption zuordnen lassen.

Kognitive Anforderungsbereiche

Die 24 NEPS-Aufgaben lassen sich ohne Ausnahme einem der drei kognitiven Anforderungsbereichen definiert nach der TIMSS-Rahmenkonzeption zuordnen. Sieben Aufgaben erfassen eine Kompetenz im Bereich *Reproduzieren*, 11 Aufgaben in *Anwenden* und 6 Aufgaben in *Problemlösen*. Die Gegenüberstellung der Verteilung der TIMSS-Aufgaben und der NEPS Aufgaben in den Anforderungsbereichen der TIMSS-Rahmenkonzeption zeigt, dass NEPS 11% weniger Aufgaben im kognitiven Anforderungsbereich *Reproduzieren* hat und dafür 6% mehr Aufgaben im Bereich *Anwenden* und 5% mehr, die die Kompetenz des *Problemlösens* erfordern. Die Unterschiede sind jedoch auch hier nicht signifikant ($\chi^2 = 1.11, df = 2$).

Tabelle 4.15: Anzahl und prozentuale Verteilung der Aufgaben auf die kognitiven Anforderungsbereiche der TIMSS-Rahmenkonzeption im TIMSS- und NEPS-Test

	NEPS in TIMSS*		TIMSS	
	N	%	N	%
Reproduzieren	7	29	71	40
Anwenden	11	46	71	40
Problemlösen	6	25	35	20
Gesamt	24	100	177	100

* PÜ = 51.8, κ = 0.1 (MW)

Die Ergebnisse zur Vergleichbarkeit der Verteilung der Aufgaben auf die kognitiven Anforderungsbereiche *Reproduzieren*, *Anwenden* und *Problemlösen* zeigen hohe Übereinstimmungen und keine bedeutsamen Unterschiede. Es zeigen sich lediglich leicht unterschiedliche Tendenzen. Bezüglich der Reihenfolge der prozentualen Häufigkeit der Aufgaben zu den kognitiven Anforderungsbereichen lässt sich festhalten, dass bei NEPS in TIMSS der kognitive Anforderungsbereich *Anwenden* mit 46% der prozentual häufigste Inhaltsbereich ist, mit deutlichem Abstand zum *Reproduzieren* (29%) und dem *Problemlösen* (25%). In TIMSS hingegen ist die Gewichtung der kognitiven Anforderungsbereichen

Reproduzieren und *Anwenden* mit jeweils 40% gleich, jedoch sind Aufgaben aus dem Bereich des Problemlösens mit 20% nur halb so oft vorzufinden wie die der anderen beiden kognitiven Anforderungsbereiche. Der NEPS-Test hat also im Vergleich zum TIMSS-Test tendenziell mehr Aufgaben, die reines Reproduzieren von Wissen erfordern und mehr Aufgaben, in denen die Schülerinnen und Schüler ihr Wissen anwenden und Probleme lösen müssen. Zusammenfassend bleibt festzuhalten, dass sich alle Aufgaben des NEPS-Tests den kognitiven Anforderungsbereichen der TIMSS-Rahmenkonzeption zuordnen lassen und dass alle hier definierten Anforderungsbereiche im NEPS Test abgedeckt werden, wenn auch in einer tendenziell unterschiedlichen Fokussierung.

NEPS-Aufgaben in der Rahmenkonzeption des Ländervergleichs

Inhaltsbereiche

Die NEPS-Mathematikaufgaben wurden ebenfalls der Rahmenkonzeption des Ländervergleichs zugeordnet. Zunächst lässt sich festhalten, dass alle 24 NEPS-Aufgaben den Inhaltsbereichen des Ländervergleichs zugeordnet werden konnten. Bei der Zuordnung der

Tabelle 4.16: Anzahl und prozentuale Verteilung der Aufgaben auf die Inhaltsbereiche der Ländervergleichs-Rahmenkonzeption im Ländervergleichs-Test und dem NEPS-Test

	NEPS im Ländervergleich*		**Ländervergleich**	
	N	%	N	%
Zahlen und Operationen	9	27	88	27
Raum und Form	5	15	60	18
Muster und Strukturen	7	21	48	15
Größen und Messen	6	18	68	21
Daten, Häufigkeit und Wahrscheinlichkeit	7	21	66	20
Gesamt	34	100	330	100

* Mehrfachzuordnung möglich; PÜ = 87.2; κ = .6 (MW)
Rundungsfehler in den prozentualen Verteilungen sind möglich

NEPS-Aufgaben in die Ländervergleichs-Rahmenkonzeption wurde eine Mehrfachzuordnung ermöglicht, da auch die Rahmenkonzeption des Ländervergleichs dies zulässt. Insgesamt konnten 9 Aufgaben dem Inhaltsbereich *Zahlen und Operationen* zugeordnet werden, 5 Aufgaben dem Bereich *Raum und Form*, 7 Aufgaben erfassen Kompetenzen im Bereich *Muster*

und Strukturen, 6 Aufgaben zählen zu *Größen und Messen* und 7 Aufgaben zu *Daten, Häufigkeit und Wahrscheinlichkeit*. Tabelle 4.17 stellt die Verteilung der Aufgaben des Ländervergleichs- und des NEPS-Tests auf die Inhaltsbereiche des Ländervergleichs dar. Die beiden Verteilungen sind sehr ähnlich. Der NEPS-Test hat tendenziell weniger Aufgaben, die dem Kompetenzbereich *Raum und Form* sowie *Größen und Messen* zugeordnet werden können, aber prozentual mehr Aufgaben im Bereich *Muster und Strukturen*. Die Unterschiede sind jedoch nicht signifikant ($\chi^2 = 1.09, df = 4$).

In der Tabelle 4.17 wird die Einordnung der NEPS Aufgaben in die Inhaltsbereiche von NEPS und dem Ländervergleich gegenübergestellt. Die Aufgaben, die in NEPS die Kompetenz im Inhaltsbereich *Quantität* erfassen sollen, verteilen sich nach der Rahmenkonzeption vom Ländervergleich auf die Inhaltsbereiche *Zahlen und Operationen* (6 Aufgaben), *Größen und Messen* (4 Aufgaben), *Muster und Strukturen* (1 Aufgabe) und Daten *Häufigkeit und Wahrscheinlichkeit* (1 Aufgabe). Die NEPS Aufgaben im Bereich *Veränderung und Beziehungen* verteilen sich auf *Zahlen und Operationen* (3 Aufgaben), *Größen und Messen* (2 Aufgaben), *Muster und Strukturen* (4 Aufgaben) sowie *Daten, Häufigkeit und Wahrscheinlichkeit* (1 Aufgabe). Im Inhaltsbereich *Raum und Form* im NEPS findet eine eins-zu-eins-Zuordnung zu dem Inhaltsbereich *Raum und Form* im Ländervergleich statt. Die *Daten und Zufall* Aufgaben fallen im Ländervergleich dem Inhaltsbereich *Muster und Strukturen* (2 Aufgaben) und dem Inhaltsbereich *Daten, Häufigkeit und Wahrscheinlichkeit* (5 Aufgaben) zu.

Tabelle 4.17: Anzahl und prozentuale Verteilung der NEPS-Aufgaben in der NEPS-Rahmenkonzeption und in der Rahmenkonzeption des Ländervergleichs

		NEPS							
		Quantität		Veränderung und Beziehungen		Raum und Form		Daten und Zufall	
		N	**%**	**N**	**%**	**N**	**%**	**N**	**%**
Ländervergleich*	Zahlen und Operationen	6	18	3	9				
	Größen und Messen	4	12	2	6				
	Muster und Strukturen	1	3	4	12			2	6
	Raum und Form					5	15		
	Daten, Häufigkeit und Wahrscheinlichkeit	1	3	1	3			5	15

* Mehrfachzuordnung möglich; PÜ = 87.2; κ = .6 (MW)

Zusammenfassend lässt sich zunächst festhalthalten, dass alle 24 NEPS-Items den Inhaltsbereichen, die in der Rahmenkonzeption des Ländervergleichs definiert werden, zugeordnet werden können (vgl. Abbildung 4.9). Es gibt demnach keine mathematischen Inhalte, die im NEPS, jedoch nicht im Ländervergleich erfasst werden. Da der NEPS-Test für die fünfte Jahrgangsstufe konzipiert wurde, ist dies ein umso relevanteres Ergebnis. Die Gewichtung der Verteilung der NEPS-Mathematikaufgaben und der im Ländervergleich eingesetzten Mathematikaufgaben in der Rahmenkonzeption des Ländervergleichs zeigt keine signifikanten Unterschiede hinsichtlich der Zuordnung zu den Inhaltsbereichen. Tendenziell hat der NEPS-Test etwas weniger Aufgaben in den Inhaltsbereichen *Raum und Form* und *Größen und Messen* des Ländervergleichs. Hingegen gibt es mehr Mathematikaufgaben im NEPS-Test, die dem Inhaltsbereich *Muster und Strukturen* im Ländervergleich zugeordnet werden können. Wird die Gegenüberstellung der Zuordnung der NEPS-Aufgaben auf die NEPS- und die Ländervergleichs-Rahmenkonzeption betrachtet, zeigt sich, dass sich die Definitionen der Inhaltsbereiche in den beiden Studien nicht eins zu eins entsprechen. In Kapitel 4.4.1.2 wurde bereits eine Annahme getroffen, welche Inhaltsbereiche der NEPS-Rahmenkonzeption denen der Ländervergleichs-Rahmenkonzeption in etwa gleichkommen. Diese Annahme wurde an dieser Stelle überprüft. Bei der Interpretation der Daten muss berücksichtigt werden, dass die NEPS-Items in der Rahmenkonzeption des Ländervergleichs zum Teil mehrfachzugeordnet wurden. Dies bedeutet, dass in der Tabelle 4.17 auch die zweite Zuordnung der Rater zu einem Inhaltsbereich zugelassen wurde. Bei 22 NEPS-Aufgaben entspricht die Zuordnung der Rater hinsichtlich der Zuordnung zum Inhaltsbereich der Rahmenkonzeption des Ländervergleichs dem in Kapitel 4.4.1.2 angenommenem Inhaltsbereich in der NEPS-Rahmenkonzeption. Beispielsweise wurden die fünf Aufgaben, die im NEPS dem Inhaltsbereich *Daten und Zufall* zugeordnet werden, in der Rahmenkonzeption des Ländervergleichs durch die Rater ebenfalls dem Inhaltsbereich *Daten, Häufigkeit und Wahrscheinlichkeit* zugeordnet. Da die Rater jedoch zusätzlich einen zweiten Inhaltsbereich pro Aufgabe auswählen konnten, der in der Aufgabe ebenfalls erfasst wird, wurden zwei der Aufgaben zusätzlich dem Inhaltsbereich *Muster und Strukturen* zugeschrieben. Es gibt jedoch zwei Aufgaben, bei denen dies nicht der Fall ist. Die Aufgaben, die in der NEPS-Rahmenkonzeption dem Inhaltsbereich *Veränderung und Beziehungen* zugeordnet wurden, zählen laut dem Urteil der Rater zu dem Ländervergleichs-Inhaltsbereich *Zahlen und Operationen*. In den beiden Aufgaben geht es sowohl um das Anwenden von Grundrechenarten bzw. um Beziehung zwischen Zahlen als auch um das

Erkennen und Umsetzen von funktionalen Beziehungen. Es wird angenommen, dass die Rater einen anderen Schwerpunkt in den beiden Aufgaben gesehen haben, als die Experten, die den NEPS-Test erstellt haben. Trotz dieser marginalen Unterschiede lässt sich konkludieren, dass die in Kapitel 4.4.1.2 angenommenen Beziehungen zwischen den Inhaltsbereichen der beiden Studien durch das Expertenreview bestätigt werden konnten und damit hohe Gemeinsamkeiten zwischen den definierten Inhaltsbereichen bestehen.

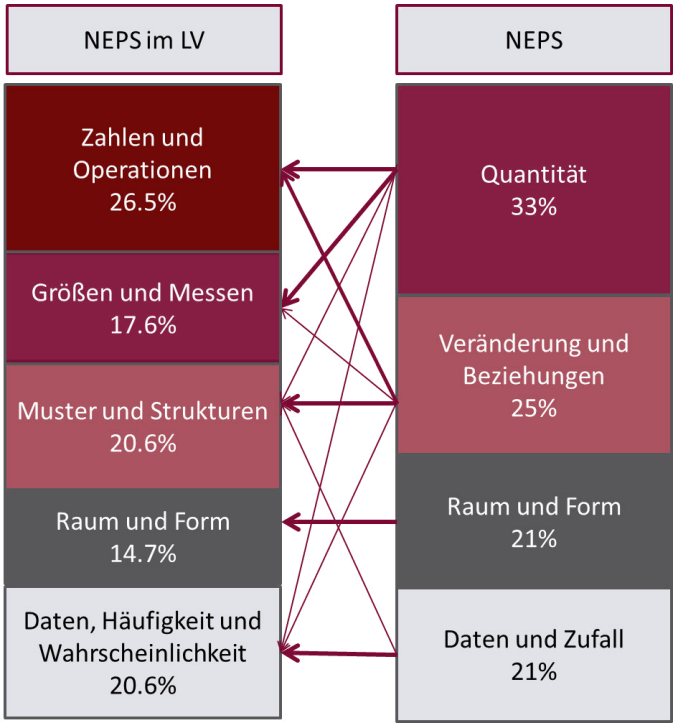


Abbildung 4.8: Verteilung der NEPS-Mathematikaufgaben auf die Inhaltsbereiche der Ländervergleichs-Rahmenkonzeption

Prozessbezogene Kompetenzen

Neben den Inhaltsbereichen unterscheidet der Ländervergleich prozessbezogene Kompetenzen. Daher wurde durch die Rater zusätzlich eine Zuordnung der NEPS-Aufgaben zu den prozessbezogenen Kompetenzen der Rahmenkonzeption des Ländervergleichs vorgenommen. Auch im Kompetenzbereich der Prozesse (*Problemlösen*, *Kommunizieren*, *Argumentieren*, *Modellieren*, *Darstellen* & *technische Grundfertigkeiten*) fand eine Mehrfachzuordnung der Aufgaben durch die Experten statt, da die Rahmenkonzeption des Ländervergleichs dies ebenfalls zulässt. Insgesamt wurden 15 Aufgaben dem Prozess des *Problemlösens* zugeordnet, 7 Aufgaben erfordern die Kompetenz des *Kommunizierens*, 6 Items erfordern das *Argumentieren*, in 14 Aufgaben wird der Prozess des *Modellierens* erwartet, in 10 Aufgaben wird die Kompetenz des *Darstellens* angesprochen und ebenfalls in

10 Aufgaben werden technischen Grundfertigkeiten benötigt. Eine Gegenüberstellung mit den Verteilungen aus dem Ländervergleich ist bezüglich der prozessbezogenen Kompetenzen nicht möglich, da diese Verteilung nicht berichtet wurde. Die Ergebnisse lassen dennoch den Schluss zu, dass im NEPS-Test alle im Ländervergleich definierten prozessbezogenen Kompetenzen abgedeckt werden. Ein Schwerpunkt des NEPS-Tests liegt auf dem Prozess des Problemlösens und des Modellierens, was den Erwartungen entspricht, da der NEPS-Test auf dem Konzept der Mathematical Literacy basiert.

Zwischenfazit

Da der Vergleich der Rahmenkonzeptionen (vgl. Kapitel 4.4.1) nur ein unzureichendes Bild hinsichtlich der Fragestellung, ob es sich eventuell um eine jingle oder jangle fallacy handelt, liefern konnte, wurden die NEPS-Aufgaben von Experten den Rahmenkonzeptionen von TIMSS und dem Ländervergleich zugeordnet und die prozentualen Verteilungen auf die unterschiedlichen mathematischen Bereiche wurde miteinander verglichen. Dies liefert Hinweise über die Ähnlichkeiten und Unterschiede in der Erfassung der mathematischen Konstrukte. Es zeigt sich, dass sich fast alle NEPS-Mathematikaufgaben in die Rahmenkonzeptionen von TIMSS und dem Ländervergleich einordnen lassen. Die Unterschiede in den prozentualen Verteilungen der NEPS-Mathematikaufgaben auf die jeweiligen Subdimensionen unterscheiden sich nicht signifikant zu denen der TIMSS und Ländervergleichsverteilungen der Hauptuntersuchungen. Es liegt damit zwar keine vollständige Überschneidung vor, jedoch bestehen sehr große Ähnlichkeiten zwischen den definierten Konstrukten bzw. den Mathematikaufgaben. Die gefundenen Unterschiede sind dabei nur tendenziell und nicht statistisch abzusichern. Grundsätzlich bleibt daher zunächst festzuhalten, dass die Ähnlichkeit der definierten Konstrukte eine gute Voraussetzung für das Linking ist. Weitere Analysen auf dimensionaler und skalenbezogener Ebene werden diesbezüglich jedoch noch einen detaillierteren Vergleich ermöglichen (vgl. Kapitel 5). Ein Equating und die damit einhergehenden Interpretationsmöglichkeiten sind jedoch aufgrund der gefundenen Unterschiede nicht empfehlenswert (vgl. Kapitel 1.1.1).

4.4.3 Vergleich der Aufgabenmerkmale

Zusätzlich zu dem Vergleich der mathematischen Aspekte in den Aufgaben des NEPS, TIMSS und Ländervergleich sollen die Aufgaben auch hinsichtlich ihrer formalen und

sprachlichen Merkmale untersucht werden (Forschungsfrage 1d). Hierzu wurde ein zweites Expertenreview durchgeführt, bei dem die Aufgaben der Studien hinsichtlich vielfältiger Merkmale (vgl. Kapitel 3.6.1.2) klassifiziert wurden. In diesem Kapitel werden die Ergebnisse der Klassifikation der drei Studien einander gegenübergestellt. Ziel der Aufgabenklassifikation ist, einen Vergleich der drei Studien auf Aufgabenebene Rater-Reliabilität wird in drei unterschiedlichen Maßen angegeben: Prozentuale Übereinstimmung, Cohens Kappa und Inter-Klassen-Korrelation (vgl. Kapitel 3.6.1.3).

Formale Testmerkmale

Die Tabelle 4.18 zeigt eine Gegenüberstellung der formalen Testmerkmale der drei Studien (vgl. Kapitel 3.6.1.2). Fett markiert sind hierbei die Merkmale, hinsichtlich derer sich der NEPS-Test signifikant vom Ländervergleich- bzw. TIMSS-Test unterscheidet. Der TIMSS- und der Ländervergleichs-Test wurden hierbei wiederum nicht einander gegenübergestellt.

Im NEPS-Test haben knapp 21% der Aufgaben einen Stimulus, d. h. die Aufgaben haben einen Einführungstext, auf den sich anschließend mehrere Aufgaben beziehen. Der Ländervergleichs-Test hat mit 49.4 % aller Aufgaben einen signifikant höheren Anteil von Aufgaben mit einem Stimulus als der NEPS-Test ($\chi^2 = 7.21,\ df = 1$). Der TIMSS-Test verfügt mit einem Anteil von 12 % der Aufgaben über den geringsten Anteil an Stimuli. Bezüglich des Kontextbezuges der Aufgaben zeigt sich, dass sich der NEPS-Mathematiktest von den beiden Studien nicht signifikant unterscheidet (NEPS/TIMSS: $\chi^2 = 1.26,\ df = 1$; NEPS/ Ländervergleich: $\chi^2 = 1.02,\ df = 1$). Der NEPS-Test hat mit 62,5 % Aufgaben mit einem Kontextbezug den höchsten Anteil. Grafische Elemente werden am häufigsten im Ländervergleichs-Test verwendet, im TIMSS-Test etwas seltener und beim NEPS-Test werden Aufgaben am wenigsten mit zusätzlichen grafischen Elementen versehen. So kommen Terme oder Formeln im NEPS-Test in nur 12.5 % der Aufgaben vor und Tabellen werden gar nicht verwendet. Die häufigste Form von grafischen Elementen in NEPS sind Graphen, Bilder und Fotos, die in jeweils 20.8 % der NEPS-Aufgaben vorkommen. Der TIMSS-Mathematiktest verwendet am häufigsten Graphen (42.3 % der Aufgaben), hierbei liegt ein signifikanter Unterschied zum NEPS-Test vor ($\chi^2 = 4.06,\ df = 1$). Bilder und Fotos werden in etwa gleich häufig verwendet und Tabellen sowie Terme und Formeln ebenfalls eher selten. Am häufigsten werden Graphen sowohl in dem Ländervergleich- als auch in dem TIMSS-Test verwendet. Bilder, Fotos, Tabellen, Terme und Formeln werden im Ländervergleichs-

Mathematiktest in etwa gleich häufig verwendet (jeweils in ca. 19 % der Aufgaben). Ein signifikanter Unterschied zwischen dem Ländervergleichs- und dem NEPS-Mathematiktest liegt in der Häufigkeit der Verwendung von Tabellen vor ($\chi^2 = 5.85,\ df = 1$). Auch hinsichtlich des Antwortformats zeigen sich einige Unterschiede. In dem NEPS-Mathematiktest kommen nur geschlossene und halboffene Antwortformate vor, wohingegen im Ländervergleich und in TIMSS auch offene Antwortformate vorkommen. Der Unterschied zwischen NEPS und dem Ländervergleich ist hinsichtlich des Antwortformats bedeutsam ($\chi^2 = 9.27,\ df = 2$). Zudem muss in TIMSS und im Ländervergleich in einigen Aufgaben die Lösung in eine vorgegebene Grafik integriert werden, dies kommt bei NEPS nicht vor.

Tabelle 4.18: Gegenüberstellung der formalen Aufgabenmerkmale I

	NEPS		**Ländervergleich**		**TIMSS**	
	N	**%**	**N**	**%**	**N**	**%**
Stimulus vorhanden	5	20.8	129	**49.4**	21	12.0
Kontextbezug vorhanden	15	62.5	135	51.7	88	50.3
Term/Formel vorhanden	3	12.5	47	18.0	18	10.3
Tabelle vorhanden	0	0.0	52	**19.9**	16	9.1
Graph usw. vorhanden	5	20.8	92	35.2	74	**42.3**
Bild/Foto vorhanden	5	20.8	49	18.8	34	19.4
Antwortformat, geschlossen	13	54.0	81	**31.0**	97	55.4
Antwortformat, halboffen	11	45.8	118	**45.2**	52	29.7
Antwortformat, offen	0	0.0	62	**23.8**	26	14.9

PÜ = 87% (MW), κ = .61 (MW)

Tabelle 4.19 stellt weitere formalen Kriterien der drei Tests gegenüber (vgl. Kapitel 3.6.1.2). Fettgedruckt sind hierbei wiederum die formalen Kriterien, bei denen sich der Ländervergleich- bzw. TIMSS-Test signifikant von dem NEPS-Test unterscheidet. Ein Vergleich von TIMSS und dem Ländervergleich wird nicht vorgenommen.

Es zeigt sich, dass der NEPS-Test signifikant mehr Wörter pro Aufgabenstellung verwendet als der Ländervergleich-Test und tendenziell, jedoch nicht signifikant ($T = 2.83,\ df = 24.55$), mehr Wörter als der TIMSS-Test. Die NEPS-Aufgabenstellungen bestehen im Mittel aus 23 Wörtern (SD = 18.51), die Aufgaben des Ländervergleichs aus etwa zwölf Wörtern (SD = 11.11) und im TIMSS-Test aus etwa 18 Wörtern (SD = 12.82). Der NEPS-Test besteht ebenfalls mit durchschnittlich 2 bis 3 Sätzen pro Aufgabenstellung aus den meisten Sätzen. Der Ländervergleichs-Test hat hingegen im Durchschnitt 1 bis 2 Sätze und der TIMSS-

Test etwa 2 Sätze pro Aufgabenstellung. Der Unterschied ist jedoch nur zwischen dem NEPS- und dem Ländervergleichs-Test signifikant ($T = 2.67,\ df = 25.50$). Hinsichtlich der Anzahl der Wörter in dem Antworttext verwendet der NEPS-Test mit knapp 12 Wörtern (SD = 11.56) die meisten Wörter. Der NEPS-Test unterscheidet sich hiermit vom TIMSS-Test, in dem im Mittel etwa 8 Wörter (SD = 8.82) in dem Antworttext vorkommen, und vom Ländervergleich-Test, der im Mittel etwa 9 Wörter (SD = 10.17) im Antworttext hat. Im unteren Teil der Tabelle sind zudem die deskriptiven Analysen zu den Anzahlen der Wörter und Sätze pro gesamte Aufgabe angegeben. Dies bedeutet, dass hier die Wörter und Sätze im Stimulus, im Aufgabentext und im Antworttext zusammengefasst wurden. Es zeigt sich, dass der NEPS-Test aus signifikant mehr Wörtern pro Aufgabe besteht als der Ländervergleichs- ($T = 2.43,\ df = 24.81$) und der TIMSS-Test ($T = 2.26,\ df = 26.04$). Hinsichtlich der Anzahl der Sätze pro Aufgabe lässt sich kein bedeutsamer Unterschied feststellen. Tendenziell hat der NEPS-Test mehr Sätze (MW = 2.83, SD = 1.52) als der Ländervergleichs- (MW = 2.3, SD = 1.24) und der TIMSS-Test (MW = 2.3, SD = 1.45). Hinsichtlich der Anzahl der Antwortalternativen ergeben sich keine Unterschiede zwischen den Studien. Alle drei Studien haben im Mittel etwa 4 Antwortalternativen pro Multiple-Choice-Aufgabe, wobei die Streuung beim Ländervergleichs-Test mit SD = .98 am höchsten ist. Zudem lässt sich festhalten, dass in allen drei Studien mit im Mittel etwa einem mathematischen Begriff pro Aufgabe eher selten mathematische Begriffe verwendet werden.

Sprachliche Komplexität

Die Tabelle 4.20 zeigt die Ergebnisse der Analyse der sprachlichen Komplexität der Aufgaben (vgl. Kapitel 3.6.1.2). Hier werden die prozentualen Häufigkeiten der Aufgaben auf die jeweiligen Niveaustufen pro Bereich und pro Studie angegeben. Zusätzlich wird der Median ausgewiesen (Md). Grundlegend lässt sich zunächst festhalten, dass alle Unterschiede zwischen den Studien hinsichtlich der sprachlichen Komplexität zwischen dem NEPS und dem Ländervergleich bzw. TIMSS nicht signifikant sind und daher nur Tendenzen aufgezeigt werden. Zusammenfassend lässt sich zunächst festhalten, dass die Stimuli- und die Antworttexte in allen drei Studien einfacher sind als der Text in der Aufgabenstellung. Zudem

Tabelle 4.19: Gegenüberstellung der formalen Aufgabenmerkmale II

	NEPS			Ländervergleich			TIMSS		
	N	**MW**	**SD**	**N**	**MW**	**SD**	**N**	**MW**	**SD**
Anzahl Wörter Stimulus	5	19.00	10.95	129	12.71	9.04	21	17.95	10.92
Anzahl Sätze Stimulus	5	1.80	1.10	129	1.53	0.89	21	1.71	0.90
Anzahl Aufgaben Stimulus	5	2.60	0.55	129	3.35	1.20	21	2.81	1.29
Anzahl Wörter Aufgabe	24	23.04	18.51	261	**12.18**	**11.11**	175	17.90	12.82
Anzahl Sätze Aufgabe	24	2.46	1.64	261	**1.54**	**1.25**	175	2.09	1.32
Anzahl Wörter Antworttext	16	11.50	11.56	125	9.45	10.17	77	8.18	8.82
Anzahl Antwortalternativen	13	4.08	0.28	81	4.23	0.98	97	4.04	0.43
Anzahl mathematischer Begriffe	24	0.96	2.29	261	0.75	1.45	175	0.97	1.79
Zusammengefasste Analysen:									
Anzahl Wörter Aufgabe	24	34.67	23.11	261	**22.98**	**14.98**	175	**23.68**	**15.83**
Anzahl Sätze Aufgabe	24	2.83	1.52	261	2.3	1.24	175	2.3	1.45

PÜ = 100% (MW), ICC = 1.0 (MW)

ist die Wortebene meist etwas komplexer hinsichtlich der Sprache als die Satz- und die Textebene. Im Folgenden werden die Ergebnisse im Detail dargestellt.

Als erstes wurde der Stimulus von den Experten eingeordnet. Wie bereits beschrieben, kommt ein Stimulus im NEPS-Test fünfmal vor, im Ländervergleichs-Test gibt es in 129 der 261 klassifizierten Aufgaben einen Stimulus und im TIMSS-Test in 21 der 175 Aufgaben. Insofern wurde auch nur für diese Aufgaben eine Schwierigkeit eingestuft, sodass sich die prozentualen Angaben hier nur auf die Aufgabenauswahl beziehen. Im NEPS-Test ist der Stimulus auf Wortebene (Md = 2) schwieriger als auf Satz- und Textebene (Md = 0). Ein ähnliches Bild zeichnet sich ebenfalls beim Ländervergleich und bei TIMSS ab, wobei die sprachliche Komplexität in dem TIMSS-Test auf Wort- und auf Satzebene im Median auf Niveaustufe 1

Die Aufgabenstellung wurde hinsichtlich der Wortebene von den Experten in allen drei Studien im Median auf Niveaustufe 1 eingestuft. Auf der Satzebene sind die meisten NEPS- und Ländervergleichs-Mathematikaufgaben auf Niveaustufe 0, hingegen bei TIMSS auf Niveaustufe 1. Auf der Textebene verhält es sich ähnlich. In TIMSS-Test gibt es etwas mehr Aufgaben, die Niveaustufe 2 oder 3 zugeordnet werden als beim NEPS-Test und beim Ländervergleichs-Test, wobei der Ländervergleichs-Test auf Textebene noch etwas einfacher ist (Md = 0) als der NEPS-Test (Md = 1).

Einen vorgegebenen Text in der Antwort gab es im NEPS-Test in 16 Aufgaben, im Ländervergleichs-Test in 125 Aufgaben und beim TIMSS-Test in 77 Aufgaben. Hinsichtlich der sprachlichen Schwierigkeit der Antworttexte zeigt sich, dass der Antworttext nicht so komplex ist, wie der Stimulus und die Aufgabenstellung. Bis auf eine Ausnahme liegt der Median in allen drei Studien auf Wort-, Satz- und Textebene auf Niveaustufe 0. Die Unterschiede zwischen den Studien sind dadurch bedingt auch sehr gering. Nur auf der Wortebene ist der Ländervergleich tendenziell etwas schwieriger (Md = 1) als der TIMSS- und der NEPS-Test.

Wiederum wurden die Aufgaben zusammengefasst betrachtet, d. h. es wurde für jede Aufgabe eine sprachliche Schwierigkeit für die Wort-, Satz- und Textebene berechnet. Hierfür wurde die maximale Niveaustufe im Stimulus, in der Aufgabenstellung und im Antworttext pro Aufgabe als Schwierigkeitsindex für die gesamte Aufgabe bestimmt. Die Ergebnisse werden im unteren Teil der Tabelle 4.20 dargestellt. Auch hier sind die Unterschiede nicht signifikant und es werden lediglich Tendenzen aufgezeigt. Bei allen drei Studien zeigt sich, dass die Wortebene von den Experten als etwas schwieriger eingestuft wurde als die Satz- und Textebene, wobei der TIMSS-Test noch etwas schwieriger auf der Wortebene ist als der NEPS- und der Ländervergleichs-Test. Auf der Satz- und der Textebene zeigen sich im Median keine Unterschiede, bei den drei Studien liegt die sprachliche Schwierigkeit im Median auf Niveaustufe 1.

Tabelle 4.20: Gegenüberstellung der sprachlichen Komplexität in den Aufgaben des NEPS, Ländervergleichs- und TIMSS-Mathematiktests

		NEPS						Ländervergleich						TIMSS					
		N	**0**	**1**	**2**	**3**	**Md**	**N**	**0**	**1**	**2**	**3**	**Md**	**N**	**0**	**1**	**2**	**3**	**Md**
Stimulus	Wortebene	5	0	40	60	0	2	129	28	32	39	2	1	21	29	33	38	0	1
	Satzebene	5	60	40	0	0	0	129	53	38	4	5	0	21	38	29	33	0	1
	Textebene	5	60	40	0	0	0	129	58	29	9	5	0	21	57	24	19	0	0
Aufgaben-stellung	Wortebene	24	17	42	29	13	1	261	38	35	24	4	1	175	17	37	38	8	1
	Satzebene	24	54	33	4	8	0	261	59	32	1	8	0	175	35	50	3	12	1
	Textebene	24	46	46	4	4	1	261	56	35	6	3	0	175	45	41	11	3	1
Antworttext	Wortebene	16	75	13	6	6	0	125	47	30	19	3	1	77	60	22	14	4	0
	Satzebene	16	94	6	0	0	0	125	91	6	1	2	0	77	94	5	1	0	0
	Textebene	16	88	13	0	0	0	125	78	20	0	2	0	77	99	1	0	0	0
Zusammen-gefasst	Wortebene	24	13	42	29	17	1	261	19	38	38	5	1	175	14	35	43	9	2
	Satzebene	24	46	42	4	8	1	261	41	46	2	11	1	175	33	49	3	16	1
	Textebene	24	38	54	4	4	1	261	42	43	10	5	1	175	43	42	13	3	1

PÜ = 0.632 (MW), κ = 0.324 (MW)

Zwischenfazit

Die Analysen haben gezeigt, dass sich die drei Studien hinsichtlich der formalen Merkmale in vielen Aspekten sehr ähnlich sind, jedoch auch einige Unterschiede existieren. Bedeutsame Unterschiede konnten zwischen dem Ländervergleichs- und dem NEPS-Mathematiktest bei der Anzahl der Aufgaben mit Stimulus sowie bei der Häufigkeit bei der Verwendung von Tabellen aufgezeigt werden. Im NEPS-Test gibt es keine Aufgabe, die eine Tabelle nutzt und der NEPS-Test hat weniger Aufgaben mit einem Stimulus als der Ländervergleichs-Test. Die Verwendung von Stimuli könnte dazu führen, dass die lokale stochastische Unabhängigkeit verletzt wird. Wird beispielsweise der Aufgabentext im Stimulus nicht verstanden, könnte dies zur Folge haben, dass alle dazu gehörenden Aufgaben nicht beantwortet werden können. Ein signifikanter Unterschied ergab sich zudem zwischen dem NEPS- und dem TIMSS-Test in der Häufigkeit der Verwendung von Graphen, Graphiken und Diagrammen. Studien haben gezeigt, dass die Verwendung von derartigen Repräsentationsformen beispielsweise schwachen Leserinnen und Lesern dabei helfen können die Aufgabe zu verstehen, wenn zur Lösung der Aufgabe lediglich das Verständnis der Repräsentationsform benötigt wird und der Text dadurch in den Hintergrund gerät (Schnotz, 2005; Prenzel et al., 2002). Bezüglich des Antwortformates lässt sich festhalten, dass sich der NEPS-Test signifikant vom Ländervergleichs-Test unterscheidet. Der NEPS-Test verwendet keine offenen Aufgabenformate und dafür mehr geschlossene Aufgabenformate als der Ländervergleich. Es ist davon auszugehen, dass sich dies auf die Schwierigkeit der Aufgabe auswirken kann. Da die Schülerinnen und Schüler bei offenen Antwortformaten selbstständig einen Text produzieren müssen, kann dieses Antwortformat beispielsweise für Schülerinnen und Schüler mit Schwierigkeiten in diesem Bereich Nachteile nach sich ziehen.

Zudem ergaben die Analysen auf der formalen Ebene, dass ein signifikanter Unterschied bei der Anzahl der Wörter sowie der Anzahl der Sätze pro Mathematikaufgabe im NEPS-Test und im Ländervergleichs-Test besteht. Der NEPS-Test hat signifikant mehr Wörter und mehr Sätze pro Aufgabenstellung als der Ländervergleich. Für schwache Leserinnen und Leser könnte sich hierdurch ein Nachteil ergeben. Studien haben herausgefunden, dass vor allem für Zweitsprachenlerner die Schwierigkeit der Aufgaben durch eine hohe Anzahl von Wörtern und Sätzen beeinflusst werden kann (Haag et al., 2013; Wolf & Leon, 2009; Wu, 2010). In den übrigen Aspekten hat sich eine hohe Übereinstimmung zwischen den Studien gezeigt.

Zusammenfassend lässt sich hinsichtlich der formalen Aspekte festhalten, dass der NEPS-Test mehr Ähnlichkeiten mit dem TIMSS-Test aufweist als mit dem Ländervergleichs-Test.

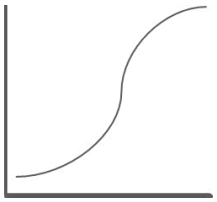
Hinsichtlich der sprachlichen Schwierigkeit der Mathematikaufgaben lassen sich zwischen dem NEPS-Test und dem Ländervergleichs- bzw. TIMSS-Test keine signifikanten Unterschiede aufzeigen. Daher sollten sich die Unterschiede in der sprachlichen Komplexität der Aufgaben auch nicht auf die weiteren Analysen und vor allem auf die Übertragung der Skalenmetriken auswirken. Tendenziell kann jedoch festhalten werden, dass die sprachliche Komplexität in allen drei Studien in der Aufgabenstellung größer ist als in den Stimuli oder in den Antwortalternativen. Zudem wurden viele Aufgaben sowohl hinsichtlich der Stimuli, der Aufgabenstellung als auch des Antworttexts auf der Wortebene als schwieriger eingestuft. Der Stimulus ist auf Wortebene im NEPS etwas komplexer als im Ländervergleich oder in TIMSS und auf Satzebene etwas leichter als im TIMSS-Test. Hinsichtlich der Wortebene unterscheiden sich die Aufgabenstellungen in den Studien nicht voneinander. Die Satzebene ist tendenziell etwas leichter im NEPS-Test als im TIMSS-Test und auf Textebene ist der NEPS-Test etwas schwieriger als der Ländervergleichs-Test. Der Antworttext wurden in den drei Studien bis auf eine Ausnahme auf Niveaustufe 0 eingestuft. Die Ländervergleichs-Aufgaben sind auf Wortebene etwas komplexer als die NEPS-Aufgaben.

Die gefundenen Gemeinsamkeiten zwischen den Mathematiktests liefern eine weitere Voraussetzung für ein stabiles und exaktes Linking. Jedoch wurden auch signifikante Unterschiede zwischen den Mathematikaufgaben der drei Studien festgestellt, die einen Einfluss auf das Linking haben könnten, da die Testleistungen der Schülerinnen und Schüler in den Mathematiktests durch die Unterschiede bedingt verschieden ausfallen können. Die Testinstrumente der drei Studien sind daher nicht als deckungsgleich und damit austauschbar anzusehen. Weitere Analysen hinsichtlich der dimensionalen und skalenbezogenen Vergleichbarkeit auf Grundlage einer gemeinsamen Stichprobe können hier weitere Hinweise zur Vergleichbarkeit liefern (vgl. Kapitel 5). Nur wenn sich hier große Zusammenhänge zwischen den Tests zeigen, ist eine Verlinkung sinnvoll.

Tabelle 4.21: Vergleich der formalen Merkmale und sprachlichen Komplexität in den Mathematikaufgaben der drei Studien

	TIMSS 2011 und NEPS	Ländervergleich 2011 und NEPS
formale Merkmale	TIMSS hat im Vergleich zu NEPS signifikant mehr Aufgaben - mit Graphen usw.	Der Ländervergleich hat im Vergleich zu NEPS signifikant mehr Aufgaben - mit einem Stimulus - mit Tabellen - mit offenem Antwortformat Der Ländervergleich hat im Vergleich zu NEPS signifikant weniger Aufgaben - mit geschlossenem Antwortformat Der Ländervergleich hat im Vergleich zu NEPS signifikant weniger - Wörter - Sätze
sprachliche Komplexität	Der Ländervergleich hat im Vergleich zu NEPS keine signifikanten Unterschiede	Der Ländervergleich hat im Vergleich zu NEPS keine signifikanten Unterschiede

4.5 Methodischer Vergleich



Dieses Kapitel soll zunächst einen Überblick über die in den Studien verwendeten Auswertungsverfahren der Mathematiktests geben. Anschließend werden die Auswertungsverfahren einander gegenübergestellt. Ziel ist es herauszustellen, ob die statistischen Methoden und die Skalierungsmodelle in der NEPS-Mathematikstudie gleich bzw. ähnlich wie in den Mathematikgrundschulstudien von TIMSS und dem Ländervergleich sind (Forschungsfrage 1e). Unterschiede zwischen den Studien sollen zudem hinsichtlich ihrer möglichen Auswirkungen auf die Ergebnisse hin beleuchtet werden.

Die Auswertung vom **TIMSS-Test** basiert auf dem Modell der Item Response Theorie. Nach Foy, Brossmann und Galia (2012) führte u. a. das implementierte Testheftdesign zu dieser Auswahl. In einem ersten Schritt wurden die Aufgaben auf ihre Messeigenschaften hin untersucht. Hierbei stellte sich heraus, dass zwei Mathematikaufgaben und sieben Naturwissenschaftsaufgaben die vorher festgelegten Kriterien (Martin & Mullis, 2012b) nicht erfüllen. Für die Schätzung der Itemparameter wurde ein mehrdimensionales 3-PL-Modell genutzt. Die Itemparameter werden mit der Software PARSCALE (Muraki & Bock, 1997) geschätzt. Für jedes Item wird hierbei ein Schwierigkeitsparameter (Location), ein Trennschärfeparameter (Slope) sowie ein Rateparameter (Guessing) bestimmt. Zudem wurde ein mehrdimensionales Antwortmodell modelliert, d. h. es gibt sogenannte polytome Aufgaben, die zusätzlich zu der Antwortkategorie ‚richtig' und ‚falsch' noch die Antwortkategorie ‚teilweise richtig' abbilden, so dass zusätzliche Schwellenparameter geschätzt werden können (vgl. hierzu u. a. Rost, 1996).

Für die Schätzung der Personenparameter wurde der Plausible-Value-Ansatz gewählt unter Verwendung der Software MGROUP (Sheehan, 1985). Insgesamt werden fünf PVs pro Schülerin bzw. pro Schüler gezogen. Zudem wurde für jede Schülerin und jeden Schüler, der an TIMSS teilgenommen hat, ein Gewicht berechnet (Wendt et al., 2012).

Für die erste TIMSS-Erhebung 1995 wurden ein Mittelwert von 500 Punkten sowie eine Standardabweichung von 100 festgelegt (Skalenmittelwert). Zusätzlich wird ein internationaler Mittelwert angegeben, der für jeden Erhebungszyklus neu berechnet wird. In die Berechnung des internationalen Mittelwertes fließen die Mittelwerte aller

Teilnehmerstaaten ein, ausgenommen sind jedoch die Benchmark-Teilnehmer. In TIMSS 2011 wurde so ein Mittelwert für Mathematik von 491 Punkten mit einer Standardabweichung von 81 berechnet und ein Mittelwert für Naturwissenschaften von 486 Punkten mit einer Standardabweichung von 85 (Wendt et al., 2012).

In TIMSS 2011 gab es zwei Aspekte, die zur Stichprobenverzerrung geführt haben und die bei der Berechnung der Stichproben- und Messfehler zu berücksichtigen waren: (1) die Stichprobenauswahl, beispielsweise bedingt durch die Anzahl der Klassen pro Klassenstufe in einer Schule, und (2) der Stichprobenausfall. Für eine korrekte Schätzung der Standardfehler trotz Cluster wurde das Jackknife-Verfahren gewählt (Wendt et al., 2012).

Der **Ländervergleich** nutzt für die Auswertung der Leistungsdaten das Rasch-Modell. Die Kompetenzwerte werden in einem vierstufigen Verfahren bestimmt: (1) Kalibrierung der Items, (2) Berechnung der Personenparameter, (3) Definition der Transformationsvorschrift und (4) Berechnung der Standardfehler. Die Schätzungen werden mit den Softwarepaketen ConQuest, der Software R und WesVar vorgenommen (Weirich et al., 2012).

Für die Bestimmung der Itemparameter wurde für jedes Fach ein gesondertes Modell geschätzt. Da die Mathematikitems zum Teil mehreren Kompetenzbereichen zugeordnet wurden, wurde diese Items zur Schätzung der Itemparameter mehrfach verwendet, so dass für diese Items auch zwei oder mehre Itemparameter geschätzt wurden. Zusätzlich wurde zur Bestimmung eines globalen Itemparameters für die Mathematikitems ein eindimensionales Modell geschätzt. Nach Weirich et al. (2012) wurde „Von einer gemeinsamen mehrdimensionalen Kalibrierung der Items […] abgesehen, da dies für Items mit Mehrfachladungsstruktur zu einer problematischen Interpretation der Itemparameter führen kann.“ Insgesamt wurden somit für den Bereich Mathematik sechs eindimensionale Modelle berechnet.

Die hier gewonnenen Itemparameter wurden für die Schätzung der Personenparameter fixiert. In jedem der sechs Modelle wurden 15 PVs gezogen. Für eine genauere Schätzung der PVs wurde ein Hintergrundmodell in die Analysen mit einbezogen, welches alle Variablen enthält, die für spätere Analysen benötigt werden. Insgesamt flossen elf Hintergrundvariablen (z.B. Geschlecht, Schulart, Alter und sozialer Hintergrund) in die Regressionsanalyse mit ein. Die fehlenden Werte in den Hintergrundvariablen wurden hierzu vorher imputiert (Weirich et al., 2012). Um Aussagen über die Population treffen zu können, wurde zudem für jede

Schülerin und jeden Schüler ein statistisches Gewicht berechnet. Durch dieses Verfahren wird die Repräsentativität der Ergebnisse gesichert. Anschließend wurde in einem dritten Schritt die Transformationsvorschrift bestimmt: Dies bedeutet, die Personenschätzer, die sich nach der Schätzung auf einer Logit-Metrik (PV_{Logit}) befinden, in besser interpretierbare Werte transformiert werden. Zunächst werden für jeden Kompetenzbereich ein gewichteter Mittelwert (M_g) sowie eine gewichtete Standardabweichung (SD_g) berechnet. Anschließend erfolgt die Transformation der PVs in die 500 Metrik mit einer Standardabweichung von 100 durch die folgende Vorschrift:

$$\mathrm{PV}_{\text{Ländervergleich}} = (\mathrm{PV}_{\text{Logit}} - \mathrm{M_g}) * \frac{100}{SDg} + 500$$

Für die Berechnung der Standardfehler wurde das Jackknife-Verfahren gewählt, um den Grad der Abweichung der Stichprobe von der Population zu berechnen. Hierbei wurde berücksichtigt, dass es sich bei der Stichprobe um eine stratifizierte Clusterstichprobe, also um die Ziehung ganzer Klassen statt einzelner Schülerinnen und Schüler, handelt (Weirich et al., 2012).

Die Auswertung der Leistungsdaten im **NEPS** erfolgt auf Grundlage des Rasch-Modells und des Partial Credit Modells mit der Computersoftware ConQuest (Wu, Adams & Wilson, 1998). Für jeden Kompetenzbereich wird eine eindimensionale Struktur der Daten angenommen. Für den Kompetenzbereich Mathematik wird diese durch die Schätzung eines mehrdimensionalen Modells – bezogen auf die Inhaltsdimensionen – überprüft. Die Modellfitwerte des mehrdimensionalen Modells sind etwas passgenauer als die des eindimensionalen Modells. Dennoch wird aufgrund der hohen Korrelationen zwischen den Inhaltsdimensionen im mehrdimensionalen Modell (.87 < r > .94), das eindimensionale Modell für die Auswertung bevorzugt (Duchhardt & Gerdes, 2012a). Zusätzlich wurden die Items hinsichtlich ihrer Fit-Statistiken und ihrer Testfairness (DIF) untersucht und ob sie den vorher definierten Kriterien genügen (Duchhardt & Gerdes, 2012a; Pohl & Carstensen, 2012). Alle Items weisen bei den Analysen akzeptable Fit-Statistiken und eine akzeptable Testfairness auf, so dass kein Item für die Auswertung der Leistungsdaten ausgeschlossen wird.

Für die Personen werden WLEs (Warm, 1989) berechnet. Die zusätzliche Ziehung von PVs soll in einem späteren Arbeitsschritt erfolgen (Duchhardt & Gerdes, 2012a; Pohl & Carstensen,

2012). Zusätzlich wurde für jede Schülerin und jeden Schüler unterschiedliche Gewichte berechnet (Skopek, Pink & Bela, 2012; Zinn, 2013).

Zwischenfazit

Die unterschiedlichen Herangehensweisen bei der Auswertung der Daten in den drei Studien sind insofern wichtig herauszustellen, als dass sie einen Einfluss auf die Ergebnisse und auf die Vergleichbarkeit haben. Bei dem methodischen Vergleich der drei Studien konnten viele Gemeinsamkeiten aber auch Unterschiede aufgezeigt werden. Tabelle 4.22 gibt einen Überblick zum methodischen Vergleich der Studien. Ein bedeutender Unterschied besteht bezüglich der Skalierung der Daten in den drei Studien. TIMSS skaliert die Leistungsdaten unter Annahme eines 3-PL Modells, wohingegen NEPS und der Ländervergleich ein 1-PL Modell annehmen. Der Unterschied zwischen dem 1-PL- und dem 3-PL-Modell ist, dass beim 3-PL-Modell sowohl die Itemtrennschärfe als auch ein Rateparameter mit modelliert werden. Beim 1-PL-Modell hingegen, werden der Diskriminationsparameter auf „1“ und der Rateparameter auf „0“ festgesetzt. Bezüglich der Skalierung unterschiedlicher Modelle konnte beispielweise in TIMSS 2007 gezeigt werden, dass sich bei einer Reskalierung der Leistungsdaten mit einem 1-PL-Modell die Rangfolgen der Länder bezüglich der mathematischen bzw. naturwissenschaftlichen Leistung nur marginal ändern im Vergleich zu einer Skalierung unter Annahme eines 3-PL-Modells. Hinsichtlich der Reskalierung des Mathematik- und Naturwissenschaftstests hat sich gezeigt, dass sich die Rangplatzverschiebungen, bedingt durch die Reskalierung der Daten unter Annahme eines 1-PL-Modells, nur zwischen Ländern ergaben, deren Leistungskennwerte nicht signifikant unterschiedlich waren (Bonsen et al., 2008). Diese Ergebnisse decken sich mit den Resultaten einer Studie von Brown, Micklewright, Schnepf und Waldmann (2005). Brown et al. fanden heraus, dass die Nutzung unterschiedlicher IRT Modelle robust ist hinsichtlich deskriptiver Kennwerte (also zentraler Tendenzen), jedoch nicht für die Verteilung der Ergebnisse. Insofern sollte die Interpretation der Linkingergebnisse nicht auf Individualebene erfolgen, sondern nur auf Gruppenebene, da sich durch die unterschiedlichen Skalierungen Rangplatzverschiebungen ergeben können.

Es gibt jedoch noch weitere Unterschiede in den Modellannahmen der drei Studien. TIMSS skaliert die Leistungsdaten in einem mehrdimensionalen Modell unter Annahme einer Between-Item-Dimensionalität. Im Ländervergleichs-Test werden sowohl eine globale

mathematische Kompetenz als auch jeweils für jeden mathematischen Inhaltsbereich einen Kompetenzwert (bzw. 15 PVs pro Schülerin und Schüler) berechnet, wobei die Items zum Teil mehreren Kompetenzbereichen zugeordnet sind. NEPS hingegen geht von einer eindimensionalen Struktur der Daten aus. TIMSS und NEPS skalieren zudem ein Partial Credit Modell wohingegen der Ländervergleichs-Test keine Items verwendet für die Schwellenparameter skaliert werden müssten. Bezüglich der Between- vs. Within-Item-Dimensionalität können Winkelmann und Robitzsch; (2009 in einer Analyse zeigen, dass die Korrelationsmuster zwischen den Teildimensionen unterschiedlich hoch bzw. niedrig ausfallen, je nachdem welche Dimensionalität der Datenstruktur angenommen wird. Für die Gegenüberstellung verwenden sie die Daten der Normierungs- und Pilotierungsstichprobe des Ländervergleichs Mathematik in der Grundschule. Im Vergleich zeigt sich, dass die Skalierung der Daten unter Annahme einer Within-Item-Dimensionalität etwas niedrigere Korrelationen aufweisen als unter Annahme einer Between-Item-Dimensionalität. Robitzsch; (2009 zeigt weiterhin in einer Simulationsstudie, mit welchen Konsequenzen zu rechnen ist, wenn die Daten zwar Between-Item-Dimensionalität aufweisen, jedoch ein Modell mit Within-Item-Dimensionaltität spezifiziert wird und umgekehrt. Er konnte zeigen, dass die Korrelationen in einem 1-PL-Modell deutlich überschätzt werden, wenn die Datenstruktur Within-Item-Dimensionalität aufweisen, jedoch ein Modell mit Between-Item-Dimensionalität skaliert wird. Umgekehrt werden die Korrelationen in der Simulationsstudie unterschätzt. Eine unverzerrte Schätzung könne jedoch mit einem 2-PL-Modell unter Annahme einer Within-Item-Dimensionalität erreicht werden. Dieses Wissen ist vor allem für den dimensionalen Vergleich relevant und wird daher in Kapitel 5.1 diskutiert.

Die Personenparameter werden in allen drei Studien mit dem PV-Ansatz skaliert. Ein Unterschied besteht lediglich hinsichtlich der Anzahl der geschätzten PVs. Im NEPS werden zusätzlich zu den PVs noch WLEs geschätzt. Zudem geben die drei Studien Gewichte für jede Schülerin bzw. jeden Schüler an

Eine Linking-Studie unter Verwendung des IRT-Ansatzes ist in diesem Fall schwierig, weil die unterschiedliche Modellierung der Daten bei einer gemeinsamen Skalierung nicht zu berücksichtigen bzw. mit Fehlern behaftet ist. Die Methode des Equipercentilen Equatings kann die unterschiedlichen Modellierungen berücksichtigen. Dennoch sind die oben genannten Unterschiede bei der Interpretation der Ergebnisse u. a. beim dimensionalen

Vergleich zu berücksichtigen, da die Unterschiede in der Skalierung zu einer Über- bzw. Unterschätzung der Zusammenhänge führen können. Darüber hinaus bleibt festzuhalten, dass das Linking eindimensional skaliert wird, was zur Folge hat, dass nicht die Komplexität, die die Daten von TIMSS und dem Ländervergleich eigentlich zulassen, abgebildet wird (vgl. Kapitel 5).

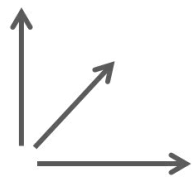
Tabelle 4.22: Methodischer Vergleich der Studien

Skalierungsmodelle	**TIMSS**	**Ländervergleich**	**NEPS**
zugrundeliegende Testtheorie	IRT	IRT	IRT
1-PL- vs. 3-PL Modell	3-PL Modell	1-PL Modell	1-PL Modell
ein- vs. mehrdimensional	mehrdimensional	6 eindimensionale Modelle	eindimensional
Between- vs. Within Item Dimensionality	Between Item Dimensionalität	Within-Item-Dimensionalität	Between-Item-Dimensionalität
Partial Credit Modell	ja	nein	ja
Bestimmung der Personenparameter	5 plausible values	15 plausible values	WLEs & PVs
Berechnung von Gewichten	ja	ja	ja

5 Dimensionale und skalenbezogene Zusammenhänge

Neben dem inhaltlichen Vergleich soll auch ein Vergleich auf Ebene der empirischen Daten der Linking-Studie (vgl. Kapitel 3.1) erfolgen, d. h. es soll überprüft werden, ob sich hohe empirische Zusammenhänge zwischen den Mathematiktests der NEPS K5 Studie und den Studien TIMSS K4 und Ländervergleich Primar zeigen (Forschungsfrage 2). Ziel ist es, differenziertere Aussagen darüber treffen zu können, ob bzw. in welchem Maße die drei Tests ein gleiches Konstrukt mathematischer Kompetenz messen. Die Analysen unterteilen sich in zwei Bereiche. Zum einen soll untersucht werden, ob eine dimensionale Äquivalenz der drei Tests vorliegt (Forschungsfrage 2a) und zum anderen ob eine skalenbezogene Äquivalenz gegeben ist (Forschungsfrage 2b). Der Vergleich auf dimensionaler Ebene erfolgt hinsichtlich der faktoriellen Struktur der Tests (vgl. Kapitel 5.1). Der Vergleich auf skalenbezogener Ebene soll die Frage beantworten, ob – bei gleicher Stichprobe – die Personenfähigkeiten, die mit den Tests geschätzt wurden, zwischen den Studien äquivalent sind, d.h. eine Linearität zwischen den Schülerinnen und Schülern in den unterschiedlichen Tests besteht. (vgl. Kapitel 5.2).

5.1 Dimensionaler Vergleich



Der zweite Aspekt, der neben der konzeptionellen Äquivalenz nach van de Vijver (1998) gegeben sein sollte, ist die dimensionale Äquivalenz der Testinstrumente. Eine dimensionale Äquivalenz wird in dem vorliegenden Fall dann als gegeben erachtet, wenn – bei gleicher Stichprobe – die faktorielle Struktur des NEPS-Mathematiktests K5 sehr ähnlich bzw. gleich ist, wie in den Mathematiktests von TIMSS K4 und vom Ländervergleich Primar. Die faktorielle Struktur der Tests soll anhand der Korrelationen der Teildimensionen in den drei Tests erfolgen. Hierbei wird jedoch eine Beschränkung auf die Gegenüberstellung der Inhaltsbereiche vorgenommen, da die Zuordnungen zu den anderen Bereichen (kognitive Anforderungsbereich und prozedurale Fähigkeiten) nicht für alle Studien verfügbar sind. Die Frage ist, ob sich die inhaltlichen Gemeinsamkeiten und Unterschiede, die bereits in Kapitel 4 aufgezeigt wurden, auch empirisch nachweisen lassen. Hierzu werden zwei unterschiedliche Herangehensweisen genutzt. Zum einen sollen mit Hilfe der Gegenüberstellung der Korrelationen der Teildimensionen im NEPS-Mathematiktest K5 mit den Korrelationen der Teildimensionen in

den Mathematiktests von TIMSS K4 und des Ländervergleichs Primar Unterschiede und Gemeinsamkeiten aufgezeigt werden (Forschungsfrage 2a, Teil 1). Zum anderen soll überprüft werden, ob hohe Korrelationen zwischen den Teildimensionen der Tests und damit hohe Überschneidungen in den Teildimensionen der Tests aufgezeigt werden können (Forschungsfrage 2b, Teil 2).

Für die Überprüfung der dimensionalen Äquivalenz werden die Daten der Linking-Studie sowohl getrennt (Forschungsfrage 2a, Teil 1) als auch gemeinsam skaliert (Forschungsfrage 2a, Teil 2). Um eine bessere Vergleichbarkeit der Ergebnisse sicherzustellen, wurden die Daten jeweils möglichst mit derselben Skalierungsmethode geschätzt. Die nachfolgenden Berechnungen erfolgen mit einem mehrdimensionalen 1-PL-Modell, welches je nach Studie auch Items enthält, die mit einem Partial Credit Modell skaliert werden. An der Entscheidung der Between- bzw. Within-Item-Dimensionalität in den Originalstudien wird festgehalten, da hier jeweils nur die Einfach- bzw. Mehrfachzuordnungen der Items vorhanden sind. Gewichte werden nicht genutzt und es fließt kein Hintergrundmodell in die Analysen mit ein. Die Berechnungen erfolgen mit der Computersoftware Acer ConQuest (Wu et al. 2007), wobei die Itemparameter frei geschätzt werden.

5.1.1 Korrelationen der Teildimensionen innerhalb der Tests

Um die faktorielle Struktur bezüglich der Korrelationen der Teildimensionen innerhalb der Tests zu untersuchen, werden die Zusammenhänge zwischen den Inhaltsbereichen der drei Studien miteinander verglichen. Hierfür werden drei separate mehrdimensionale Rasch-Modelle geschätzt.

TIMSS unterscheidet drei Inhaltsbereiche (vgl. Kapitel 4.4.1): *Arithmetik, Geometrie und Messen sowie Daten* (Selter et al., 2012). Die Daten wurden unter Verwendung des Partial Credit Ansatzes und unter der Annahme der Between-Item-Dimensionalität skaliert. In der Tabelle 5.1 sind unterhalb der Diagonalen die latenten Korrelationen zwischen den Inhaltsbereichen dargestellt. Am höchsten korreliert der Inhaltsbereich *Geometrie und Messen* mit *Arithmetik* (r = .85). Etwas niedriger sind die latenten Korrelationen zwischen *Daten* und *Arithmetik* sowie zwischen *Daten* und *Geometrie und Messen*. Die Unterschiede in der Höhe der Koeffizienten sind jedoch marginal. Oberhalb der Diagonalen werden die latenten Korrelationen bei zufälliger Zuordnung zu den Inhaltsbereichen angegeben

(randomisierte Itemklassifikation). Die zufälligen Zuordnungen liefern, nach Winkelmann und Robitzsch (2009, S. 187), „Obergrenzen für die erwartbaren Korrelationen unter der Annahme der Eindimensionalität“. Es zeigt sich, dass die randomisierte Zuordnung zu etwas höheren Korrelationen führt, als die originale Zuordnung zu den Inhaltsbereichen ($.89 < r < .92$). Die Inhaltsbereiche lassen sich demnach separieren, der Anteil an gemeinsamer Varianz ist jedoch sehr hoch.

Tabelle 5.1: Korrelationen zwischen den TIMSS-Inhaltsbereichen

Dimension	Arithmetik	Geometrie & Messen	Daten
Arithmetik	1.00	.92	.90
Geometrie & Messen	.85	1.00	.89
Daten	.83	.83	1.00

Der **Ländervergleich Mathematik Primar** unterscheidet fünf Inhaltsbereiche (vgl. hierzu auch Kapitel 4.4.1): *Zahlen und Operationen*, *Geometrie und Messen, Muster und Strukturen, Raum und Form sowie Daten, Häufigkeiten und Wahrscheinlichkeiten* (Roppelt & Reiss, 2012). Bei der Skalierung wurde eine Mehrfachzuordnung der Items zu den Inhaltsbereichen berücksichtigt (Within-Item-Dimensionalität). Die latenten Korrelationen zwischen den Inhaltsbereichen werden unterhalb der Diagonale angegeben und liegen zwischen .48 und .70 (vgl. Tabelle 5.2).

Tabelle 5.2: Korrelation zwischen den Ländervergleichs-Inhaltsbereichen

Dimension	Zahlen & Operationen	Geometrie & Messen	Muster & Strukturen	Raum & Form	Daten, Häufigk. & Wahr-scheinlichk.
Zahlen & Operationen	1.00	.78	.82	.84	.80
Geometrie & Messen	.69	1.00	.80	.86	.83
Muster & Strukturen	.50	.51	1.00	.80	.82
Raum & Form	.63	.64	.53	1.00	.83
Daten, Häufig-keiten & Wahr-scheinlichkeiten	.61	.65	.48	.70	1.00

Am höchsten korreliert der Inhaltsbereich *Daten, Häufigkeiten und Wahrscheinlichkeiten* mit *Raum und Form* (r = .70). Der kleinste Koeffizient ergibt sich zwischen den Inhaltsbereichen *Daten, Häufigkeiten und Wahrscheinlichkeiten* und *Muster und Strukturen* (r = .48). Oberhalb der Diagonalen sind die latenten Korrelationen zwischen den zufällig zugeordneten Inhaltsbereichen angegeben. Die Korrelationen bei der Zufallszuordnung sind deutlich höher, als bei der originalen Zuordnung zu den Inhaltsbereichen (.78 < r < .86). Dies lässt den Schluss zu, dass sich die Inhaltsbereiche des Ländervergleichs separieren lassen, es jedoch auch einen Anteil an gemeinsamer Varianz gibt.

Im **Nationalen Bildungspanel** werden vier Inhaltsbereiche unterschieden (vgl. hierzu auch Kapitel 4.4.1): *Quantität, Veränderung und Beziehungen, Raum und Form* sowie *Daten und Zufall* (Ehmke et al., 2009). Die Daten wurden unter Annahme der Between-Item-Dimensionalität in einem mehrdimensionalen Modell geschätzt. Der Tabelle 5.3 sind die latenten Korrelationen zwischen den vier Inhaltsbereichen zu entnehmen (unterhalb der Diagonalen). Die Korrelationen liegen zwischen .88 und .93. Am höchsten korrelieren die beiden Inhaltsbereiche *Veränderung und Beziehungen* und *Daten und Zufall* (r = .93). Am niedrigsten korrelieren die Inhaltsbereiche *Daten und Zufall* mit *Quantität* sowie *Veränderung und Beziehungen mit Quantität* (*r* = .88). Oberhalb der Diagonalen sind die latenten Korrelationen dargestellt, die bei einer zufälligen Zuordnung der Items zu den Inhaltsbereichen berechnet wurden. Die Analysen zeigen, dass die Korrelationen bei der randomisierten Zuordnung tendenziell etwas höher liegen, dass jedoch der Unterschied zwischen der randomisierten Zuordnung und der inhaltlichen Zuordnung nicht sonderlich hoch ist.

Tabelle 5.3: Korrelation zwischen den NEPS-Inhaltsbereichen

Dimension	Quantität	Veränderung & Beziehungen	Raum & Form	Daten & Zufall
Quantität	1.00	.94	.96	.93
Veränderung & Beziehungen	.88	1.00	.93	.94
Raum & Form	.92	.92	1.00	.94
Daten & Zufall	.88	.93	.91	1.00

Zwischenfazit

Problematisch für einen Vergleich der faktoriellen Struktur der Items anhand der Korrelationen der Teildimensionen innerhalb der Tests ist die unterschiedliche Skalierung der Daten. Wie bereits in Kapitel 4.5 aufgezeigt, kann beispielsweise die Annahme einer Between bzw. Within-Item-Dimensionalität Einfluss auf die Höhe der latenten Korrelationen haben. Die Daten des NEPS- und des TIMSS-Tests wurden unter Annahme der Between-Item-Dimensionalität skaliert. Falls den Daten jedoch eigentlich eine Within-Item-Dimensionalität zu Grunde läge, würden die latenten Korrelationen überschätzt werden. Andersherum verhält es sich mit den Ergebnissen des Ländervergleichs-Test. Hier wird eine Within-Item-Dimensionalität angenommen. Läge aber stattdessen eine Between-Item-Dimensionalität vor, würden die latenten Korrelationen unterschätzt werden (Robitzsch, 2009). Zudem haben Winkelmann und Robitzsch (2009) in einer Untersuchung gezeigt, dass die latenten Korrelationen zwischen den inhaltsbezogenen Kompetenzbereichen in der Normierungs- und Pilotierungsstichprobe des Ländervergleichs Mathematik Primar unter Annahme einer Within-Item-Dimensionalität niedriger ausfallen als unter Annahme einer Between-Item-Dimensionalität. Dieses könnte die Unterschiede in der Höhe der Korrelationen erklären, die sich bei einer Gegenüberstellung der Studien zeigt. Der Ländervergleichs-Test weist deutlich niedrigere Korrelationen der Inhaltsbereiche auf ($.48 < r < .70$) als der TIMSS- und der NEPS-Test (TIMSS: $.83 < r < .85$; NEPS: $.88 < r < .93$), wobei die latenten Korrelationen des NEPS-Tests im Vergleich zum TIMSS-Test tendenziell noch höher ausfallen. Dies könnte daran liegen, dass der NEPS-Test nur wenige Items pro Inhaltsbereich hat und sich dadurch die Inhaltsbereiche eventuell nicht eindeutig voneinander abgrenzen lassen. Bedingt durch diese und weitere mögliche methodische Artefakte, die beispielsweise durch die unterschiedliche Skalierung und die Anlage der Studien entstehen können, bleibt dieser Vergleich der faktoriellen Struktur auf einer oberflächlichen Ebene und kann nur Tendenzen abbilden. Mit einer gemeinsamen Skalierung des NEPS-Tests mit dem TIMSS- und dem Ländervergleichs-Test soll ein weiterer Vergleich der faktoriellen Struktur auf Ebene der Inhaltsbereiche erfolgen. Die Ergebnisse werden im folgenden Kapitel dargestellt.

Zusammenfassend kann festgehalten werden, dass die latenten Korrelationen zwischen den Inhaltsbereichen in den drei Mathematiktests – vor allem im TIMSS- und im NEPS-Test – sehr hoch sind und damit einen hohen Anteil gemeinsamer Varianz haben. Dennoch zeigt sich, dass die Inhaltbereiche zusätzlich zu einer globalen, stoffgebietsunabhängigen

mathematischen Kompetenz ihren Anteil beitragen. Die faktorielle Struktur des NEPS-Mathematiktests weist dabei höhere Ähnlichkeiten mit dem TIMSS-Test als mit dem Ländervergleichstest auf. Eine Aussage darüber zu treffen, ob in den drei Studien die jeweiligen Pendants ähnlich hoch bzw. niedrig mit anderen Inhaltsbereichen korrelieren, lässt sich nicht treffen, da die Unterschiede in der Höhe der Korrelationen zwischen den Inhaltsbereichen im NEPS und in TIMSS nur klein sind.

5.1.2 Korrelationen der Teildimensionen zwischen den Tests

Für die Berechnung latenter Korrelationen der Teildimensionen zwischen den Tests werden die Daten von NEPS jeweils mit den Daten vom Ländervergleich und von TIMSS ebenfalls mit einem mehrdimensionalen Rasch-Modell gemeinsam skaliert. Sowohl das neun-dimensionale Modell für NEPS und den Ländervergleich als auch das sieben-dimensionale Modell für NEPS und TIMSS konnten nicht stabil geschätzt werden, so dass die Inhaltsbereiche von NEPS einzeln mit den fünf Inhaltsbereichen des Ländervergleichs bzw. mit den drei Inhaltsbereichen von TIMSS in einem mehrdimensionalen Rasch-Modell geschätzt werden.

Gemeinsame Skalierung des NEPS- und TIMSS-Test

Für einen Vergleich der faktoriellen Struktur der Tests wurden die TIMSS- und die NEPS-Daten zusammen in einem mehrdimensionalen Rasch-Modell skaliert. Die dunkelgrau hinterlegten Zellen in der Tabelle 5.4 geben an, welche Zusammenhänge zwischen den Inhaltsbereichen auf einer theoretischen Ebene angenommen werden (vgl. hierzu Kapitel 4.4.1). Die hellgrau hinterlegten Zellen wurden anhand des ersten Expertenreviews ermittelt, bei dem die Experten die NEPS-Items in die Rahmenkonzeptionen von TIMSS eingeordnet haben (vgl. hierzu Kapitel 4.4.2). Wurden in dem Expertenreview mindestens zwei NEPS-Items eines Inhaltsbereiches einem anderen Inhaltsbereich zugeordnet, wurden die Zellen in der Tabelle 5.4 hellgrau hinterlegt. Das Expertenreview ergab, dass die Inhaltsbereiche in den beiden Studien, die zwar ähnlich benannt sind und theoretisch Ähnliches messen, nicht umfassend übereinstimmen. Einige der NEPS-Items wurden in der TIMSS-Rahmenkonzeption einem anderen Inhaltsbereich zugeordnet als theoretisch angenommen. Die fett markierten Korrelationen in Tabelle 5.4 zeigen die höchste Korrelation pro Zeile an. Es zeigt sich, dass der NEPS-Inhaltsbereich *Quantität* am höchsten mit dem TIMSS-Inhaltsbereich *Geometrie und Messen* korreliert (r = .84) sowie mit dem Inhaltsbereich *Arithmetik* (r = .81). Der NEPS-

Inhaltsbereich *Veränderung und Beziehungen* korreliert am höchsten mit *Arithmetik* (r = .79) sowie mit *Geometrie und Messen* (r = .78). *Raum und Form* (NEPS) korreliert am höchsten mit dem TIMSS-Inhaltsbereich *Geometrie und Messen* (r = .91) und *Daten und Zufall* (NEPS) korreliert mit allen drei TIMSS-Inhaltsbereichen ähnlich hoch (.82 < r < .84).

Tabelle 5.4: Korrelation der Inhaltsbereiche zwischen NEPS und TIMSS

		TIMSS		
	Dimension	**Arithmetik**	**Geometrie & Messen**	**Daten**
NEPS	**Quantität**	.81	**.84**	.73
	Veränderung & Beziehungen	**.79**	.78	.71
	Raum & Form	.85	**.91**	.77
	Daten & Zufall	**.84**	.83	.82

Gemeinsame Skalierung des NEPS- und Ländervergleichs-Test

Für den Vergleich der faktoriellen Struktur des NEPS- und des Ländervergleichs-Tests wurden die Daten der beiden Studien in einem mehrdimensionalen Rasch-Modell zusammen skaliert (der Ländervergleichs-Test wiederum unter Annahme einer Within-Item-Dimensionalität). Die Markierung der Zellen in der Tabelle 5.5 ist identisch mit der des Vergleichs von NEPS und TIMSS. Die dunkelgrau hinterlegten Zellen geben Zusammenhänge

Tabelle 5.5: Korrelation der Inhaltsbereiche zwischen NEPS und dem Ländervergleich

		Ländervergleich				
		Zahlen & Operationen	**Größen & Messen**	**Muster & Strukturen**	**Raum & Form**	**Daten, Häufigkeiten & Wahrscheinlichkeiten**
NEPS	**Quantität**	.65	.69	.42	**.80**	.72
	Veränderung & Beziehungen	.64	**.73**	.52	.68	.64
	Raum & Form	.61	.79	.43	**.69**	.61
	Daten & Zufall	.63	**.74**	.43	**.74**	.71

auf einer theoretischen Ebene an (vgl. Kapitel 4.4.1) und die hellgrau hinterlegten Zellen wurden anhand des ersten Expertenreviews ermittelt (vgl. Kapitel 4.4.2). Bei den hellgrau hinterlegten Zellen wurden von den Experten mindestens zwei NEPS-Items eines Inhaltsbereiches einem anderen Inhaltsbereich als theoretisch angenommen der

Rahmenkonzeption des Ländervergleichs zugeordnet. Die Korrelationen liegen zwischen r = .42 und r = .80. Die höchsten latenten Korrelationen sind zwischen den Inhaltsbereichen *Quantität* (NEPS) und *Raum und Form* (LV) (r = .80), *Veränderung und Beziehungen* (NEPS) und *Größen und Messen* (LV) (r = .73), *Raum und Form* (NEPS) und *Raum und Form* (LV) (r = .69) sowie *Daten und Zufall* (NEPS) mit *Größen und Messen* sowie *Raum und Form* (LV) (r = .74).

Zwischenfazit

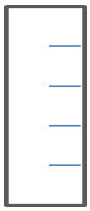
Durch die gemeinsame Skalierung der Daten von NEPS und TIMSS sowie NEPS und Ländervergleich konnte ein direkter Vergleich der faktoriellen Struktur der Tests erfolgen. Bezüglich des Vergleichs von NEPS und TIMSS hat die Analyse der Frameworks ergeben (vgl. Kapitel 4.4.1), dass hohe Übereinstimmungen zwischen dem TIMSS-Inhaltsbereich *Arithmetik* und *Quantität* sowie *Veränderung und Beziehungen* vorliegen. Durch das Expertenreview I (vgl. Kapitel 4.4.2) wurden darüber hinaus hohe Übereinstimmungen zwischen dem NEPS-Inhaltsbereich *Quantität* und *Geometrie und Messen* (TIMSS) aufgezeigt. Die latenten Korrelationen zwischen den Inhaltsbereichen bestätigen dieses Bild. Weiterhin ergaben sich hohe Korrelationen zwischen *Raum und Form* bei NEPS und *Geometrie und Messen* bei TIMSS. Dieses Ergebnis bestätigt ebenfalls die theoretischen Annahmen sowie die Ergebnisse des Expertenreviews I. Beim Inhaltsbereich *Daten und Zufall* liegen nur marginal unterschiedliche Korrelationen mit den TIMSS-Inhaltsbereichen vor. Die höchsten Korrelationen wären hier zwischen dem Inhaltsbereich *Daten und Zufall* (NEPS) und *Daten* (TIMSS) zu erwarten gewesen.

Die Korrelationen zwischen den Inhaltsbereichen in NEPS und im Ländervergleich sind im Vergleich etwas geringer, was durch die Within-Item-Dimensionalität des Ländervergleich-Tests bedingt sein kann. Darüber hinaus zeigt sich, dass die Inhaltsbereiche *Raum und Form* im NEPS und im Ländervergleich hoch miteinander korrelieren. Der Inhaltsbereich *Veränderung und Beziehungen* (NEPS) korreliert am höchsten mit *Größen und Messen* (LV). Die Ergebnisse entsprechen damit den Erwartungen. Beim Inhaltsbereich *Quantität* sind die Ergebnisse hingegen nicht erwartungskonform. Die höchsten Korrelationen wäre mit den Inhaltsbereichen *Zahlen und Operationen* sowie *Größen und Messen* zu erwarten gewesen (vgl. Kapitel 4.4.1). Im NEPS-Inhaltsbereich *Daten und Zufall* zeigt sich, dass dieser Inhaltsbereich am höchsten mit den Inhaltsbereichen *Größen und Messen* sowie *Raum und Form* korreliert, dies entspricht wiederum nicht den Erwartungen. Die Korrelationen des

Inhaltsbereichs *Muster und Strukturen* (LV) mit den Inhaltsbereichen des NEPS-Tests fallen insgesamt deutlich niedriger aus als die Korrelationen zwischen den anderen Inhaltsbereichen. Bereits das Expertenreview hat gezeigt, dass sowohl NEPS-Items aus dem Inhaltsbereich *Quantität*, *Daten und Zufall* sowie *Veränderung und Beziehungen* diesem Inhaltsbereich zusätzlich zugeordnet wurden und damit keine große Übereistimmung mit einem einzelnen Inhaltsbereich festgestellt werden konnte.

Zusammenfassend betrachtet lässt sich bezüglich der latenten Korrelationen der Inhaltsbereiche zwischen den Tests – wie bereits bei den Korrelationen innerhalb der Tests – ein größerer Zusammenhang des NEPS-Tests mit dem TIMSS-Test feststellen. Obwohl die Inhaltsbereiche hohe Übereinstimmungen aufweisen, sind die Testinstrumente jedoch nicht als gegenseitig austauschbar zu betrachten. Bei beiden Vergleichen (NEPS-TIMSS und NEPS-Ländervergleich) wird ersichtlich, dass hinsichtlich der latenten Korrelationen die angenommenen Überschneidungen der NEPS-Inhaltsbereiche *Veränderung und Beziehungen* sowie *Raum und Form* mit den jeweiligen Pendants der TIMSS- und Ländervergleich-Inhaltsbereiche aufgezeigt werden können. Beim Vergleich der faktoriellen Struktur konnte der angenommene Zusammenhang des NEPS-Inhaltsbereichs *Quantität* mit den Ländervergleichs-Inhaltsbereichen *Zahlen und Operationen* sowie *Größen und Messen* jedoch nicht nachgewiesen werden. Im Inhaltsbereich *Daten und Zufall* lässt sich die angenommene Entsprechung weder in TIMSS noch im Ländervergleich aufzeigen. Bei den Vergleichen ist jedoch zu berücksichtigen, dass der NEPS-Test nur aus insgesamt 24 Items besteht. Dies bedeutet, dass jeder Inhaltsbereich nur durch 5 bis 8 Items repräsentiert wird. Wie bereits bei den Analysen zur faktoriellen Struktur innerhalb der Tests deutlich wurde, lassen sich die Inhaltsbereiche im NEPS-Test empirisch nicht gut separieren. Der NEPS-Test ist zudem auch nicht dazu entwickelt wurden, unterschiedliche Kompetenzen der Schülerinnen und Schüler auf den Inhaltsbereichen zu berechnen, sondern der NEPS-Test wurde entwickelt, um eine eindimensionale globale mathematische Kompetenz zu erfassen.

5.2 Skalenbezogener Vergleich



Zusätzlich zu dem Vergleich auf dimensionaler Ebene soll an dieser Stelle die Vergleichbarkeit bezüglich der Skalen erfolgen. Nach van de Vijver (1998) und Pietsch et al. (2009) gilt es diesbezüglich zu prüfen, ob die geschätzten

Personenfähigkeiten in den drei Tests nahezu äquivalent sind. Wenn eine Linearität zwischen den Schülerinnen und Schülern in den unterschiedlichen Tests besteht, sind die drei Tests auf der Ebene der Skalen vergleichbar. Diese Vergleichbarkeit soll bezüglich unterschiedlicher Aspekte untersucht werden. Eine der Grundvoraussetzungen für ein Equating ist, dass die Reliabilitäten der Tests gleich sind (vgl. Kapitel 1.1.1). Im Rahmen dieser Arbeit soll zwar aufgrund der bereits gefundenen Interschiede in den Testkonzeptionen kein Equating erfolgen, dennoch können die Reliabilitäten der jeweiligen Tests einen Einfluss auf die Güte des Linking haben (Dorans & Holland, 2000). Daher soll in Kapitel 5.2.1 zunächst eine Darstellung und Gegenüberstellung der Testreliabilitäten erfolgen.

Anschließend sollen die Zusammenhänge zwischen den drei Tests anhand latenter Korrelationen beschrieben werden. Nach Kolen und Brennan (2010) sowie Dorans und Holland (2000) kann ein Korrelationskoeffizient in einem SG-Design dazu verwendet werden, Aussagen über die Vergleichbarkeit von Tests zu treffen. Die Ergebnisse liefern weitere Hinweise darauf, ob es sich bei den Tests um ein gemeinsames Konstrukt mathematischer Kompetenz handelt oder ob die Tests ein unterschiedliches Konstrukt mathematischer Kompetenz abbilden.

Es soll zunächst überprüft werden, ob es sich bei den Tests um unterschiedliche Konstrukte mathematischer Kompetenz handelt oder ob die Tests ein gemeinsames Konstrukt mathematischer Kompetenz abbilden. Wenn, wie in dem vorliegenden Fall, ein SG-Design verwendet wird und die Daten IRT skaliert werden, kann ein Korrelationskoeffizient dazu genutzt werden, Aussagen über die Vergleichbarkeit von zwei Tests zu treffen (Kolen & Brennan, 2010; Dorans & Holland, 2000). Zudem ist es im IRT-Kontext möglich zu überprüfen, ob eher ein ein- oder ein zweidimensionales Modell die Daten der beiden Tests besser abbildet. Ein Vergleich der Modellgüte eines eindimensionalen oder eines zweidimensionalen Modells, d.h. entweder werden der NEPS- und der TIMSS-Test eindimensional oder zweidimensional modelliert, soll zusätzlich zur Klärung der Frage beitragen (vgl. Kapitel 3.6.2.3).

5.2.1 Reliabilitäten

Eine Voraussetzung für ein Equating ist, dass die Teste eine adäquate und eine möglichst ähnliche Reliabilität aufweisen (vgl. Kapitel 1.1.1; Dorans & Holland, 2000; Holland & Dorans, 2006; Dorans & Walker, 2007). In der vorliegenden Arbeit soll zwar kein Equating, sondern ein

Linking der Tests erfolgen, dennoch kann auch bei einem Linking die Reliabilität der Tests einen Einfluss auf die Güte des Linking, beispielsweise auf die Invarianz über Subgruppen, haben (Dorans & Holland, 2000). Einen Einfluss auf die Güte hat dabei nicht nur die Ähnlichkeit, sondern auch die Höhe der Reliabilitäten. Daher werden im Folgenden die Tests hinsichtlich der Höhe und der Ähnlichkeit ihrer Reliabilitäten hin verglichen. Die Reliabilitätskoeffizienten sind dabei den Ergebnisberichten der jeweiligen Studien entnommen wurden.

TIMSS

Die Reliabilität des internationalen TIMSS 2011 Mathematiktest wird bei Foy, Martin, Mullis und Stanco (2012) mit dem Koeffizienten Cronbach's Alpha angegeben. Der Test hat international eine Reliabilität von im Median $\alpha = .82$ für die vierte Jahrgangsstufe. National liegt die Reliabilität des TIMSS-Mathematiktests in Deutschland bei $\alpha = .81$.

Ländervergleich

Für den Ländervergleichs-Mathematiktest Primar liegen zum jetzigen Zeitpunkt keine Informationen über die Reliabilität vor.

NEPS

Für den NEPS 2010 Mathematiktest für die fünfte Jahrgangsstufe wurde eine EAP/PV Reliabilität berechnet (Duchhardt & Gerdes, 2012b). Die EAP/PV Reliabilität für den NEPS-Test liegt bei .80.

Zwischenfazit

Zusammenfassend lässt sich festhalten, dass die Reliabilitäten des NEPS- und TIMSS-Mathematiktests in einem zuverlässigen Bereich liegen. Am höchsten ist die Messgenauigkeit beim TIMSS-Test ($\alpha = .81$) und nur leicht geringer beim NEPS-Test mit .80. Die Reliabilitätsmaße für den NEPS- und den TIMSS-Test sind sich damit hinreichend ähnlich und entsprechen somit der Vorgabe gleicher Reliabilitäten für ein Equating. Jedoch geben Dorans und Holland (2000) zu bedenken, dass auch die Höhe der Reliabilität Einfluss auf das Equating haben kann. Bezüglich des TIMSS- und des NEPS-Tests lässt sich festhalten, dass die Höhe der Reliabilitäten ein gewisses Maß an Messungenauigkeit vermuten lässt. Diese Messungenauigkeit kann folglich einen Einfluss auf die Güte des Linking haben. Beispielsweise

kann die Höhe der Reliabilitäten die Zuordnungsgenauigkeit (Klassifikationskorrektheit) auf die Kompetenzstufen beeinflussen (Pietsch et al., 2009). Für die Vergleichbarkeit des NEPS- und des Ländervergleichs-Tests lässt sich diesbezüglich mangels Daten keine Aussage treffen.

5.2.2 Zusammenhänge zwischen den Tests

Grundlegendes Ziel dieses Kapitels ist, zu überprüfen, ob die Skalenkennwerte in den drei Tests eine gleichwertige Interpretation erlauben. Dafür soll zunächst überprüft werden, ob die drei Tests eine gemeinsame Skala mathematischer Kompetenz abbilden. In einem ersten Schritt werden die Ergebnisse des NEPS-Tests dem TIMSS-Test gegenübergestellt und in einem zweiten dem Ländervergleichs-Test. Hierbei werden jeweils zwei Modelle geschätzt. Beim ersten Modell wird angenommen, dass der NEPS-Test und der TIMSS- bzw. Ländervergleichs-Test ein leicht unterschiedliches Konstrukt mathematischer Kompetenz erfassen. Daher werden die jeweiligen Tests zwar gemeinsam, jedoch auf zwei unterschiedlichen Dimensionen modelliert und die Korrelationen zwischen den Dimensionen werden berechnet, um Aussagen darüber treffen zu können, wie hoch der Anteil an gemeinsamer Varianz ist. In einem zweiten Modell wird angenommen, dass die jeweiligen Tests ein gemeinsames Konstrukt mathematischer Kompetenz konstituieren und es wird ein eindimensionales Modell geschätzt, bei dem die Items aus jeweils beiden Studien auf einer Dimension laden. Zusätzlich werden Modellgeltungstests berechnet, um die Modelle zu vergleichen. Die deskriptiven Statistiken der jeweiligen Tests, die ebenfalls Hinweise zur skalenbezogenen Vergleichbarkeit liefern können - werden in Kapitel 6 gegenübergestellt.

Schritt 1: Vergleich von NEPS und TIMSS

In einem ersten Schritt wird der NEPS-Test mit dem TIMSS-Test verglichen. Dafür werden zwei Modelle geschätzt (vgl. Kapitel 3.6.2.2). Im ersten Modell werden der NEPS-Test und der TIMSS-Test auf zwei unterschiedlichen Dimensionen skaliert. Die Daten werden unter der Annahme eines 1-PL-Modells skaliert. Die Partial-Credit-Items werden ebenfalls berücksichtigt und es wird eine Between-Item-Dimensionalität angenommen. Insgesamt fließen 24 NEPS-Items und 175 TIMSS-Items in die Analysen mit ein, die von insgesamt 733 Schülerinnen und Schülern bearbeitet wurden. Die Itemparameter wurden frei geschätzt. Die latente Korrelation zwischen den beiden Dimensionen liegt bei $r = .90$.

Tabelle 5.6: Ergebnisse der Modellgeltungstests und Modellvergleiche für das ein- und das zweidimensionale Modell für die Skalierung des NEPS- und TIMSS-Tests

	N	Parameter	Devianz	AIC	BIC	CAIC
NEPS-TIMSS 1Dim	733	**210**	39156	39576	40541	40751
NEPS-TIMSS 2Dim	733	212	**39109**	**39533**	**40508**	**40720**

Im zweiten Modell laden alle Items des NEPS- und des TIMSS-Tests auf einer Dimension mathematischer Kompetenz. Die Ergebnisse der Modellgeltungstest werden in Tabelle 5.6 gegenübergestellt. Es ist das Modell zu bevorzugen, welches die geringeren Werte aufweist. Weiterhin sollte das Einfachheitskriterium Berücksichtigung finden (vgl. Kapitel 3.6.2.3). Es zeigt sich, dass das eindimensionale Modell hinsichtlich der Anzahl der Parameter zu bevorzugen wäre, jedoch die Modellgeltungstests etwas niedrigere Werte bei dem zweidimensionalen Modell ergeben. Die Unterschiede in der Devianz zwischen dem ein- und dem zweidimensionalen Modell sind jedoch signifikant ($\chi^2 = 47$, $df = 2$). Insofern passt das zweidimensionale Modell besser auf die Daten als das eindimensionale Modell.

Schritt 2: NEPS und der Ländervergleich

In einem zweiten Schritt wurde der NEPS-Test gemeinsam mit dem Ländervergleichs-Test modelliert (vgl. Kapitel 3.6.2.3). Um die Höhe des Zusammenhangs der beiden Tests zu überprüfen, werden zunächst die latenten Korrelationen zwischen den Tests berechnet. Hierfür werden die Ergebnisse von 752 Schülerinnen und Schüler in den beiden Testverfahren gemeinsam skaliert, wobei die NEPS-Items die Faktorladung 1 aufweisen und die Ländervergleichs-Items auf Faktor 2 laden. Insgesamt fließen 24 NEPS-Items und 209 Ländervergleichs-Items in die Analysen mit ein. Die Tests werden unter der Annahme eines mehrdimensionalen Partial-Credit-Modells geschätzt. Die messfehlerbereinigte Korrelation zwischen den Test liegt bei $r = .92$. Die Höhe der Korrelation weist darauf hin, dass die Tests möglicherweise ein gemeinsames Konstrukt mathematischer Kompetenz abbilden.

Tabelle 5.7: Ergebnisse der Modellgeltungstests und Modellvergleiche für das Ein- und das Zweidimensionale Modell für die Skalierung des NEPS- und Ländervergleichs-Tests

	N	Parameter	Devianz	AIC	BIC	CAIC
NEPS-LV 1Dim	752	**236**	51375	51847	52054	52290
NEPS-LV 2Dim	752	238	**51338**	**51814**	**52023**	**52261**

Daher wird ein zweites Modell geschätzt, in dem die Items aus beiden Tests auf einen gemeinsamen Faktor laden. Um die beiden Modelle miteinander vergleichen zu können, werden Modellgeltungstests berechnet. Die Ergebnisse werden in Tabelle 5.7 gegenübergestellt. Der Modellvergleich zeigt, dass das eindimensionale Modell hinsichtlich des Einfachheitskriteriums zu bevorzugen wäre, weil es weniger Parameter schätzt, die Modellgeltungstests weisen jedoch für das zweidimensionale Modell tendenziell niedrigere Werte auf. Da die Unterschiede zwischen den Modellgeltungstests augenscheinlich nicht sehr hoch sind, wird zusätzlich ein Likelihood-Differenztest herangezogen, um zu überprüfen, ob die zwei zusätzlichen Parameter einen signifikanten Zugewinn bringen. Der Test ergibt, dass die Unterschiede zwischen den Devianzen signifikant sind $(\chi^2 = 37,\ df = 2)$. Das zweidimensionale Modell beschreibt demnach die empirischen Daten besser als das eindimensionale Modell.

Zwischenfazit

Ziel dieses Kapitels war weiterhin der Frage nachzugehen, ob es sich bei den drei Tests um unterschiedliche Konstrukte oder um ein gemeinsames Konstrukt mathematischer Kompetenz handelt. Nach Kolen und Brennan (2010) sowie Dorans und Holland (2000) können sowohl Korrelationskoeffizienten als auch ein Test auf Ein- bzw. Zweidimensionalität der Datenstruktur dazu genutzt werden, um Aussagen über die Vergleichbarkeit von Tests zu treffen. Die Ergebnisse zeigen, dass der NEPS-Mathematiktest sowohl mit dem TIMSS-Mathematiktest (r = .90) als auch mit dem Ländervergleichs-Mathematiktest hoch korreliert (r = .92). Die Höhe der Korrelationen ist ähnlich wie in anderen Vergleichsstudien. Beispielsweise ermittelten Blum et al. (2004) eine Korrelation von r = .92 zwischen dem nationalen und internationalen PISA-Mathematiktest 2003. Böhme et al. (2014) berechneten eine Korrelation zwischen dem Mathematiktest des Ländervergleichs 2011 und TIMSS 2011 in Höhe von r = .91 und Hartig und Frey (2012) stellten eine Korrelation in Höhe von r = .94 zwischen den Mathematiktests des Ländervergleichs (Normierungsstichprobe 2006) und PISA 2006 fest.

Die hohen Korrelationen legen die Vermutung nahe, dass ein erheblicher Anteil eines globalen Faktors sowohl zwischen dem NEPS- und dem TIMSS-Test als auch zwischen dem NEPS- und dem Ländervergleichs-Test existiert. Dennoch zeigen die Analysen hinsichtlich der Passung der Daten auf ein eindimensionales Modell, dass die beiden Konstrukte nicht

identisch sind und entsprechend der Empfehlungen von Rost (2004) ein zweidimensionales Modell zu bevorzugen wäre.

Zusammenfassend lässt sich also festhalten, dass eine zufriedenstellende Ähnlichkeit in den Konstrukten der Tests besteht, gleichzeitig zeigt sich jedoch auch, dass die beiden Konstrukte nicht gänzlich ein und dasselbe messen. Diese Konstruktverschiedenheit kann – neben den Messfehlern – zu Ungenauigkeiten im Linking führen (Kolen und Brennan, 2010). Die Unterschiede in Testergebnissen können damit beispielsweise zu einer fehlerbehafteten individuellen Zuordnung zu den Kompetenzstufen führen. Dies sein nach Kolen und Brennan (2010) jedoch kein Grund dafür, das Linking nicht vorzunehmen. Vielmehr sind hiermit Limitationen bei der Interpretation der Ergebnisse des Linking verbunden, die jedoch in weiteren Analysen (siehe folgendes Kapitel) noch näher spezifiziert werden.

6 Linking der Studien

In diesem Kapitel werden die Linking-Ergebnisse der NEPS-Studie mit der TIMSS- und der Ländervergleichs-Studie vorgestellt (vgl. Kapitel 3.6.3). Hierbei erfolgt zudem die Überprüfung der Ergebnisse hinsichtlich deskriptiver Statistiken (z. B. Mittelwerte und Standardabweichungen) als auch hinsichtlich der Verteilung auf die Kompetenzstufen. Darüber hinaus wird analysiert, ob die Ergebnisse stabil über Subgruppen sind (Forschungsfrage 3). Hierfür wird der NEPS-Test zuerst mit dem TIMSS-Test verlinkt und es erfolgt eine Analyse hinsichtlich der Exaktheit und Stabilität (vgl. Kapitel 6.1), anschließend wird der NEPS-Test mit dem Ländervergleichs-Test verlinkt (vgl. Kapitel 6.2).

6.1 Linking des NEPS- und des TIMSS-Tests

Um die Ergebnisse des NEPS-Tests auf der internationalen Metrik der TIMSS-Studie zu verorten, wurden drei Schritte durchgeführt. Zunächst wurden die Ergebnisse aus der Linking-Studie von dem NEPS- und dem TIMSS-Tests auf den Metriken der jeweiligen Hauptuntersuchungen fixiert. Eine detaillierte Beschreibung des Vorgehens und der Skalierungsmethoden ist dem Kapitel 3.6.3 zu entnehmen.

Die 175 TIMSS-Items, die in der Linking-Studie Verwendung fanden, wurden in einem 3-PL-Modell skaliert, wobei die Itemparameter aus der internationalen Hauptuntersuchung fixiert wurden. Die somit gewonnenen Personenparameter wurden mit einer linearen Transformation auf die internationale TIMSS-Metrik gebracht. Die NEPS-Items wurden im Gegenzug unter Verwendung der Itemparameter der NEPS-Hauptuntersuchung skaliert. Somit fand auch hier eine Übertragung auf die Metrik der Hauptuntersuchung statt. Der Tabelle 6.1 sind die deskriptiven Statistiken der beiden Tests für die 733 Schülerinnen und Schüler zu entnehmen, die an beiden Tests teilgenommen haben. In dem NEPS-Test erreichen die Schülerinnen und Schüler im Mittel einen WLE von θ = -.47 (SD = 1.18). Der höchste WLE der erreicht wurde liegt bei θ = 4.00 und der niedrigste bei θ = -3.55. Es zeigen sich geschlechtsspezifische Unterschiede. Die Schüler erreichen im Durchschnitt einen etwas besseren WLE (MW = -0.34, SD = 1.21) als die Schülerinnen (MW = -0.61, SD = 1.13). Die Verteilung ist tendenziell steiler und rechtschief und weicht damit leicht von einer Normalverteilung ab. Sowohl für die Gesamtgruppe als auch für die Gruppe der Schüler ist die

Abweichung in Schiefe und Kurtosis signifikant. Da die Schiefe unter 1 und die Kurtosis unter 4 liegt, ist diese Abweichung nach Miles (2001) jedoch noch als moderat zu beurteilen. In dem TIMSS-Test erreichen die Schülerinnen und Schüler im Mittel 545 Kompetenzpunkte (SD = 64). Der höchste Kompetenzwert der erreicht wurde sind 735 Kompetenzpunkte und der niedrigste sind 355 Kompetenzpunkte. Die Verteilung weicht nicht signifikant von der Normalverteilung ab. Auch im TIMSS-Test sind die Schüler mit einem Mittelwert von 550 (SD = 65) etwas besser als die Schülerinnen (MW = 540, SD = 62). Die Unterschiede zwischen den Geschlechtern sind im NEPS-Test etwas höher (d = .23) als im TIMSS-Test (d = .16). Die Ergebnisse weisen darauf hin, dass das Verlinken der beiden Skalen, welches im nächsten Schritt erfolgen soll, inkonsistent hinsichtlich des Geschlechts sein kann.

Weiterhin wurden die Korrelationen zwischen dem Mathematiktest von NEPS und TIMSS berechnet. Es zeigt sich, dass die Korrelationen für die Gesamtgruppe bei r = .72 (latente Korrelation r = .90, vgl. Kapitel 5.2.2), für die Gruppe der Schüler bei r = .73 und für die Gruppe der Schülerinnen bei r = .71 liegen. Diese moderaten Korrelationen liefern einen ersten Hinweis darauf, dass die Tests Gemeinsamkeiten aufweisen. Jedoch weisen sie auch darauf hin, dass die Tests zu einem gewissen Grad auch etwas Unterschiedliches messen. Dies gilt für die Gruppe der Schülerinnen tendenziell etwas stärker als für die Gruppe der Schüler.

Tabelle 6.1: Deskriptive Statistiken für den NEPS- und den TIMSS-Test

		N	Min	Max	MW	SD	Schiefe	Kurtosis
NEPS	Gesamt	733	-3.55	4.00	-0.47	1.18	0.28	3.36
	Männlich	381	-3.55	4.00	-0.34	1.21	0.39	3.52
	Weiblich	352	-3.55	2.16	-0.61	1.13	0.08	2.93
TIMSS	Gesamt	733	355	735	545	64	0.07	3.06
	Männlich	381	380	735	550	65	0.11	3.00
	Weiblich	352	355	710	540	62	0.00	3.08

Die dargestellten Ergebnisse der beiden Studien lassen sich jedoch, bedingt durch die unterschiedlichen Metriken, nicht ohne weiteres miteinander in Beziehung setzen. Daher erfolgt in einem dritten Schritt die Verlinkung der beiden Metriken durch das equipercentile

Linking mit der Computersoftware LEGS 2.0.1 (vgl. Kapitel 3.6.3). Da die Software ganzzahlige, positive Werte voraussetzt, wurden die Ergebnisse des NEPS-Tests (x) transformiert (X_T).

$$X_T = rnd ((NEPS_{WLE} *100) + 500) \quad \text{(Formel 1)}$$

Weiterhin wurden die Kompetenzwerte der Schülerinnen und Schüler im TIMSS-Test auf Fünferstellen gerundet, da die Computersoftware LEGS dies als Maximum vorgibt.

Für das equipercentile Linking werden die Häufigkeitsverteilungen sowie die Perzentilränge des NEPS (X_T) - und des TIMSS-Tests (Y) auf den Metriken der jeweiligen Hauptuntersuchung benötigt. Ein Auszug der Häufigkeitsverteilung und der Perzentilränge aus den unteren, mittleren und oberen Kompetenzwertbereichen wird in der Tabelle 6.2 dargestellt. Der Tabelle sind die Kompetenzwerte von NEPS (X und X_T) zu entnehmen, mit der jeweiligen Häufigkeit des Vorkommens (N) und dem Perzentilrang (P(X)), jeweils für die Gesamtgruppe, die Subpopulation der Jungen und der Subpopulation der Mädchen. Die Kompetenzwerte für TIMSS (Y) sind aus der Tabelle 6.3 abzulesen, wiederum mit den gleichen Angaben wie für die NEPS Kompetenzwerte (N und P(Y)) und ebenfalls getrennt nach Gesamtgruppe und differenziert nach Geschlecht.

Ein Vergleich der beiden Tabellen zeigt, dass der NEPS-Mathematiktest – bedingt dadurch, dass er aus nur 24 Items besteht – deutlich weniger Punktschätzer aufweist als der TIMSS-Test, der aus 175 Mathematikitems besteht. Die Konsequenz hieraus ist, dass nicht alle möglichen Punktschätzer vorkommen und damit Leerstellen in der Tabelle entstehen. Beispielsweise gibt es bei NEPS den Punktschätzer 145 und anschließend erst wieder den Punktschätzer 174. Die Werte dazwischen kommen nicht vor und somit erhalten alle Werte, die dazwischen liegen, einen gleichen Perzentilrang von P(X) = .409.

Tabelle 6.2: Verteilung für die Ergebnisse des NEPS-Mathematiktests für die Gesamtgruppe und getrennt nach Geschlecht

		Gesamt		Männlich		Weiblich	
X	**X_T**	**N**	**P(X)**	**N**	**P(X)**	**N**	**P(X)**
-3.55	145	3	0.205	1	0.132	2	0.282
-3.26	174	2	0.546	0	0.265	2	0.845
-3.02	198	7	1.160	3	0.661	4	1.690
-2.81	219	8	2.183	5	1.720	3	2.676
-2.62	238	7	3.206	3	2.778	4	3.662
-2.44	256	6	4.093	1	3.307	5	4.930
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
-0.74	426	24	43.520	11	39.021	13	48.310
-0.64	436	22	46.658	14	42.328	8	51.268
-0.54	446	18	49.386	9	45.370	9	53.662
-0.44	456	29	52.592	13	48.280	16	57.183
-0.34	466	27	56.412	14	51.852	13	61.268
-0.24	476	19	59.550	9	54.894	10	64.507
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
1.56	656	6	95.362	5	94.048	1	96.761
1.74	674	12	96.589	8	95.767	4	97.465
2.16	716	13	98.295	6	97.619	7	99.014
2.43	743	1	99.250	1	98.545	0	100.000
2.77	777	2	99.454	2	98.942	0	100.000
4.00	900	3	99.795	3	99.603	0	100.000

Tabelle 6.3: Verteilung für die Ergebnisse des TIMSS-Mathematiktests für die Gesamtgruppe und getrennt nach Geschlecht

	Gesamt		**Männlich**		**Weiblich**	
Y	**N**	**P(Y)**	**N**	**P(Y)**	**N**	**P(Y)**
355	1	0.068	0	0.000	1	0.142
375	2	0.273	0	0.000	2	0.570
380	1	0.478	1	0.131	0	0.855
385	1	0.615	0	0.262	1	0.997
395	3	0.888	2	0.525	1	1.282
405	2	1.230	1	0.919	1	1.567
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
560	18	59.699	8	57.218	10	62.393
565	27	62.773	17	60.499	10	65.242
570	24	66.257	11	64.173	13	68.519
575	19	69.194	10	66.929	9	71.652
580	20	71.858	10	69.554	10	74.359
585	21	74.658	11	72.310	10	77.208
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
700	2	98.770	2	98.163	0	99.430
705	1	98.975	0	98.425	1	99.573
710	2	99.180	1	98.556	1	99.858
715	2	99.454	2	98.950	0	100.000
720	2	99.727	2	99.475	0	100.000
735	1	99.932	1	99.869	0	100.000

Die Perzentilränge der beiden Studien können nun miteinander in Beziehung gesetzt werden. Dies erfolgt mit Hilfe der Computersoftware LEGS. Das Ergebnis der Analysen ist eine sogenannte Konkordanz-Tabelle (vgl. Tabelle 6.3), hier für die Verlinkung der Ergebniswerte des NEPS auf die Metrik des TIMSS-Tests (X → Y), in der die sich entsprechenden

Ergebniswerte der beiden Tests einander gegenüberstellt werden. Auch an dieser Stelle wird für die Übersichtlichkeit nur ein Teil der Tabelle angezeigt. In der gekürzten Tabelle werden die äquivalenten Ergebniswerte für geringe, mittlere und hohe Kompetenzwerte ausgegeben. Darüber hinaus werden in der Konkordanz-Tabelle die gerundeten Werte angegeben, um einen in TIMSS existierenden Kompetenzwert zu erhalten. Das Linking wurde jeweils getrennt für die Gesamtgruppe und die Subpopulationen differenziert nach Geschlecht geschätzt. Zudem sind der Tabelle die nicht geglättet äquivalenten Ergebniswerte (no Smoothing) und die im Nachhinein geglättet äquivalenten Ergebniswerte (Postsmoothing) für S = .3 und S = 1.0 zu entnehmen. Dabei sind die Werte folgendermaßen zu interpretieren (vgl. Kapitel 3.6.3): Angenommen S = 0, so ergäbe sich keine Änderung an der Transformationsfunktion und die äquivalenten Ergebniswerte würden sich nicht verändern. Je höher S gesetzt wird, desto glatter wird die Funktion und wird so annähernd zu einer Geraden.

Die Abbildungen 6.1 und 6.2 liefern eine grafische Gegenüberstellung der äquivalenten Ergebniswerte des NEPS-Tests im Vergleich zu den Ergebniswerten des NEPS-Tests. Die erste Abbildung zeigt das Ergebnis für die nicht geglätteten Wert und die zweite Abbildung die Ergebnisse für die im Nachhinein geglätteten Werte (S = 1.0). Bereits für die nicht geglätteten Werte zeigt sich, dass die Funktion annähernd linear ist. Im mittleren Kompetenzbereich scheint das Linking jedoch etwas besser zu passen als in den Randbereichen. Hier wird deutlich, dass die Gruppen differenziert nach Geschlecht jeweils leicht von der Gesamtgruppe abweichen. Im unteren Kompetenzbereich bekommen die Jungen etwas höhere Kompetenzwerte zugeordnet als die Mädchen. Im oberen Kompetenzbereich dreht sich dieser Befund um. Bei der geglätteten Funktion sind kaum noch Gruppenunterschiede zu erkennen. Nur in den oberen Kompetenzbereichen liegen weiterhin leichte Abweichungen der Subgruppen von der Gesamtgruppe vor.

Die leichten Abweichungen zwischen den Gruppen in den Randbereichen könnten dadurch entstehen, dass die Kompetenzwerte in diesen Bereichen nur von wenigen Schülerinnen und Schülern erlangt werden. Zudem wird durch die Angabe der gerundeten Kompetenzwerte bereits ein Rauschen verursacht.

Tabelle 6.4: Äquivalente Ergebniswerte des NEPS-Tests auf der TIMSS-Metrik

		no smoothing			S=0.30			S=1.00		
X	X_T	Gesamt	Männlich	Weiblich	Gesamt	Männlich	Weiblich	Gesamt	Männlich	Weiblich
-3.55	145	375	380	355	355	355	358	355	355	358
-3.54	146	375	395	375	356	356	365	356	356	365
-3.53	147	375	395	375	357	356	366	357	356	366
-3.52	148	375	395	375	358	357	366	358	357	366
-3.51	149	375	395	375	359	358	366	359	358	366
-3.50	150	375	395	375	360	359	367	360	359	367
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
-0.25	475	555	555	560	556	555	561	556	555	562
-0.24	476	560	555	565	559	555	563	559	555	563

[Fortsetzung der Tabelle 6.4.]

		no smoothing			S=0.30			S=1.00		
X	**X_T**	**Gesamt**	**Männlich**	**Weiblich**	**Gesamt**	**Männlich**	**Weiblich**	**Gesamt**	**Männlich**	**Weiblich**
-0.22	478	560	555	565	560	556	565	560	556	565
-0.21	479	560	555	565	560	556	565	560	556	565
-0.20	480	560	555	565	560	556	565	560	556	565
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
3.95	895	720	715	730	735	716	734	735	716	734
3.96	896	720	715	731	735	716	734	735	716	734
3.97	897	720	715	732	735	716	735	735	716	735
3.98	898	720	715	733	735	716	735	735	716	735
3.99	899	720	715	734	735	716	735	735	716	735
4.00	900	720	720	735	735	729	735	735	729	735

Abbildung 6.1: Equipercentiles Linking für die Gesamtgruppe und differenziert nach Geschlecht – no smoothing

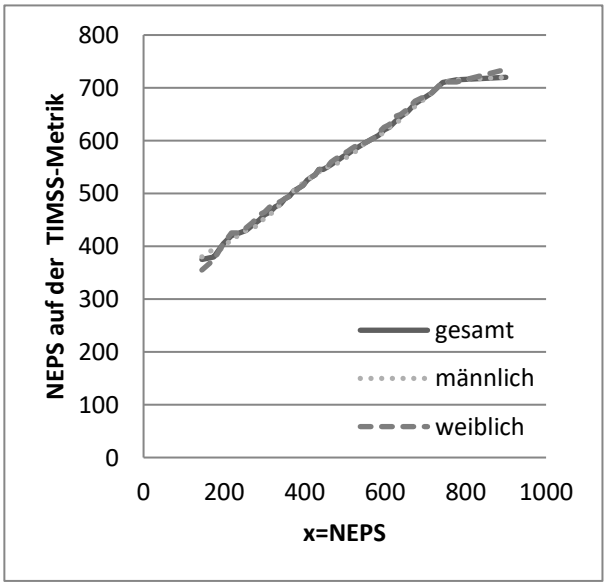
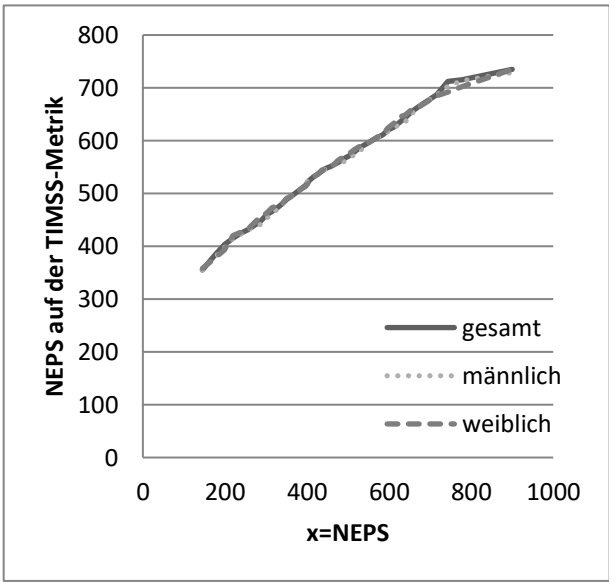


Abbildung 6.2: Equipercentiles Linking für die Gesamtgruppe und differenziert nach Geschlecht - postsmoothing (S = 1.0)



Die deskriptiven Ergebnisse des equipercentilen Linking und der Skalierung der beiden Einzeltests werden in der Tabelle 6.4 dargestellt. Die Statistiken des equipercentilen Linking basieren auf gerundeten und nicht geglätteten Werten. Werden die Ergebnisse des NEPS-Tests auf der Metrik des TIMSS-Tests verortet, ergibt sich ein Mittelwert von 545 Kompetenzpunkten (SD = 63). Dies entspricht exakt den mit dem TIMSS-Test ermittelten Mittelwerten (vgl. Tabelle 6.5) und ist ein Ergebnis und auch ein Vorteil des equipercentilen Linking (Kolen & Brennan, 2010). Auch die Unterschiede zwischen den Jungen und den Mädchen bleiben bei der Verlinkung im gleichen Maße erhalten (d = .16). Leichte Unterschiede zwischen den Ergebnissen des TIMSS-Tests und dem NEPS-Test auf der Metrik von TIMSS ergeben sich lediglich in der Verteilung (Schiefe und Kurtosis). Die Verteilungen weichen jedoch beide nicht signifikant von der Normalverteilung ab. Die Methode des equipercentile Linking ist daher für die vorliegenden Daten eine angemessene Linking-Methode.

Tabelle 6.5: Deskriptive Statistiken für den NEPS- und den TIMSS-Test sowie für die äquivalenten Ergebniswerte

		N	Min	Max	MW	SD	Schiefe	Kurtosis
NEPS	Gesamt	733	-3.55	4.00	-0.47	1.18	0.28	3.36
	Männlich	381	-3.55	4.00	-0.34	1.21	0.39	3.52
	Weiblich	352	-3.55	2.16	-0.61	1.13	0.08	2.93
TIMSS	Gesamt	733	355	735	545	64	0.07	3.06
	Männlich	381	380	735	550	65	0.11	3.00
	Weiblich	352	355	710	540	62	0.00	3.08
Linking (rnd/no smoothing)	Gesamt	733	375	720	545	63	0.06	2.99
	Männlich	381	380	720	550	65	0.08	2.99
	Weiblich	352	355	735	540	62	-0.04	3.10

Die Ergebnisse des NEPS-Tests sind nun auf der Metrik des TIMSS-Tests verortet. Damit lassen sich die NEPS-Kompetenzwerte den Kompetenzstufen von TIMSS zuordnen (vgl. Kapitel 3.6.3). Die TIMSS-Leistungsskala ist in fünf Abschnitte (Kompetenzstufen) unterteilt. In diesen werden die mathematischen und kognitiven Anforderungen der Schülerinnen und Schüler beschrieben, die nötig sind, um die jeweilige Kompetenzstufe zu erreichen. Die Ergebnisse der Zuordnung zu den Kompetenzstufen von sowohl dem TIMSS-Test als auch dem NEPS-Test auf der Metrik des TIMSS-Tests werden in Tabelle 6.6 dargestellt. Die Tabelle zeigt die prozentuale Verteilung der Schülerinnen und Schüler auf die fünf Kompetenzstufen: (1) rudimentär (unter 400 Punkte), (2) niedrig (400 – 474 Punkte), (3) durchschnittlich (475 – 549 Punkte), (4) hoch (550 – 624 Punkte) und (5) fortgeschritten (ab 625 Punkte). Der ersten Kompetenzstufe wurden etwa 1% der Schülerinnen und Schüler zugeordnet. Sie verfügen über ein rudimentäres schulisches Anfangswissen. Die zweite Kompetenzstufe erreichen etwa 12% (NEPS) bzw. 13% (TIMSS) der Schülerinnen und Schüler. Um diese Kompetenzstufe zu erreichen, müssen die Schülerinnen und Schüler grundlegende mathematische Fertigkeiten und Fähigkeiten besitzen, um die Aufgaben zu lösen. Der dritten Kompetenzstufe konnten bei beiden Tests aufgerundet 39% der Schülerinnen und Schüler zugeordnet werden. Damit können die Schülerinnen und Schüler einfache situationsgetreue Mathematikaufgaben lösen. Etwa 38% der Schülerinnen und Schüler erreichen die vierte Kompetenzstufe. Hierfür ist es

erforderlich, dass sie ihr Wissen nutzen können, um mathematische Probleme zu lösen. Der höchsten Kompetenzstufe fallen rund 9% (TIMSS) bzw. 11% (NEPS) der Schülerinnen und Schüler zu und sind damit in der Lage komplexe mathematische Probleme zu lösen und ihr Vorgehen zu erklären.

Tabelle 6.6: Prozentuale Verteilung auf die Kompetenzstufen der TIMSS-Studie für den TIMSS-Test und die äquivalenten Ergebniswerte des NEPS-Tests

		TIMSS 2011 Kompetenzstufen						
		Stufe 1	**Stufe 2**	**Stufe 3**	**Stufe 4**	**Stufe 5**	**Gesamt**	**Cohen's Kappa**
TIMSS	Gesamt	1.2%	13.1%	38.6%	37.7%	9.4%	100%	
	Männlich	0.8%	13.1%	34.6%	40.7%	10.8%	100%	
	Weiblich	1.7%	13.1%	42.9%	34.4%	8.0%	100%	
Linking	Gesamt	0.7%	11.6%	38.3%	38.5%	10.9%	100%	0.339
	Männlich	0.3%	9.7%	37.0%	39.6%	13.4%	100%	0.370
	Weiblich	1.1%	13.6%	39.8%	37.2%	8.2%	100%	0.303

Anmerkungen: Die Werte in der Tabelle sind gerundet. Dadurch kann die Summe der Prozente minimal von der Gesamtsumme 100% abweichen.

Die höchste Differenz der prozentualen Verteilungen beträgt absolut nicht mehr als 1.5%. Der Tabelle sind ebenfalls die Verteilungen für die männlichen und die weiblichen Probanden zu entnehmen, die auf Grundlage der equipercentilen Ergebniswerte der Gesamtgruppe berechnet wurden. Auch diese scheinen eine gute Annäherung der NEPS-Ergebnisse auf der TIMSS-Metrik an die Verteilung der TIMSS-Ergebnisse zu bieten. Mit einem Cohen's Kappa von $\kappa = .34$ für die Gesamtgruppe besteht eine ausreichende Übereinstimmung zwischen den beiden Verteilungen (vgl. Kapitel 3.6.1.3). Der Wert zeigt an, dass zu einem gewissen Grad Unterschiede zwischen den Verteilungen bestehen. Ein χ^2-Test zeigt jedoch bei einem Vergleich der Verteilungen auf die Kompetenzstufen, dass sowohl für die Gesamtgruppe ($\chi^2 = 2.70$, df = 4) als auch für die Gruppe der Schüler ($\chi^2 = 4.38$, df = 4) und der Schülerinnen ($\chi^2 = 1.27$, df = 4) keine signifikanten Unterschiede bestehen.

Die Tabelle 6.7 zeigt die prozentuale Zuordnung auf die jeweiligen Kompetenzstufen an. Dabei wird die Zuordnung zu den Kompetenzstufen gemessen mit dem TIMSS-Test und

gemessen mit dem NEPS-Test (auf der Metrik von TIMSS) einander gegenübergestellt. Insgesamt 55% der Schülerinnen und Schüler konnten in beiden Tests der gleichen Kompetenzstufe zugeordnet werden. In 23.8% der Fälle wurde die Leistung der Schülerinnen und Schüler im NEPS-Test (auf der TIMSS-Metrik) um eine Stufe überschätzt und in 1.8% der Fälle um zwei Stufen. Hingegen wurde die Leistung der Schülerinnen und Schüler gemessen mit dem NEPS-Test in 18.2% um eine Stufe und in 1.4% der Fälle zwei Stufen unterschätzt.

Tabelle 6.7: Prozentuale Zuordnung zu den Kompetenzstufen - Vergleich der Schülerkompetenzen im TIMSS-Test und im NEPS-Test (TIMSS-Metrik)

		NEPS-Ergebnisse auf TIMSS-Metrik					
		Stufe 1	**Stufe 2**	**Stufe 3**	**Stufe 4**	**Stufe 5**	**Gesamt**
TIMSS-Test	**Stufe 1**	**0.0%**	1.0%	0.3%	0.0%	0.0%	1.3%
	Stufe 2	0.4%	**5.5%**	6.3%	1.0%	0.0%	13.2%
	Stufe 3	0.3%	4.4%	**21.6%**	11.9%	0.5%	38.7%
	Stufe 4	0.0%	0.8%	10.0%	**22.2%**	4.6%	37.6%
	Stufe 5	0.0%	0.0%	0.3%	3.4%	**5.7%**	9.4%
	Gesamt	0.7%	11.7%	38.5%	38.5%	10.8%	100.0%

Anmerkungen: Die Werte in der Tabelle sind gerundet. Dadurch kann die Summe der Prozente minimal von der Gesamtsumme 100% abweichen.

Die Tabelle 6.8 gibt einen weiteren Einblick in die Klassifikationskorrektheit der Verteilung auf die Kompetenzstufen für den NEPS-Test (TIMSS-Metrik) und den TIMSS-Test. Die mittlere Klassifikationskorrektheit liegt bei 43.5%. Unterschiede zeigen sich vor allem in den niedrigeren Kompetenzstufen. Die Schülerinnen und Schüler, die im TIMSS-Test der Kompetenzstufe 1 zugeordnet wurden, wurden mit dem NEPS-Test (TIMSS-Metrik) der Kompetenzstufe 2 oder 3 zugeordnet. Hierbei ist allerdings zu beachten, dass es sich lediglich um 1.3% der Fälle handelt (vgl. Tabelle 6.8). Hingegen konnten 60.9 % der Schülerinnen und Schüler, die im TIMSS-Test die Kompetenzstufe 5 erreicht haben, im NEPS-Test ebenfalls dieser Kompetenzstufe zugeordnet werden. Es zeigt sich, dass mit ansteigender Kompetenzstufe die Zuordnungsgenauigkeit zwischen den Tests zunimmt.

Tabelle 6.8: Klassifikationskorrektheit – Vergleich der Schülerkompetenzen im TIMSS-Test und im NEPS-Test (TIMSS-Metrik)

		NEPS-Ergebnisse auf TIMSS-Metrik				
		Stufe 1	**Stufe 2**	**Stufe 3**	**Stufe 4**	**Stufe 5**
TIMSS-Test	**Stufe 1**	0.0%	77.8%	22.2%	0.0%	0.0%
	Stufe 2	3.1%	41.7%	47.9%	7.3%	0.0%
	Stufe 3	0.7%	11.3%	55.8%	30.7%	1.4%
	Stufe 4	0.0%	2.2%	26.4%	59.1%	12.3%
	Stufe 5	0.0%	0.0%	2.9%	36.2%	60.9%

Anmerkungen: Die Werte in der Tabelle sind gerundet. Dadurch kann die Summe der Prozente minimal von der Gesamtsumme 100% abweichen.

Invarianz über Subgruppen

Werden die aufgezeigten Ergebnisse hinsichtlich der unterschiedlichen Gruppen (Gesamt, Männlich, Weiblich) verglichen, zeigt sich, dass es Unterschiede zwischen den Gruppen gibt. Es stellt sich die Frage, ob sich diese Unterschiede auch statistisch absichern lassen. Um dies zu überprüfen, werden Berechnungen durchgeführt (vgl. Kapitel 3.6.3.4), die sich auf drei Gruppenvergleiche beziehen: Männlich – Gesamt (1 – 0), Weiblich – Gesamt (2 – 0) und Männlich – Weiblich (1 -2).

Für einen Vergleich werden zum einen die paarweisen Statistiken angegeben (vgl. Tabelle 6.8). Der Tabelle sind die gewichtete mittlere Differenz (wMD), die gleichgewichtete mittlere Differenz (ewMD), die gewichtete mittlere Differenz der absoluten Werten (wMAD) sowie die gleichgewichtete mittlere Differenz der absoluten Werten (ewMAD) zwischen den skalenäquivalenten Ergebnissen aus dem equipercentilen Linking zu entnehmen. Es werden hier die Ergebnisse für die nicht geglätteten Werte (no smoothing) und die geglätteten Werte für S = .3 und S = 1.0 angezeigt. Läge eine Gruppen-Invarianz vor, so wären die Differenzen zwischen den Gruppen gleich Null (vgl. Kapitel 3.6.3.4).

Zum anderen werden für einen Vergleich die Unterschiede zwischen den Gruppen der gesamten Stichprobe und differenziert nach Geschlecht in den Abbildungen 6.3 und 6.4 veranschaulicht. Die erste Abbildung zeigt einen Vergleich der Gruppen auf Grundlage der

nicht geglätteten Werte und die zweite Abbildung für die im Nachhinein geglätteten Werte (S = 1.0). Für jeden NEPS-Wert (X) wird die Höhe der Differenz zwischen den äquivalenten Ergebniswerten des jeweiligen Gruppenvergleichs angegeben. Die größten Differenzen treten vor allem in den Randbereichen auf. Die höchste Differenz – in Bezug auf die nicht geglätteten Werte – ergibt sich beim Vergleich der äquivalenten Ergebniswerte getrennt nach Geschlecht (25 Punkte Unterschied).

Im Vergleich zur Gesamtgruppe sind die Subgruppen für die Jungen zwischen dem Kompetenzwert 314 und 624 und für die Mädchen zwischen 288 und 610 relativ stabil. Stabil bedeutet in diesem Zusammenhang, dass die Kompetenzwerte zwischen den Subgruppen und der Gesamtgruppe nicht um mehr als 5 Kompetenzwerte abweicht. In den Randbereichen sind die Abweichungen mit einem Maximum von 20 Kompetenzwerten deutlich höher. Jedoch betrifft dies bei den Jungen im Vergleich zur Gesamtgruppe nur diejenigen, die auf der untersten Kompetenzstufe liegen und auch mit 20 Punkten mehr, nicht der nächsthöheren Kompetenzstufe zugeordnet werden würden. Gleiches gilt für den oberen Randbereich. Wobei hier eine Zuordnung zur nächstkleineren Kompetenzstufe möglich wäre, da die Kompetenzwerte im Gruppenvergleich zwischen der Gesamtgruppe und der Jungen nur bis 624 Kompetenzpunkte stabil sind und erst ab 625 Kompetenzpunkten die oberste Kompetenzstufe erreicht wird. Bei dem Vergleich der Gesamtgruppe und der Gruppe der Mädchen sind die Kompetenzwerte bis etwa 610 Kompetenzpunkte stabil. Hier besteht demnach ebenfalls in dem oberen Randbereich eine gewisse Instabilität über Gruppen.

Die paarweisen Statistiken in der Tabelle 6.9 zeigen, dass die Gruppe der Jungen in der Transformation getrennt nach Geschlecht niedrigere Werte erhalten als bei der Transformation der Gesamtgruppe. Dies bedeutet, dass die Männer bei der Transformation der Gesamtgruppe etwas übervorteilt werden (eMD = -1.940, ewMD = -1.667). Genau andersherum verhält es sich für die Mädchen. Die Mädchen erhalten in der Transformation

Abbildung 6.3: Differenzen der Ergebniswerte zwischen der Gesamtgruppe und den Geschlechtern – no smoothing (0 = Gesamt; 1 = Männlich; 2 = Weiblich)

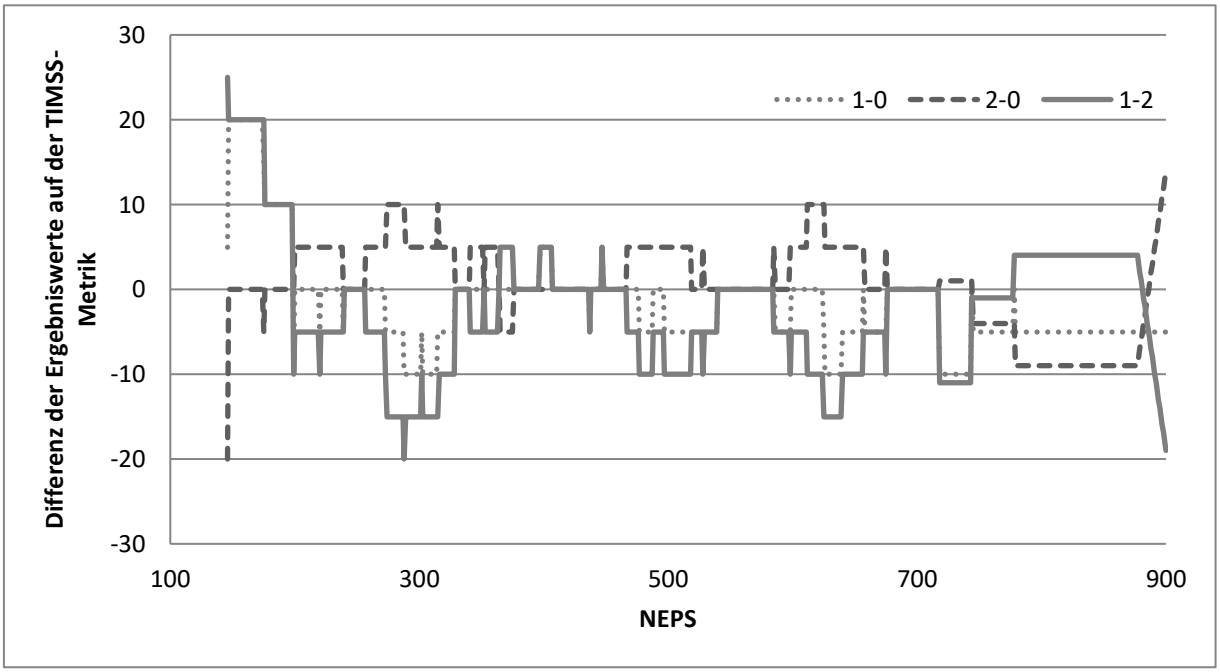
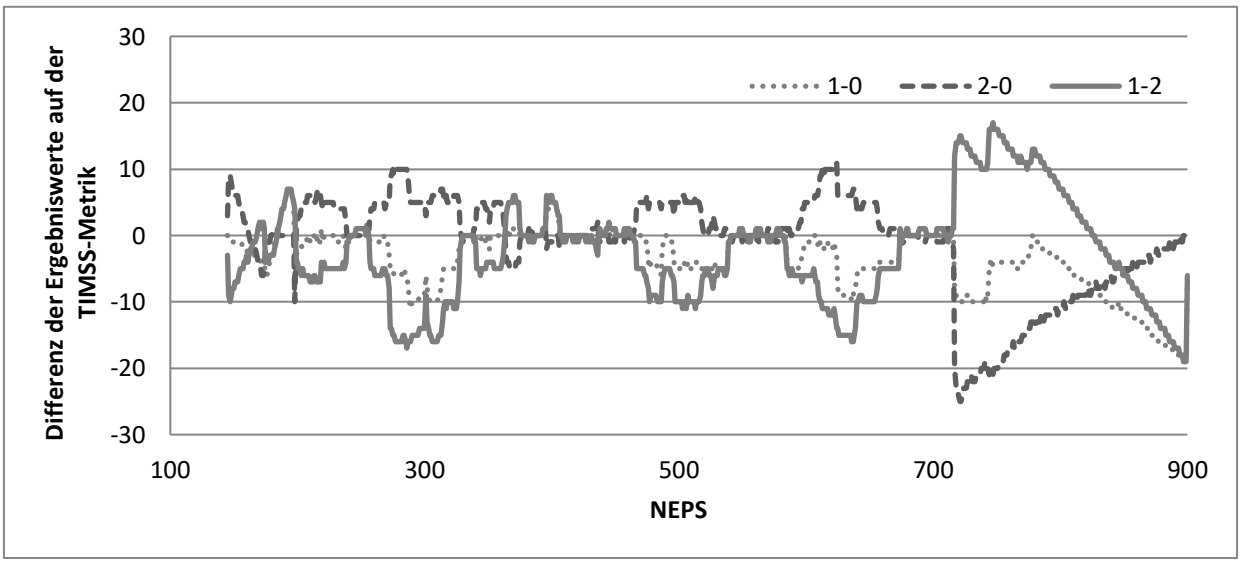


Abbildung 6.4: Differenzen der Ergebniswerte zwischen der Gesamtgruppe und den Geschlechtern – postsmoothing (S = 1.00; 0 = Gesamt; 1 = Männlich; 2 = Weiblich)



differenziert nach Geschlecht etwas höhere Werte als in der Gesamtgruppe, werden also bei der Transformation der Gesamtgruppe etwas benachteiligt (wMD = 2.392, ewMD = 0.521). Die größten Unterschiede sind jedoch beim Vergleich der Geschlechter zu beobachten (wMAD = 5.688; ewMAD = 5.762). Hier zeigt sich, dass die Gruppe der Jungen im Mittel niedrigere äquivalente Ergebniswerte zugewiesen bekommen als die Mädchen.

Weiterhin ist der Tabelle 6.9 zu entnehmen, dass die gleich gewichteten Mittelwertsdifferenzen (ewMD) etwas höher ausfallen als die zweifach gewichteten (wMD). Dies kann daran liegen, dass die Vorkommenshäufigkeit der jeweiligen äquivalenten Ergebniswerte bei der Berechnung der zweifach gewichteten Mittelwertsdifferenzen mit einfließt, wohingegen dies bei der Berechnung der gleich gewichteten Mittelwertsdifferenzen dies nicht der Fall ist. Hier geht lediglich ein konstanter Wert als Gewicht mit ein. Daher werden die Randbereiche, in denen zwar die höchsten Abweichungen vorkommen, jedoch gleichzeitig die Vorkommenshäufigkeit am geringsten ist, bei den zweifach gewichteten Mittelwertsdifferenzen nicht so stark gewichtet.

Tabelle 6.9: Paarweise Statistiken für die Subgruppen Gesamt, Männlich und Weiblich (no smoothing, S=.3 und S=1.0):

Untergruppen	Methode	wMD	ewMD	wMAD	ewMAD
1-0	no smoothing	-1.940	-1.667	2.723	3.955
	S = 0.3	-1.943	-3.575	2.712	3.914
	S = 1.0	-1.850	-3.563	2.376	3.947
2-0	no smoothing	2.392	0.521	2.977	3.545
	S = 0.3	1.588	-0.816	2.454	4.692
	S = 1.0	1.479	-0.869	2.584	4.861
1-2	no smoothing	-4.329	-2.188	5.688	5.762
	S = 0.3	-3.487	-2.759	4.568	6.116
	S = 1.0	-3.310	-2.694	4.696	6.250

0 = Gesamt; 1 = Männlich; 2 = Weiblich

Zusammenfassend lässt sich festhalten, dass Unterschiede zwischen den Transformationen getrennt nach Geschlecht im Vergleich zur Gesamtgruppe bestehen. Es stellt sich jedoch die Frage nach der Bedeutsamkeit dieser Unterschiede für das Linking. Für eine Interpretation bzw. die Evaluation dieses Linkings bezüglich der Gruppen-Invarianzen können zwei Kennwerte berechnet werden (vgl. Kapitel 3.6.3.4): (1) die standardisierte Root Mean Square Difference (RMSD), welche für jede einzelne Differenz der äquivalenten Ergebniswerte zwischen zwei Subgruppen berechnet wird, und (2) die Root Expected Mean Square Difference (REMSD), welche alle Differenzen für die jeweiligen Gruppen

zusammenfasst. Im Gegensatz zu den vorherigen Ergebnissen, die alle auf den gerundeten äquivalenten Ergebniswerten basieren, basieren diese Kennwerte auf den ungerundeten äquivalenten Ergebniswerten. Nur wenn die ungerundeten Werte verwendet werden, kann eine Interpretation der Ergebnisse in Bezug auf die Relevanz erfolgen. Hierfür kann eine sogenannte score difference that matters (DTM) berechnet werden (vgl. hierzu Kapitel 3.6.3.4). Die DTM entspricht einem halben transformierten Ergebniswert. In diesem Vergleich liegt die DTM dementsprechend bei DTM = .50. Zusätzlich kann eine standardisierte DTM (sDTM) berechnet werden, die zusätzlich die Standardabweichung berücksichtigt und damit auf der gleichen Metrik wie der REMSD liegt. In dem vorliegenden Fall ist die sDTM = .5/64 = .01. Wird von diesen Werten ausgegangen, wird deutlich, warum die Relevanz besteht, die ungerundeten Werte heranzuziehen. Soll beispielsweise die Differenz zwischen Jungen und Mädchen betrachtet werden und die Jungen weisen einen ungerundeten äquivalenten Ergebniswert von 374.7 auf und die Mädchen von 374.4, dann ist die Differenz mit .3 nicht bedeutend, wenn man den Referenzwert von DTM = .5 anlegt. Werden hingegen die gerundeten äquivalenten Ergebniswerte herangezogen entsteht eine Differenz von 1.0 und wäre damit bedeutend. Wenn die Jungen hingegen einen äquivalenten Ergebniswert von 374.6 und die Mädchen von 375.4 hätten, läge die Differenz bei .8 und damit über dem Referenzwert von .5. Werden hingegen hier die Werte gerundet, dann wären die äquivalenten Ergebniswerte zwischen den Geschlechtern gleich und damit nicht bedeutend.

Tabelle 6.10: wREMSD und ewREMSD Statistiken für das equipercentile Linking (no smoothing, S=.3 und S=1.0)

	no smoothing	**s=0.3**	**s=1.0**
wREMSD	0.06	0.05	0.06
ewREMSD	0.08	0.10	0.10

Der Tabelle 6.11 sind die REMSD (sowohl für den zweifach gewichteten als auch für den gleich gewichteten REMSD) für die ungerundeten äquivalenten Ergebniswerte zu entnehmen. Es zeigt sich, dass zwar alle Werte deutlich unter dem Richtwert der DTM von .50, jedoch leicht über der sDTM von .01 liegen. Die Berücksichtigung der Gewichte hat zwar einen Einfluss auf die Höhe des REMSD, jedoch ist der Unterschied zwischen den Werten nur sehr klein und hat keinen Einfluss darauf, ob die Werte unterhalb oder oberhalb der DTM bzw. der sDTM liegen.

Zwischenfazit

Ziel dieses Kapitels war zunächst, die Ergebnisse aus dem NEPS- und dem TIMSS-Test hinsichtlich deskriptiver Statistiken und hinsichtlich der Zuordnung zu Kompetenzstufen zu vergleichen, um anschließend Aussagen zur Exaktheit und Stabilität des Linking treffen zu können. Es zeigt sich, dass die Schülerinnen und Schüler im NEPS-Test im Mittel einen WLE von θ = -0.47 erreichen (SD = 1.18). Die Schülerinnen erreichen im Durchschnitt einen etwas geringeren WLE als die Schüler. Der Unterschied zwischen den Geschlechtern weist eine Effektstärke von d = .23 auf. Im TIMSS-Test erreichen die Schülerinnen und Schüler im Mittel einen Kompetenzwert von 545 (SD = 64). Die Effektstärke hinsichtlich der Unterschiede zwischen den Geschlechtern beträgt d = .16 und ist damit etwas geringer als im NEPS-Test. Bei beiden Tests schneiden die Schüler etwas besser ab als die Schülerinnen. Durch die unterschiedlichen Metriken der Tests sind die Ergebnisse jedoch nicht vergleichbar. Das Linking des NEPS-Tests auf die Metrik des TIMSS-Tests ermöglicht einen Vergleich der deskriptiven Statistiken auf der gemeinsamen Metrik von TIMSS. Das Linking erfolgte mit der Methode des equipercentilen Linking unter Verwendung der Computersoftware LEGS 2.0.1. Die grafische Darstellung der Verteilung der Ergebnisse zeigt, dass sich die Verteilungen sehr ähnlich sind, dass jedoch Abweichungen, vor allem zwischen den Subgruppen und innerhalb der Randbereiche bestehen. Das Smoothing hat in den Randbereichen den größten Effekt und gleicht diese Abweichungen etwas aus. Die Abweichungen in den Randbereichen entstehen beispielsweise durch die geringen prozentualen Anzahlen der Schülerinnen und Schülern mit entweder einer sehr geringen oder einer sehr hohen Kompetenzausprägung. Zudem wurden in diesem Kapitel die gerundeten äquivalenten Ergebniswerte dargestellt, die an sich bereits ein gewisses ‚Rauschen' in den Daten verursachen.

Die Ergebnisse des equipercentilen Linking zeigen, dass der NEPS-Test auf der Metrik des TIMSS-Tests einen Mittelwert von 545 Kompetenzpunkt mit einer Standardabweichung von SD = 63 aufweist. Dies entspricht genau dem Mittelwert, den die Schülerinnen und Schüler im TIMSS-Test erreicht haben und ebenfalls fast der Standardabweichung. Diese weicht lediglich um einen Punkt ab (SD = 64).

Auch die Verteilung auf die Kompetenzstufen ist für die äquivalenten Ergebniswerte ausreichend gut. Die höchste Abweichung zwischen den prozentualen Verteilungen auf den

Kompetenzstufen zwischen den Ergebnissen, die mit dem TIMSS-Test ermittelt wurden, und den Ergebnissen, die mit den äquivalenten Ergebniswerten von NEPS auf der TIMSS-Metrik geschätzt wurden, beträgt 1.5 %. Die Übereinstimmung der Zuordnungsgenauigkeit ist mit einem Cohen's Kappa von .34 für die Gesamtgruppe zufriedenstellend. Der χ^2-Test konnte darüber hinaus keine signifikanten Gruppenunterschiede aufzeigen. Ein detaillierterer Vergleich hat gezeigt, dass die Abweichungen vor allem in den unteren Kompetenzstufen zu finden sind. Beispielsweise wurden die Schülerinnen und Schüler, die im TIMSS-Test der ersten Kompetenzstufe zugeordnet werden, im NEPS-Test (TIMSS-Metrik) im NEPS-Test der zweiten oder sogar dritten Kompetenzstufe zugeordnet. Nichtsdestotrotz ist die Zuordnungsgenauigkeit mit einer korrekten Zuordnung von 43.5% der Fälle bei fünf Kompetenzstufen als angemessen einzustufen. Andere Linking-Studien, wie beispielsweise die Studie von Pietsch et al. (2009), erreichten im Vergleich hierzu nur eine mittlere Klassifikationskorrektheit von 33%. Dennoch lässt die Ungenauigkeit in der Zuordnung zu den Kompetenzstufen nur eine vorsichtige Interpretation der Ergebnisse zu.

Anschließend wurde das Linking bezüglich der Invarianz über Subgruppen evaluiert. Der Fokus lag hierbei auf Unterschieden hinsichtlich der Geschlechter. Für die Evaluation wurden zunächst paarweise Statistiken berechnet und darauf aufbauend grafische Darstellungsformen zur Gegenüberstellungen der Gruppen genutzt. Es hat sich gezeigt, dass die Jungen bei einer Transformation der Gesamtgruppe im Gegensatz zu den Mädchen etwas übervorteilt werden. Die Gesamtgruppe scheint demnach inkonsistent bezüglich der Subgruppe Geschlecht. Jedoch sagen die Zahlen nichts darüber aus, ob die Unterschiede tatsächlich bedeutend sind. Hierfür wurde der wREMSD und der ewREMSD berechnet, der mit Hilfe des Richtwerts DTM bzw. sDTM interpretiert werden kann. Sowohl der wREMSD als auch der ewREMSD liegen deutlich unter der Grenze der DTM von .5. Wird jedoch die sDTM herangezogen, zeigt sich, dass die Werte des wREMSD und des ewREMSD leicht über der sDTM liegen und damit nennenswerte Gruppenunterschiede vorliegen. Kolen und Brennan (2010) geben allerdings zu bedenken, dass es sich bei der DTM nur um einen Richtwert, jedoch nicht um ein Bewertungsurteil handelt.

Die grafische Darstellung der Gruppenunterschiede verdeutlicht, dass die höheren Abweichungen zwischen den Gruppen vor allem in den Randbereichen auftreten und dass die Abweichungen zwischen den Geschlechtergruppen am höchsten sind. Die Unterschiede

zwischen der Gesamtgruppe und der Jungen bzw. der Mädchen sind zwischen den Kompetenzwerten 314 und 624 bzw. 288 und 610 relativ stabil. Die größten Unterschiede in diesen Bereichen liegen bei fünf Kompetenzpunkten.

Zusammenfassend lässt sich festhalten, dass die Ergebnisse des equipercentilen Linking bezüglich deskriptiver Statistiken sehr robust sind, die Verteilung auf die Kompetenzstufen hingegen nicht stabil ist. Diese Ergebnisse gehen konform mit den Studien von beispielsweise Brown et al. (2005) und Pietsch et al. (2009). Wenn die NEPS-Ergebnisse auf der internationalen TIMSS-Metrik interpretiert werden sollen, ist dies zu berücksichtigen, d. h. die Ergebnisse des NEPS-Tests auf der TIMSS-Metrik lassen einen Vergleich auf Populationsebene zu, auf Einzelebene – also um Aussagen über einzelne Schülerinnen und Schüler zu treffen - sollten die Ergebnisse hingegen nicht bzw. nur sehr vorsichtig interpretiert werden.

Die Ergebnisse zeigen, dass das equipercentile Linking für die vorliegenden Daten eine geeignete Methode ist. Nissen et al. (2015) konnten darüber hinaus zeigen, dass mit der Methode des IRT-Linking sehr ähnliche Ergebnisse erzielt werden konnten. Da die Nutzung unterschiedlicher Methoden zu unterschiedlichen Ergebnissen führen kann (Lord & Wingersky, 1984) gilt ein Vergleich der Ergebnisse zweier Linking-Methoden als eine zusätzliche Form der Qualitätskontrolle (Kolen & Brennan, 2010).

6.2 Linking des NEPS- und Ländervergleichs-Tests

Im vorherigen Kapitel wurden die Ergebnisse des NEPS-Tests auf der internationalen Metrik der TIMS-Studie verortet. In diesem Kapitel soll nun analog hierzu eine Verankerung der NEPS-Ergebnisse aus der Linking-Studie auf der nationalen Metrik des Ländervergleichs erfolgen. Dieses Linking umfasst ebenfalls drei Schritte. In einem vierten Schritt wäre es möglich, die Ergebnisse aus der NEPS-Hauptuntersuchung auf der nationalen Metrik des Ländervergleichs zu verorten. Dies ist jedoch nicht Teil dieser Arbeit.

In einem ersten Schritt werden die Ergebnisse des Ländervergleichstests in der Linking-Studie auf der Metrik der Hauptuntersuchung des Ländervergleichs 2011 verortet. Hierfür wurden bei der Skalierung die Itemparameter der Hauptuntersuchung fixiert. In die Skalierung gingen insgesamt 209 Mathematikitems des Ländervergleichs mit ein und die Daten von 752 Schülerinnen und Schülern, die an beiden Tests teilgenommen haben. Wie in der

Hauptuntersuchung wurde bei der Skalierung der Daten der Linking-Studie ein 1-PL-Modell angenommen. Da kein Hintergrundmodell verwendet wurde, werden in der Linking-Studie WLEs statt PVs berechnet. Dies erfolgte aus Gründen der besseren Vergleichbarkeit. (vgl. Kapitel 3.6.3). Anschließend wurden die Ergebnisse in die 500er-Metrik des Ländervergleichs transferiert. In einem zweiten Schritt wurde der NEPS-Test der Linking-Studie auf der Metrik der NEPS-Hauptuntersuchung verortet, indem die Itemparameter für die 24 Mathematikitems aus der Hauptuntersuchung übernommen wurden. Insgesamt wurden die Daten von 752 Schülerinnen und Schülern in einem 1-PL-Modell geschätzt.

Die Tabelle 6.11 zeigt die deskriptiven Statistiken des NEPS- und des Ländervergleichstests. Die Schülerinnen und Schüler erreichen im NEPS-Test im Mittel einen WLE von -.48 (SD = 1.18). Der höchste WLE der erreicht wurde liegt bei θ = 4.00, der niedrigste bei θ = -3.55. Die Verteilung ist leicht rechtsschief und steiler als eine Normalverteilung. Sowohl für die Gesamtgruppe als auch für die Gruppe der Schüler ist dies Abweichung in der Schiefe signifikant. Da die Schiefe kleiner als 1 ist, ist diese Abweichung nach Miles (2001) jedoch noch als moderat zu beurteilen. Hinsichtlich der NEPS-Ergebnisse zeigen sich geschlechtsspezifische Unterschiede. Die Stichprobe besteht aus 391 Schülern (52%) und 361 Schülerinnen (48%). Die Schüler erreichen im Mittel einen etwas besseren WLE (MW = -.35, SD = 1.22) als die Schülerinnen (MW = -.62, SD = 1.13). Dies ergibt eine Effektstärke von d = .23, d. h. die Schüler erreichen im Durchschnitt einen um etwa ein Viertel der Standardabweichung der Gesamtgruppe höheren Wert als die Mädchen.

Im Ländervergleichs-Test erzielen die Schülerinnen und Schüler im Mittel einen Kompetenzwert von 508 (SD = 96). Der höchste Kompetenzwert, der erreicht wurde, liegt bei 897 Kompetenzpunkten und der niedrigste bei 201 Kompetenzpunkten. Die Verteilung ist geringfügig rechtsschief und hat eine positive Wölbung, weicht aber nicht signifikant von der Normalverteilung ab. Auch beim Ländervergleichstest zeigen sich geschlechtsspezifische Unterschiede. Die Schüler erreichen im Mittel einen höheren Kompetenzwert (MW = 519, SD = 99) als die Schülerinnen (MW = 495, SD = 91). Der Unterschied ist beim Ländervergleich (d = .25) tendenziell etwas größer als bei NEPS (d = .23).

Die Korrelationen zwischen dem NEPS- und dem Ländervergleichstest liegen für die Gesamtstichprobe bei r = .77 (latente Korrelation r = .92, vgl. Kapitel 5.2.2), für die Schüler bei

r = .76 und die Schülerinnen bei r = .77. Aus diesen moderaten Ergebnissen kann angenommen werden, dass die Tests Gemeinsamkeiten verbinden, aber sie zeigen auch, dass die Tests etwas Unterschiedliches messen. Dies gilt für die Schüler tendenziell etwas mehr als für die Schülerinnen.

Tabelle 6.11: Deskriptive Statistiken für den NEPS- und den Ländervergleichstest

		N	Min	Max	MW	SD	Schiefe	Kurt
NEPS	Gesamt	752	-3.55	4.00	-0.48	1.18	0.27	3.33
	männlich	391	-3.55	4.00	-0.35	1.22	0.35	3.49
	weiblich	361	-3.55	2.16	-0.62	1.13	0.1	2.92
Länder-vergleich	Gesamt	752	201	897	508	96	0.03	3.39
	männlich	391	201	897	519	99	-0.16	3.46
	weiblich	361	237	795	495	91	0.21	3.51

Offensichtlich wird, dass die Ergebnisse des NEPS-Tests auch nicht direkt mit den Ergebnissen des Ländervergleichs-Test verglichen werden können, da die Ergebnisse auf unterschiedlichen Metriken liegen. Die Verlinkung erfolgt in einem dritten Schritt, wiederum mit der Methode des equipercentilen Linking mit der Computersoftware LEGS 2.0.1. Die Ergebnisse aus dem NEPS-Test wurden hierfür in ganzzahlige, positive Werte mit der Formel 1 (vgl. Kapitel 6.1) transferiert.

Das equipercentile Linking erfolgt über die Häufigkeitsverteilungen sowie die Perzentilränge des NEPS- (x) und des Ländervergleichs-Tests (y). Die Tabelle 6.13 stellt einen Ausschnitt dieser Häufigkeitsverteilung (N) und der Perzentilränge (P(x)) für den NEPS-Test dar, sowohl für die Gesamtpopulation als auch getrennt nach Geschlecht. Die gleichen deskriptiven Statistiken für den Ländervergleich sind der Tabelle 6.12 zu entnehmen. Der NEPS-Test hat insgesamt deutlich weniger unterschiedliche Punktschätzer als der Ländervergleichs-Test. Dies liegt daran, dass der NEPS-Test nur 24 Aufgaben und der Ländervergleichs-Test in der Linking-Studie 209 Aufgaben hat. Daher kommen auch nicht alle theoretisch möglichen Punktschätzer vor. So entstehen die Leerstellen zwischen den erreichten Punktwerten. So gibt

Tabelle 6.12: Verteilung für die Ergebnisse des NEPS-Mathematiktests für die kombinierte Gruppe und getrennt nach Geschlecht

		gesamt		**männlich**		**weiblich**	
X	**X_T**	**N**	**P(x)**	**N**	**P(x)**	**N**	**P(x)**
-3.55	145	3	0.199	1	0.128	2	0.277
-3.26	174	3	0.598	1	0.384	2	0.831
-3.02	198	7	1.263	3	0.895	4	1.662
-2.81	219	8	2.261	5	1.918	3	2.632
-2.62	238	7	3.258	3	2.941	4	3.601
-2.44	256	8	4.255	2	3.581	6	4.986
	.	.	.	.	.	.	.
	.	.	.	.	.	.	.
	.	.	.	.	.	.	.
-0.74	426	25	43.684	12	39.642	13	48.061
-0.64	436	22	46.809	14	42.967	8	50.970
-0.54	446	19	49.535	9	45.908	10	53.463
-0.44	456	31	52.859	14	48.849	17	57.202
-0.34	466	27	56.715	14	52.430	13	61.357
-0.24	476	20	59.840	10	55.499	10	64.543
	.	.	.	.	.	.	.
	.	.	.	.	.	.	.
	.	.	.	.	.	.	.
1.56	656	6	95.346	5	94.246	1	96.537
1.74	674	13	96.609	8	95.908	5	97.368
2.16	716	13	98.338	6	97.698	7	99.030
2.43	743	1	99.269	1	98.593	0	100
2.77	777	2	99.468	2	98.977	0	100
4	900	3	99.801	3	99.616	0	100

es beim NEPS-Test beispielsweise den Punktschätzer $X_T = 145$ ($X = 3.55$) und als nächstes folgt der Wert $X_T = 174$ ($X = -3.26$). Die Werte dazwischen kommen nicht vor und erhalten daher alle den gleichen Perzentilrang. In diesem Fall erhalten die Punktschätzer 146 bis 173 alle den Perzentilrang .399.

Tabelle 6.13: Verteilung für die Ergebnisse des Ländervergleichs-Mathematiktests für die kombinierte Gruppe und getrennt nach Geschlecht

	gesamt		**männlich**		**weiblich**	
Y	**N**	**Q(y)**	**N**	**Q(y)**	**N**	**Q(y)**
201	1	0.066	1	0.128	0	0
221	1	0.199	1	0.385	0	0
237	2	0.399	1	0.641	1	0.138
257	1	0.598	1	0.897	0	0.276
265	1	0.731	1	1.154	0	0.276
268	1	0.864	0	1.282	1	0.414
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
489	6	40.957	2	35.385	4	46.961
490	8	41.888	3	36.026	5	48.204
491	4	42.686	2	36.667	2	49.171
493	6	43.351	3	37.308	3	49.862
494	7	44.215	3	38.077	4	50.829
495	5	45.013	1	38.590	4	51.934
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
737	2	98.670	2	98.974	0	98.343
738	3	99.003	1	99.359	2	98.619
740	1	99.269	1	99.615	0	98.895
758	3	99.535	0	99.744	3	99.309
795	1	99.801	0	99.744	1	99.862
897	1	99.934	1	99.872	0	100

Über die Perzentilränge werden nun die Scores miteinander in Beziehung gesetzt. Diese Verlinkung erfolgt mit der Computersoftware LEGS. Die gerundete Konkordanz-Tabelle für die

Tabelle 6.14: Äquivalente Ergebniswerte des NEPS-Tests auf der Ländervergleichsmetrik

	no smoothing			S = .3			S = 1.0		
X	**ges.**	**m**	**w**	**ges.**	**m**	**w**	**ges.**	**m**	**w**
145	221	201	268	201	201	226	201	201	226
146	237	201	278	203	201	277	203	201	277
147	237	201	278	205	202	277	205	202	277
148	237	201	278	207	202	277	207	202	277
149	237	201	278	209	203	277	209	203	277
150	237	201	278	211	204	277	211	204	277
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
475	528	536	521	530	536	521	531	536	521
476	534	536	521	534	536	521	532	536	522
477	536	537	527	535	536	525	535	536	525
478	536	537	527	535	536	526	535	536	525
479	536	537	527	535	536	526	535	537	526
480	536	537	527	536	537	527	536	537	527
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
895	758	738	892	892	738	893	892	738	893
896	758	738	893	893	738	894	893	738	894
897	758	738	894	894	738	895	894	738	895
898	758	738	895	895	738	896	895	738	895
899	758	738	896	896	738	896	896	738	896
900	795	740	897	897	844	897	897	844	897

Verlinkung des NEPS-Tests auf die Metrik des Ländervergleichstests (X → Y) wird in Tabelle 6.13 dargestellt. Die Tabelle stellt einen Ausschnitt, selektiert nach geringen, mittleren und hohen NEPS-Ergebniswerten, dar. Die Tabelle gibt die äquivalenten Ergebniswerte von dem NEPS-Test auf der Metrik des Ländervergleichs aus, sowohl für die nicht geglätteten

Ergebniswerte (no Smoothing) als auch für die im Nachhinein geglätteten Werte (Postsmoothing) für S = .3 und S = 1.0 (vgl. Kapitel 3.6.3). Wenn S = 0 wäre, so würde sich an der Transformationsfunktion nichts ändern und die äquivalenten Ergebniswerte wären gleich der nicht geglätteten Ergebniswerte. Desto höher der Wert S ist, desto mehr wird die Funktion geglättet und wird annähernd zu einer Gerade. Es wurden jeweils drei Transformationen vorgenommen: (1) für die gesamte Gruppe, (2) für die Stichprobe der Schüler und (3) für die Stichprobe der Schülerinnen. Zusätzlich werden die Ergebnisse in den Abbildungen 6.5 und 6.6 graphisch veranschaulicht.

Es zeigt sich, dass die Verteilungen im Bereich der mittleren Kompetenzwerte annähernd linear sind. Nur im unteren und im oberen Kompetenzbereich gibt es Abweichungen, die vor allem für die Stichproben getrennt nach Geschlecht unterschiedlich sind. Dies sind auch die Bereiche, wo sich die geglättete Verteilung von der nicht geglätteten Verteilung unterscheidet. Die Abweichungen in den Randbereichen könnten daraus resultieren, dass die Kompetenzwerte in den Randbereichen von weniger Schülerinnen und Schülern erreicht wurden, als die mittleren Kompetenzwerte, dass es Schülerinnen und Schüler gibt die entweder alles falsch oder alles richtig beantwortet haben, was mit einer ungenaueren Schätzung einhergehen kann, und dass zum anderen die gerundeten Kompetenzwerte bereits ein gewisses Maß an Rauschen verursachen. Auffällig ist, dass – bei einer Differenzierung nach Geschlecht – die Schüler im unteren und im oberen Randbereich weniger Kompetenzpunkte bei den äquivalenten Ergebniswerten erhalten als die Schülerinnen, obwohl die Schüler im Mittel einen höheren Kompetenzwert erhalten als die Schülerinnen. Ein Grund hierfür könnte ebenfalls die geringe Vorkommenshäufigkeit der unteren und höheren Kompetenzwerte vor allem bei den Schülern sein. Dadurch steigt die Höhe der Perzentilränge bei den Schülern zu Beginn und am Ende langsamer an als bei den Schülerinnen. Dadurch bekommt bei einer Transformation getrennt nach Geschlecht im Vergleich ein Schüler bei gleichem Kompetenzwert in NEPS einen niedrigeren äquivalenten Kompetenzwert im Ländervergleich.

Abbildung 6.5: Equipercentiles Linking für die Gesamte Gruppe und differenziert nach Geschlecht – no smoothing

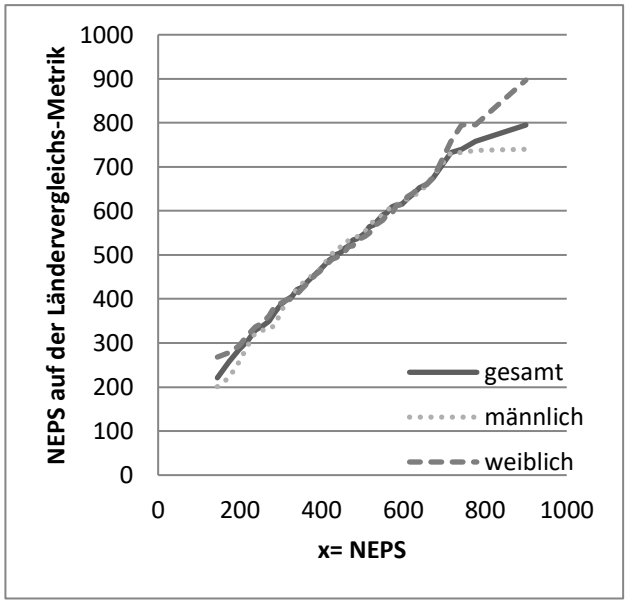
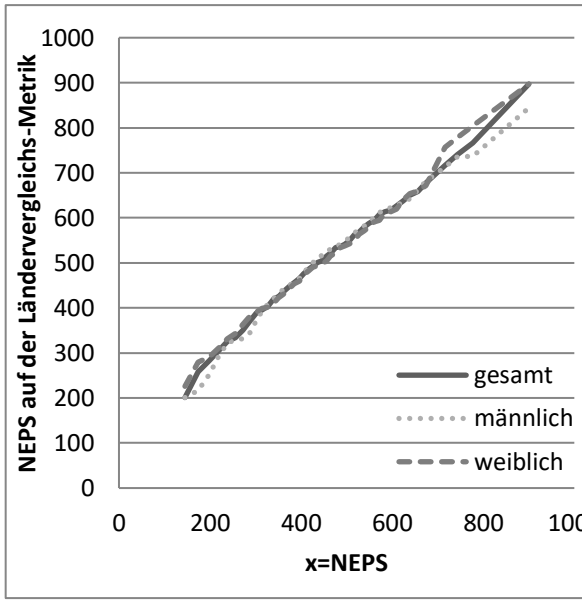


Abbildung 6.6: Equipercentiles Linking für die Gesamte Gruppe und differenziert nach Geschlecht - postsmoothing (S = 1.0)



Wie bereits aufgeführt, liegt die Annahme nahe, dass die Ungenauigkeiten in den Randbereichen der Stichprobe geschuldet sind. Es besteht die Möglichkeit sogenannte Ausreißer aus der Stichprobe zu entfernen oder die Daten zu trimmen. Eine Optimierung der Daten und damit auch der Modellpassung vorzunehmen ist beispielsweise empfehlenswert, wenn abweichende Antwortmuster analysiert werden sollen oder wenn mit bearbeitungsbedingten Ungenauigkeiten wie fehlender Testmotivation oder mangelnder Konzentration zu rechnen ist (Rost, 1996). Um aufzuzeigen, dass die Abweichungen in den Randbereichen nur eine geringe Anzahl von Schülerinnen und Schülern betrifft und das Linking für den Großteil (90%) der Stichprobe akzeptable Werte liefert, wurden zur Veranschaulichung die oberen und unteren 5% der Stichprobe getrimmt. Die Ergebnisse sind den Abbildungen 6.7 und 6.8 zu entnehmen. Es zeigt sich, dass das Linking für 90% der Stichprobe kaum noch Abweichungen in den Randbereichen aufweist und damit eine gute Passung liefert.

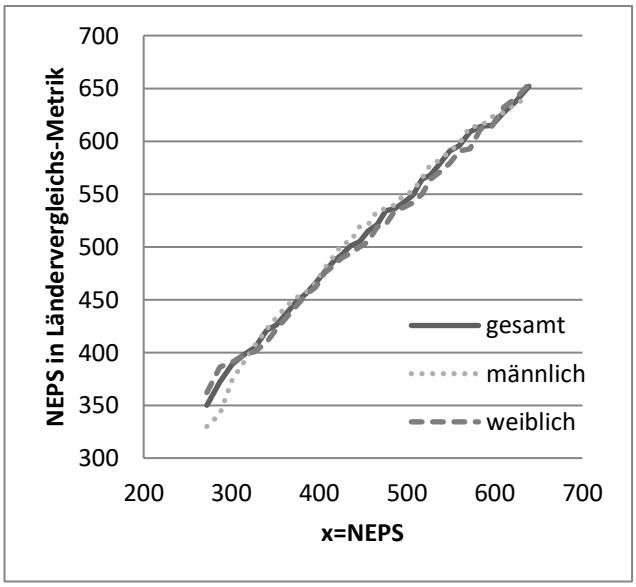


Abbildung 6.7: Equipercentiles Linking für die gesamte Gruppe und differenziert nach Geschlecht – no smoothing – trimming 5 % - 95%

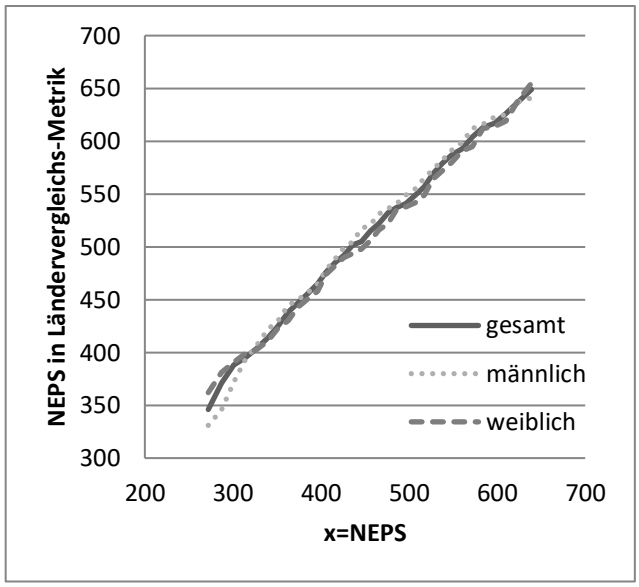


Abbildung 6.8: Equipercentiles Linking für die gesamte Gruppe und differenziert nach Geschlecht - postsmoothing (S = 1.0) – trimming 5 % - 95%

Die deskriptiven Statistiken für die nicht geglätteten und ungetrimmten Ergebnisse werden in Tabelle 6.15 dargestellt und im Vergleich zu den deskriptiven Ergebnissen der nicht verlinkten Daten gesetzt. Es zeigt sich, dass die Ergebnisse des Linking bezüglich der deskriptiven Statistiken nahezu identisch mit den Ergebnissen aus dem Ländervergleichs-Test sind. Dies ist eine Folge und auch ein Vorteil der Methode des equipercentilen Linking (Kolen & Brennan, 2010). Geringfügige Unterschiede lassen sich nur bei der Verteilung der Ergebnisse hinsichtlich der Schiefe und der Kurtosis finden. Die Betrachtung der minimalen Ergebniswerte getrennt für die Geschlechter zeigt, dass hier die Jungen einen geringeren minimalen Ergebniswert erreichen als die Mädchen, obwohl die Jungen im Durchschnitt etwa 24 Punkte mehr erreichen als die Mädchen. Bei den maximalen Ergebniswerten zeigt sich ein ähnliches Bild.

Die äquivalenten Ergebniswerte lassen sich in einem letzten Schritt den Kompetenzstufen des Ländervergleichs 2011 zuordnen. Die Ergebnisse werden in der Tabelle 6.17 den Ergebnissen des Ländervergleichs aus der Linking-Studie gegenübergestellt. In der Tabelle werden die Ergebnisse für die Transformation der Gesamtpopulation dargestellt. Die Verteilung nach

Tabelle 6.15: Deskriptive Statistiken für den NEPS- und den Ländervergleichs-Test sowie für die äquivalenten Ergebniswerte

		N	Min	Max	MW	SD	Schiefe	Kurt
NEPS	Gesamt	752	-3.55	4.00	-0.48	1.18	0.27	3.33
	Männlich	391	-3.55	4.00	-0.35	1.22	0.35	3.49
	Weiblich	361	-3.55	2.16	-0.62	1.13	0.1	2.92
Länder-vergleich	Gesamt	752	201	897	508	96	0.03	3.39
	Männlich	391	201	897	519	99	-0.16	3.46
	Weiblich	361	237	795	495	91	0.21	3.51
Linking (rnd/no smoothing)	Gesamt	752	221	795	508	95	0	3.2
	Männlich	391	201	740	519	98	-0.27	3.16
	Weiblich	361	268	897	495	91	0.23	3.4

Geschlecht erfolgte ebenfalls anhand der Gesamtpopulation. Es zeigt sich, dass beim Ländervergleich insgesamt knapp 10% der Schülerinnen und Schüler der Kompetenzstufe 1 zugeordnet wurden und bei den äquivalenten Ergebniswerten des NEPS knapp 11%. Diese Schülerinnen und Schüler beherrschen die technischen Grundlagen, um Routineprozeduren auszuführen, erreichen jedoch noch nicht die in den Bildungsstandards beschriebenen Mindeststandards. Auf der Kompetenzstufe 2 befinden sich etwa 21% der Stichprobe, wenn der Ländervergleichs-Test herangezogen wird. Werden die äquivalenten Ergebniswerte des NEPS-Tests zugrunde gelegt sind etwa 21% der Schülerinnen und Schüler auf dieser Kompetenzstufe und können somit Grundlagenwissen in einem klar strukturierten Kontext anwenden (Mindeststandards). Auf der dritten Kompetenzstufe befinden sich etwa 27% (Ländervergleich) bzw. 28% (NEPS) Schülerinnen und Schüler. Die dritte Kompetenzstufe wird erreicht, wenn die Schülerinnen und Schüler Zusammenhänge in einem vertrauten Kontext erkennen und nutzen können. Dies bedeutet, dass die Regelstandards erreicht werden. Der vierten Kompetenzstufe konnten in beiden Studien etwa 25% der Stichprobe zugeordnet werden und können daher das im curricularen Rahmen erforderliche Wissen sowie die Prozeduren sicher und flexibel anwenden (Regelstandards plus). Auf der höchsten Kompetenzstufe befinden sich etwa 16% der Schülerinnen und Schüler, egal welcher Test zugrunde gelegt wird. Die Schülerinnen und Schüler, die diese Kompetenzstufe erreichen,

können komplexe Probleme lösen und für die Lösung eigene Strategien entwickeln. Damit erreichen sie die Optimalstandards. Werden die Verteilungen des Ländervergleichs der Linking-Studie mit den äquivalenten Testwerten des NEPS aus dem Linking verglichen, zeigt sich, dass die Verteilungen hohe Übereinstimmungen aufweisen. Der größte Unterschied zwischen den beiden Studien beträgt 2%. Das Cohen's Kappa gibt die Übereinstimmung zwischen den Zuordnungen an. Für die Gesamtpopulation konnte ein Kappa von $\kappa = .31$ berechnet werden, welches ein ausreichendes Maß an Übereinstimmung darstellt. Darüber hinaus ergab der χ^2-Test keine signifikanten Unterschiede ($\chi^2 = 1.17$, df = 4). Hinsichtlich der Subpopulationen lässt sich festhalten, dass sich die Verteilungen für die Geschlechter etwas unterscheiden. Dies ist erwartungskonform, da die Jungen im Durchschnitt etwas besser abgeschnitten haben als die Mädchen (vgl. Tabelle 6.15). Auch bei der Verteilung auf die Kompetenzstufen zeigt sich, dass die Jungen eher den höheren Kompetenzstufen zugeordnet werden als die Mädchen. Zwischen den Verteilungen auf die Kompetenzstufen der beiden Tests wurden auf Gruppenebene keine signifikanten Unterschiede gefunden (Schüler: $\chi^2 = 3.97$, df = 4; Schülerinnen: $\chi^2 = 3.23$; df = 4).

Tabelle 6.16: Prozentuale Verteilung auf die Kompetenzstufen des Ländervergleichs für den Ländervergleichs-Test und die äquivalenten Ergebniswerte des NEPS-Tests

		Ländervergleich 2011						
		Nationale Kompetenzstufen						Cohen's Kappa
		Stufe 1	Stufe 2	Stufe 3	Stufe 4	Stufe 5	Gesamt	
Länder-	Gesamt	9.8%	21.4%	27.4%	25.3%	16.1%	100%	
vergleich	Männlich	9.2%	18.9%	23.5%	28.4%	19.9%	100%	
	Weiblich	10.5%	24.1%	31.6%	21.9%	11.9%	100%	
Linking	Gesamt	10.8%	19.4%	28.3%	25.3%	16.2%	100%	0.31
	Männlich	8.2%	19.7%	26.3%	27.1%	18.7%	100%	0.33
	Weiblich	13.6%	19.1%	30.5%	23.3%	13.6%	100%	0.28

Anmerkungen: Die Werte in der Tabelle sind gerundet. Dadurch kann die Summe der Prozente minimal von der Gesamtsumme 100% abweichen.

Die Tabelle 6.17 zeigt die prozentuale Zuordnung zu den Kompetenzstufen gemessen mit dem Ländervergleichs-Test und dem NEPS-Test auf der Ländervergleichs-Metrik im Vergleich. Es zeigt sich, dass 46% der Schülerinnen und Schüler im Ländervergleichs-Test und im NEPS-Test (Ländervergleichs-Metrik) der gleichen Kompetenzstufe zugeordnet werden können. Im Vergleich zu der Kompetenzmessung mit dem Ländervergleichs-Test wurden im NEPS-Test (Ländervergleichs-Metrik) 22% der Schülerinnen und Schüler eine Kompetenzstufe höher eingestuft und 4% der Schülerinnen und Schüler mehr als zwei Kompetenzstufen höher eingestuft. Hingegen wurden in 25% der Fälle die Leistungen der Schülerinnen und Schüler um eine Kompetenzstufe unterschätzt und in 3% der Fälle um zwei Kompetenzstufen unterschätzt.

Tabelle 6.17: Prozentuale Zuordnung zu den Kompetenzstufen - Vergleich der Schülerkompetenzen im Ländervergleichs-Test und im NEPS-Test (Ländervergleichs-Metrik)

		NEPS-Test auf Ländervergleich-Metrik					
		Stufe 1	**Stufe 2**	**Stufe 3**	**Stufe 4**	**Stufe 5**	**Gesamt**
Ländervergleichs-Test	**Stufe 1**	5.7%	4.3%	0.8%	0.0%	0.0%	10.8%
	Stufe 2	3.7%	8.6%	5.7%	1.1%	0.3%	19.4%
	Stufe 3	0.4%	7.4%	11.4%	7.3%	1.7%	28.3%
	Stufe 4	0.0%	1.1%	8.1%	11.0%	5.1%	25.3%
	Stufe 5	0.0%	0.0%	1.3%	5.9%	9.0%	16.2%
	Gesamt	9.8%	21.4%	27.4%	25.3%	16.1%	100%

Anmerkungen: Die Werte in der Tabelle sind gerundet. Dadurch kann die Summe der Prozente minimal von der Gesamtsumme 100% abweichen.

Die Tabelle 6.18 gibt einen weiteren Überblick über die Klassifikationskorrektheit zu den Kompetenzstufen gemessen mit dem Ländervergleichs-Test und dem NEPS-Test (Ländervergleich-Metrik). Der Tabelle ist zu entnehmen, wieviel Prozent der Schülerinnen und Schüler, die im Ländervergleichs-Test einer Kompetenzstufe zugeordnet wurden, auch im NEPS-Test dieser Kompetenzstufe zugeordnet wurden und wieviel Prozent einer anderen Stufe zugeordnet wurden. Damit kann die mittlere Klassifikationskorrektheit über die Kompetenzstufen hinweg bestimmt werden, die in dem vorliegenden Fall bei 47.5% liegt. Die höchsten Übereinstimmungen sind in Kompetenzstufe 1 (53 %) und 5 (55.7%) zu finden. Etwas

niedriger fallen die Übereinstimmungen hingegen für die mittleren Kompetenzstufen 2, 3 und 4 aus (zwischen 40.4% und 44.5%).

Tabelle 6.18: Klassifikationskorrektheit - Vergleich der Schülerkompetenzen im Ländervergleichs-Test und im NEPS-Test (Ländervergleich-Metrik)

		NEPS-Test auf Ländervergleichs-Metrik					
		Stufe 1	**Stufe 2**	**Stufe 3**	**Stufe 4**	**Stufe 5**	**Gesamt**
Ländervergleichs -Test	**Stufe 1**	53.1%	39.5%	7.4%	0.0%	0.0%	100%
	Stufe 2	19.2%	44.5%	29.5%	5.5%	1.4%	100%
	Stufe 3	1.4%	26.3%	40.4%	25.8%	6.1%	100%
	Stufe 4	0.0%	4.2%	32.1%	43.7%	20.0%	100%
	Stufe 5	0.0%	0.0%	8.2%	36.1%	55.7%	100%

Anmerkungen: Die Werte in der Tabelle sind gerundet. Dadurch kann die Summe der Prozente minimal von der Gesamtsumme 100% abweichen.

Invarianz über Subgruppen

Um Aussagen darüber treffen zu können, ob das Linking invariant bezüglich von Subgruppen ist, wurden unterschiedliche Berechnungen getätigt. Der Vergleich erfolgt hinsichtlich drei Untergruppen: Männlich – Gesamt (1 – 0), Weiblich – Gesamt (2 – 0) und Männlich – Weiblich (1 -2).

Die Unterschiede zwischen den Gruppen der gesamten Stichprobe und differenziert nach Geschlecht ohne Glättung werden in der Abbildung 6.9 dargestellt und mit Glättung im Nachhinein (S = 1.00) werden in Abbildung 6.10 dargestellt. Zudem werden die paarweisen Statistiken in der Tabelle 6.19 zusammengefasst. Der Tabelle sind die gewichtete mittlere Differenz (wMD), die gleichgewichtete mittlere Differenz (ewMD), die gewichtete mittlere Differenz der absoluten Werte (wMAD) sowie die gleichgewichtete mittlere Differenz der absoluten Werte (ewMAD) zwischen den skalenäquivalenten Ergebnissen aus dem equipercentilen Linking zu entnehmen. Die Ergebnisse werden für die nicht geglätteten Werte

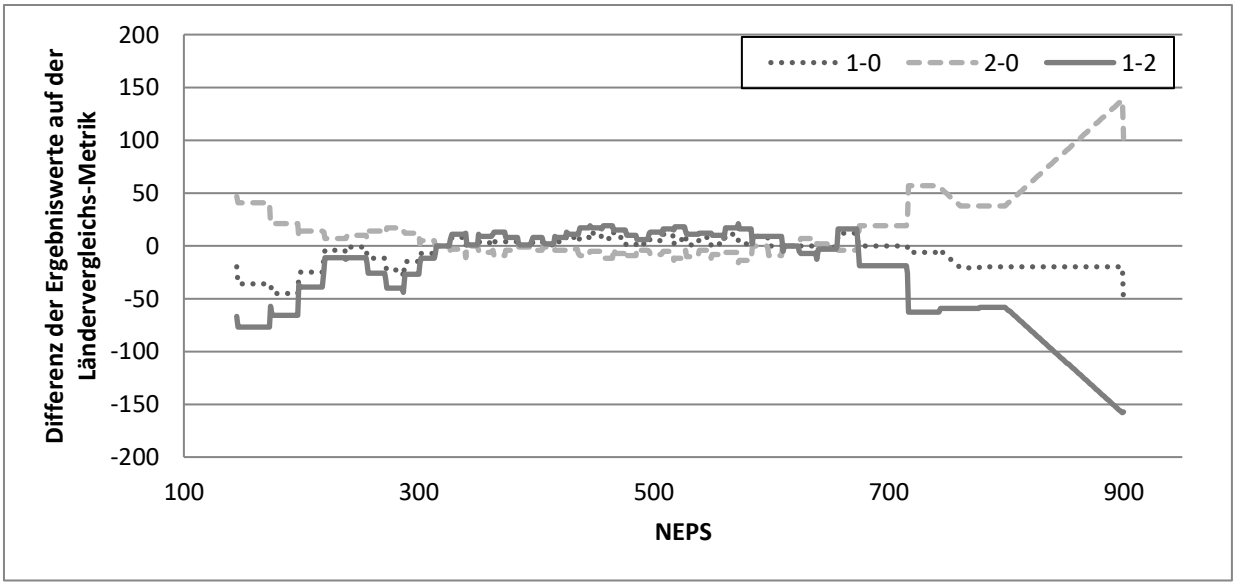


Abbildung 6.9: Differenzen der Ergebniswerte zwischen der Gesamtgruppe und den Geschlechtern – no smoothing (0 = Gesamt; 1 = Männlich; 2 = Weiblich)

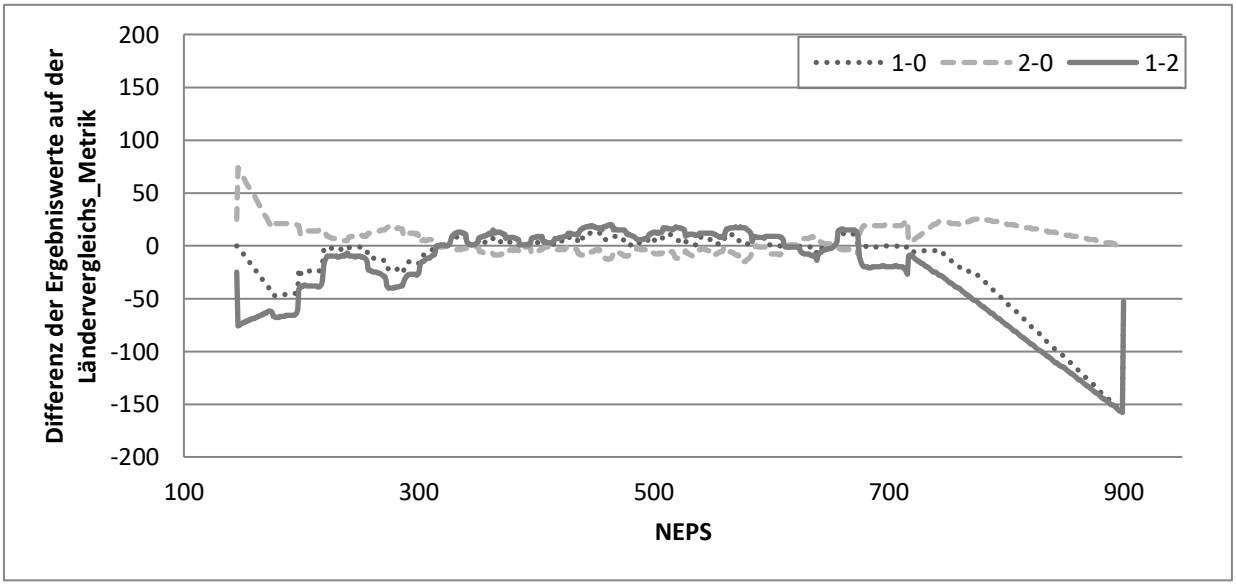


Abbildung 6.10: Differenzen der Ergebniswerte zwischen der Gesamtgruppe und den Geschlechtern – postsmoothing S = 1.0 (0 = Gesamt; 1 = Männlich; 2 = Weiblich)

und für die im Nachhinein geglätteten Werte mit S = .3 und S = 1.0 angegeben. Wenn eine Gruppen-Invarianz vorliegen würde, wären die Differenzen gleich Null. Jedoch zeigen die Grafiken und die paarweisen Statistiken, dass die Gruppen nicht invariant sind.

Beim Vergleich der Gruppen zeigt sich, dass zwischen den Transformationen getrennt nach Geschlecht die höchsten Differenzen auftreten. Für die nicht geglätteten Ergebniswerte liegt die Differenz zwischen den Geschlechtern im Mittel bei wMD = 2.88 (ewMD = -26.70,

wMAD = 13.04, ewMAD = 35.63). Weiterhin zeigt sich, dass in den Randbereichen die Schülerinnen positive Differenzen aufweisen, wohingegen die Schüler negative Differenzen im Vergleich mit der Gesamtgruppe aufweisen. Im mittleren Kompetenzbereich verhält es sich gegensätzlich. Die Differenzen zwischen den Transformationen für die Gesamtgruppe und den Schülerinnen ist im Mittel für diese Untergruppe größer (wMD = -2.00, ewMD = 19.78 bzw. wMAD = 6.53, ewMAD = 24.45) als die Differenzen der Untergruppe für die Gesamtgruppe und die Schüler (wMD = .92, ewMD = -6.93 bzw. wMAD = 6.31, ewMAD = 11.18). Auffällig ist, dass die Richtungen, in der die Gruppen sich im Mittel unterscheiden, unterschiedlich ist, je nachdem, ob der gewichtete oder der ungewichtete Mittelwert herangezogen wird. Wird der gewichtete Mittelwert betrachtet, also der stichprobenabhängige Mittelwert, zeigt sich, dass die Jungen im Vergleich zur Gesamtgruppe tendenziell etwas besser abschneiden. Dies bedeutet, dass die Jungen in der Gesamtgruppe etwas benachteiligt werden. Der Vergleich der Gruppe der Mädchen mit der Gesamtgruppe zeigt ein gegenteiliges Bild. Hier sind die Mädchen im Vergleich zur Gesamtgruppe etwas schlechter, werden demnach in der Gesamtgruppe etwas besser eingestuft und damit bevorteilt. Werden die ungewichteten Mittelwerte der Gruppendifferenzen betrachtet, zeigt sich jedoch, dass es sich hier andersherum verhält. Die Problematik der geringen Häufigkeitsverteilung in den Randbereichen wurde bereits im vorherigen Kapitel diskutiert und kann auch hier wieder eine Rolle spielen. Zudem zeigt sich, dass die ungewichteten Mittelwertsdifferenzen höher ausfallen als die gewichteten. Dies liegt daran, dass bei den gewichteten Mittelwertsdifferenzen die Vorkommenshäufigkeiten der jeweiligen Ergebniswerte mit in die Berechnung eingehen. Dadurch werden die Randbereiche, in denen die höchsten Abweichungen vorkommen, hier nicht so stark gewichtet.

Die Ergebnisse legen die Vermutung nahe, dass das Linking nicht invariant bezüglich der Geschlechter ist. Dies soll jedoch an dieser Stelle noch durch die Berechnung des REMSD als Maß für Gruppenunterschiede überprüft werden.

Die bisherigen Ergebnisse basieren auf den gerundeten äquivalenten Ergebniswerten. Zur Evaluation des Linking bezüglich von Gruppen-Invarianzen wird der REMSD und der ewREMSD angegeben (vgl. Kapitel 3.6.4). Die Ergebnisse beziehen sich hierbei jedoch auf die

Tabelle 6.18: Paarweise Statistiken für die Untergruppen Gesamt, Männlich und Weiblich (no smoothing, s=.3 und S=1.0)

Untergruppen	Methode	wMD	ewMD	wMAD	ewMAD
1-0	Equi (no smoothing)	0.92	-6.93	6.31	11.18
	Equi (S = .3)	1.24	-18.36	6.09	22.66
	Equi (S = 1.0)	1.72	-18.36	6.11	22.66
2-0	Equi (no smoothing)	-2.00	19.78	6.53	24.45
	Equi (S = .3)	-2.61	8.82	6.74	13.50
	Equi (S = 1.0)	-2.18	6.72	6.12	11.33
1-2	Equi (no smoothing)	2.88	-26.70	13.04	35.63
	Equi (S = .3)	3.92	-27.17	12.76	36.15
	Equi (S = 1.0)	3.97	-25.08	12.16	33.98

0 = Gesamt; 1 = Männlich; 2 = Weiblich

ungerundeten Werte. Das Nutzen der ungerundeten Werte ist Voraussetzung für eine Interpretation der REMSD hinsichtlich ihrer Höhe. Für die Interpretation der Werte kann eine sogenannte ‚score difference that matters' (DTM) berechnet werden (vgl. hierzu Kapitel 3.6.4). Die DTM entspricht einem halben transformierten Ergebniswert, in dem vorliegenden Fall ist die DTM = .50. Um die DFM zu standardisieren (sDFM) und sie damit auf die Metrik des REMSD zu bringen, wird die DFM durch die Standardabweichung geteilt. In diesem Fall ist die sDTM = .5/96 = .01. Die Werte verdeutlichen, warum es relevant ist, die ungerundeten Werte heranzuziehen. Wird beispielsweise die Differenz zwischen Jungen und Mädchen betrachtet und die Jungen weisen einen äquivalenten Ergebniswert von 237.3 auf und die Mädchen von 237.6, dann ist die Differenz mit .3 zwar nicht bedeutend, werden die Werte jedoch gerundet entsteht eine Differenz von 1.0 und wäre damit bedeutend. Hingegen wäre bei einem äquivalenten Ergebniswert von 236.6 für die Jungen und von 237.4 für die Mädchen die Differenz von .8 bedeutend, werden die Werte jedoch gerundet, so sind die äquivalenten Ergebniswerte von Jungen und Mädchen gleich, so dass sie nicht bedeutend wären.

Werden nun also die REMSD für die ungerundeten Ergebniswerte betrachtet (vgl. Tabelle 6.19), zeigt sich, dass alle Werte unter der DTM von .50 jedoch nicht unter der sDTM von .01

liegen. Zudem zeigt sich, dass die Berücksichtigung der Gewichte zwar einen Einfluss auf die Höhe des REMSD hat, jedoch nicht darauf, ob die Werte unterhalb oder oberhalb der DTM bzw. sDTM liegen.

Tabelle 6.19: wREMSD und ewREMSD Statistiken für das equipercentile Linking (no smoothing, S=.3 und S=1.0)

	no smoothing	**s=0.3**	**s=1.0**
wREMSD	0.105	0.094	0.085
ewREMSD	0.307	0.346	0.338

Zwischenfazit

In dem vorliegenden Kapitel wurde der Frage nachgegangen, ob das Linking der Mathematiktests des Ländervergleichs und des NEPS eine hohe Exaktheit aufweist hinsichtlich deskriptiver Statistiken und der Verteilung auf die Kompetenzstufen. Wird der NEPS-Test unter Annahme eines 1-PL-Modells mit fixierten Itemparametern aus der Hauptuntersuchung skaliert, ergibt sich für die Stichprobe von 752 Schülerinnen und Schülern ein mittlerer WLE von θ = -0.48 mit einer Standardabweichung von SD = 1.18. Die gleichen Schülerinnen und Schüler erhalten im Ländervergleichstest einen Mittelwert von 508 Kompetenzpunkten mit einer Standardabweichung von SD = 96. Im Durchschnitt sind die Kompetenzen der Jungen gemessen mit beiden Tests etwas besser als die der Mädchen. Der Unterschied ist im Ländervergleich etwas höher als im NEPS. Weitere Aussagen über die Vergleichbarkeit der beiden Tests können anhand dieser Werte jedoch nicht getroffen, da die beiden Tests keine vergleichbare Metrik aufweisen. Um die beiden Tests auf eine Metrik zu bringen, wurden die Tests mit Hilfe der Methode des equipercentilen Linking verlinkt. Die Verlinkung erfolgte mit der Computersoftware LEGS 2.0.1 unter Verwendung der Häufigkeitsverteilungen. Das equipercentile Linking wurde für die nicht geglätteten Werte und die geglätteten Werte mit S = 0.3 und S = 1.0 durchgeführt. Die grafische Darstellung der Ergebnisse zeigt, dass die Verteilungen der NEPS-Werte und der äquivalenten Ergebniswerte auf der Ländervergleichs-Metrik im mittleren Kompetenzbereich annähernd linear sind. In den Randbereichen, also für sehr hohe und sehr niedrige Kompetenzen, zeigen sich jedoch Abweichungen vor allem hinsichtlich der Unterschiede zwischen der Gesamtgruppe und den Subgruppen getrennt nach

Geschlecht. Dies könnte dem Umstand geschuldet sein, dass die Randbereiche geringe Vorkommenshäufigkeiten aufweisen, dadurch, dass es Schülerinnen und Schüler gibt, die im NEPS-Mathematiktest entweder alle Aufgaben oder keine Aufgabe richtig gelöst haben, oder dass die Abweichungen durch die Rundung der Kompetenzwerte auftreten. Werden die Ausreißer ignoriert und nur die Schülerinnen und Schüler betrachtet, die Kompetenzen im mittleren Bereich aufweisen, also, wenn die oberen und unteren 5 % der Stichprobe getrimmt werden, dann sind kaum noch Abweichungen in den Randbereichen sichtbar. Zusammenfassend lässt ich also festhalten, dass sich die Verteilungen sehr ähnlich sind und das Linking in den mittleren Kompetenzbereichen stabil ist, jedoch innerhalb der Randgruppen (z. B. für Ausreißer) Abweichungen bzw. Ungenauigkeiten vorhanden sind, die zu Fehlinterpretationen führen könnten. Dies ist ein Nachteil des equipercentilen Linking. Diese Methode ist sehr ‚sensibel' hinsichtlich Irregularitäten in den Verteilungen der Tests. Eine Möglichkeit, diesem Problem zu begegnen, ist das Smoothing (Dorans et al., 2011; Livingston, 2004; Yin et al., 2004). Werden die geglätteten Werte betrachtet, zeigt sich, dass die Abweichungen in den Randbereichen durch das Glätten der Daten annähernd verschwinden. Hier hat das Glätten der Werte den höchsten Effekt. Daher bietet es sich an, die geglätteten äquivalenten Ergebniswerte bei der Interpretation der Daten heranzuziehen.

Werden die zentralen Tendenzen betrachtet, zeigt sich, dass das Linking auf Gruppenebene sehr exakt ist. Ein Vergleich der Ergebnisse aus dem Ländervergleichs-Test mit den äquivalenten Ergebniswerten des NEPS-Tests zeigt, dass die Verteilungskennwerte wie Mittelwert, Standardabweichung, Schiefe und Kurtosis nahezu identisch sind. Der NEPS-Test weicht jedoch in der Schiefe für die Gesamtgruppe und auch für die Gruppe der Schüler signifikant von der Normalverteilung ab. Da die Schiefe unter 1 liegt, ist diese Abweichung nach Miles (2001) jedoch noch als moderat zu beurteilen. Die Übereinstimmungen zwischen den Verteilungsmerkmalen der Tests sind ein Ergebnis und auch ein großer Vorteil, den das equipercentile Linking mit sich bringt.

Hinsichtlich der Verteilung auf die Kompetenzstufen kann konkludiert werden, dass die Klassifikationskorrektheit zu den Kompetenzstufen für die beiden Tests im Vergleich akzeptable und nicht signifikant voneinander verschiedene Werte liefert, jedoch tendenzielle Unterschiede zwischen den Verteilungen offenbart. Für die Gesamtgruppe konnte ein Cohen's Kappa von $\kappa = .31$ ermittelt werden. Dies entspricht einer mittleren Klassifikationskorrektheit

von 47.5%. Für einen Vergleich der Höhe der Klassifikationskorrektheit kann die zu erwartende Klassifikationskorrektheit nach Pietsch et al. (2009) ermittelt werden, wobei zu berücksichtigen bleibt, dass Pietsch et al. eine IRT-Transformation der Daten vornimmt und die Reliabilität eines Testes dabei auf 1 fixiert. Der NEPS-Test weist eine Reliabilität von r = .8 auf und die Tests korrelieren latent zu .92, d. h. nach Pietsch et al. ist maximal eine Klassifikationskorrektheit von etwa 42 % zu erwarten. Diese wird in dem vorliegenden Fall übertroffen. Das bedeutet, dass die Zuordnung zu den Kompetenzstufen zwar so genau wie möglich erfolgt, dennoch aber eine Ungenauigkeit von 52.5% aufweist. Damit lassen sich zwar auf Gruppenebene relativ stabile Interpretationen vornehmen, von der Interpretation auf Individualebne wird hingegen abgeraten.

Anschließend wurde untersucht, wie stabil das Linking hinsichtlich der Subgruppe Geschlecht ist. Der Vergleich erfolgte auf drei Ebenen: (1) Vergleich zwischen den Schülern und der Gesamtgruppe (1-0), (2) Vergleich zwischen den Schülerinnen und der Gesamtgruppe (2-0) und (3) Vergleich zwischen den Schülern und den Schülerinnen (1–2). Wenn eine Gruppeninvarianz vorliegen würde, wären die Differenzen zwischen den Gruppen gleich Null. Die Analysen zeigen jedoch, dass Unterschiede zwischen den Gruppen bestehen. Je nachdem, ob die zweifach gewichteten oder die gleichgewichteten Mittelwertsdifferenzen betrachtet werden, ergibt sich jedoch ein anderes Bild. Der stichprobenabhängige Mittelwert, also der zweifach gewichtete, zeigt an, dass die Jungen im Vergleich zur Gesamtgruppe tendenzielle etwas besser abschneiden, wenn die Stichprobe nach Geschlecht aufgeteilt werden würde. Die Jungen werden in der Gesamtgruppe demnach etwas benachteiligt. Im Gegensatz hierzu werden die Mädchen in der Gesamtgruppe etwas bevorteilt, da sie bei Aufteilung der Stichprobe nach Geschlecht etwas schlechter abschneiden würden. Die Werte sagen jedoch nichts darüber aus, ob die Abweichungen bezüglich der Subgruppen tatsächlich bedeutsam sind. Daher wurde zusätzlich der REMSD und der ewREMSD ermittelt. Dieser kann in Zusammenhang mit der DTM bzw. der sDTM interpretiert werden. Wird die DTM als Richtwert angelegt, sind die Unterschiede zwischen den Gruppen als nicht bedeutsam einzustufen. Wird jedoch die sDTM angelegt, liegen die Werte des REMSD und des ewREMSD leicht über der sDTM. Kolen und Brennan (2010) geben jedoch zu bedenken, dass es sich bei der DTM nur um einen Richtwert und nicht um ein Bewertungsurteil handelt. Insofern bleibt festzuhalten, dass das Linking nicht invariant bezüglich der Subgruppe Geschlecht ist, dass die Unterschiede jedoch nicht sehr groß sind. Sollen Vergleiche auf Gruppenebene erfolgen, kann die

Möglichkeit genutzt werden, die äquivalenten Ergebniswerte aus den Gruppen getrennt nach Geschlecht zu nutzen.

7 Diskussion

Ziel der vorliegenden Forschungsarbeit war der Frage nachzugehen, ob sich die kriterialen Interpretationen, die in TIMSS und im LV mit den Kompetenzstufen gegeben sind, auf die Testergebnisse der NEPS Mathematikstudie K5 übertragen lassen. Voraussetzung hierfür ist jedoch zunächst eine ausreichende Übereinstimmung zwischen den Testinstrumenten (Forschungsfrage 1) und den Verteilungen der Testwertergebnisse (Forschungsfrage 2) sowie eine hohe statistische Exaktheit und Stabilität beim Verlinken der Testinstrumente (Forschungsfrage 3). Im Folgenden sollen zunächst die Ergebnisse zu den jeweiligen Forschungsfragen zusammengefasst und diskutiert werden. Anschließend werden Limitationen aufgezeigt, die sich bei der Übertragung der Ergebnisse aus der Linking-Studie auf die NEPS-Hauptuntersuchung ergeben. Ein abschließender Ausblick soll weitere Analysemöglichkeiten sowie weitere Möglichkeiten, die Daten aus der vorliegenden Studie zu nutzen, aufzeigen.

7.1 Inhaltliche Vergleichbarkeit (F 1)

Die Exaktheit und Stabilität eines Linking hängt von der Ähnlichkeit der Rahmenkonzeptionen der zu verlinkenden Studien und den hier verwendeten Testinstrumenten ab (Kolen & Brennan, 2010; Feuer et al., 1999; Mislevy, 1992; Linn, 1993). In einem ersten Schritt wurde daher untersucht, inwiefern eine inhaltliche Vergleichbarkeit der NEPS-Mathematikstudie K5 mit den Studien TIMSS K4 und dem Ländervergleich Mathematik Primar gegeben ist (F 1). In diesem Zusammenhang wurde die Vergleichbarkeit der drei Studien hinsichtlich der folgenden Aspekte analysiert: (1a) Anlage und Ziele, (1b) Stichprobe, (1c) Messbedingungen, (1d) Konstrukte (konzeptionell) und (1e) Konstrukte (methodisch).

Zusammenfassung der Ergebnisse

(1a) Hinsichtlich der Vergleichbarkeit der Anlage und Ziele der Studien konnte aufgezeigt werden, dass die Studien viele Gemeinsamkeiten aufweisen. Beispielsweise verfolgen die Studien das gemeinsame Ziel, die mathematische Kompetenz von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe bzw. am Anfang der fünften Jahrgangsstufe zu messen. Zudem erfassen die Studien Hintergrunddaten, wie beispielsweise das Geschlecht oder den

Migrationshintergrund der Schülerinnen und Schüler, um Unterschiede zwischen Subpopulationen aufdecken zu können.

Jedoch konnten auch Unterschiede zwischen den Studien beispielsweise in den Testkonzeptionen aufgezeigt werden. TIMSS basiert auf einem internationalen Curriculum Modell, der Ländervergleich auf den nationalen Bildungsstandards und NEPS orientiert sich am Literacy Konzept und an den nationalen Bildungsstandards. Hieraus resultieren auch Unterschiede hinsichtlich der curricularen Validität der Testinstrumente. Weiterhin konnte festgestellt werden, dass NEPS die Leistungsdaten im Bereich Mathematik im Längsschnitt erfasst und keine kriteriale Interpretation anhand von Kompetenzstufen zulässt, wohingegen TIMSS und der Ländervergleich im Querschnitt erfasst werden und die Ergebnisse kriterial in einem nationalen bzw. internationalen Kontext interpretiert werden können. Dies bedeutet, dass sich der Fokus der Studien hinsichtlich der Messintention unterscheidet.

(1b) In einem zweiten Schritt wurden die Studien hinsichtlich ihrer Stichproben verglichen. Dies beinhaltete sowohl den Prozess der Stichprobenziehung als auch die Merkmale der Stichproben an sich. Es kann festgehalten werden, dass die drei untersuchten Studien TIMSS, Ländervergleich und NEPS die Kompetenzen der Schülerinnen und Schüler jahrgangsbasiert und nicht altersbasiert erfassen. Der Prozess der Stichprobenziehung weist dabei viele Gemeinsamkeiten auf. In allen Studien wurden unterschiedliche Strata berücksichtigt (wie z.B. Bundesländer) und es handelt sich um Clusterstichproben. Ein wichtiger Unterschied ist, dass die Teilnahme an TIMSS und dem Ländervergleich für die zufällig gezogenen Schulen in öffentlicher Trägerschaft – abgesehen von einigen Ausnahmefällen - verbindlich ist, wohingegen die Teilnahme an NEPS freiwillig ist. Weiterhin unterscheiden sich die Studien bezüglich der Stichprobengröße für Deutschland.

(1c) Hinsichtlich der Messbedingungen finden sich ebenfalls Gemeinsamkeiten und Unterschiede zwischen den Studien. Die Messzeitpunkte vom Ländervergleich und TIMSS sind nahezu identisch und liegen im Frühsommer des Jahres 2011. Die NEPS Mathematik K5 Erhebung erfolgte hingegen im Herbst/Winter 2010. Damit liegt etwa ein halbes Schuljahr zwischen den teilnehmenden Schülerinnen und Schülern, wobei zu berücksichtigen bleibt, dass NEPS in der fünften Jahrgangstufe und TIMSS und der Ländervergleich in der vierten Jahrgangsstufe die Kompetenzen der Schülerinnen und Schüler erhebt. Weiterhin hat sich gezeigt, dass Unterschiede in den Gesamt-Testzeiten (TIMSS: 72 Min.; LV: 160 Min.; NEPS: 100

Min.) und in den Mathematik-Testzeiten (TIMSS: 36 Min.; LV: 80 Min.; NEPS: 31 Min.) vorliegen (Bos et al., 2012; Stanat et al., 2012; Duchhardt & Gerdes, 2012a). Für die Messung der mathematischen Kompetenz nutzen zudem sowohl TIMSS als auch der Ländervergleich ein Multi-Matrix-Design. NEPS hingegen verwendet nur eine Testheftvariante zur Erfassung der mathematischen Kompetenz. Hilfsmittel sind in allen drei Studien nicht erlaubt.

(1d) Um zu analysieren, inwieweit die in den Studien definierten Konstrukte konzeptionell vergleichbar sind, wurde ein Vergleich der Rahmenkonzeptionen, ein Vergleich der erfassten mathematischen Inhalte durch ein Expertenreview sowie ein Vergleich ausgewählter Aufgabenmerkmalen ebenfalls durch ein Expertenreview vorgenommen.

Vergleich der Rahmenkonzeptionen: Die drei Studien definieren die mathematische Kompetenz in den Rahmenkonzeptionen nicht in gleicher, aber doch in einer sehr ähnlichen Weise. So unterscheiden nicht alle Studien zwischen den Bereichen Inhalt, Prozess und Anforderung und definieren hier jeweils auch nicht dieselben Subdimensionen. Die Definition hinsichtlich der inhaltsbezogenen und prozessbezogenen Kompetenzen weisen hohe Übereinstimmungen vor allem zwischen NEPS und dem Ländervergleich auf. Im Ländervergleichs-Test wird jedoch eine inhaltliche Subdimension mehr definiert als im NEPS- und TIMSS-Test. Die Definition der prozessbezogenen Kompetenzen entspricht sich bei NEPS und dem Ländervergleich approximativ. Der Ländervergleich definiert zusätzlich Anforderungsbereiche. TIMSS unterscheidet hingegen zwischen einer inhaltlichen Dimension und einer kognitiven Dimension, wobei die kognitiven Dimensionen per Definition sowohl Anforderungsbereiche als auch prozessbezogene Kompetenzen umfassen.

Es bleibt jedoch die Frage offen, ob die gleichen Unterscheidungen und Bezeichnungen auch tatsächlich gleiche Inhalte abdecken oder ob es sich um eine jingle fallacy handelt. Zusätzlich stellt sich die Frage, ob sich hinter unterschiedlichen Bezeichnungen auch tatsächlich unterschiedliche Inhalte verbergen oder ob es sich um eine jangle fallacy handelt. Inwiefern die genannten Unterschiede und Definitionen daher tatsächlich zu Unterschieden in der Erfassung des mathematischen Konstrukts führen, wurde mit einem Expertenreview untersucht.

Vergleich der mathematischen Inhalte: Die NEPS-Mathematikaufgaben lassen sich fast ausschließlich in die Rahmenkonzeptionen von TIMSS und dem Ländervergleich einordnen. Zudem unterscheiden sich die prozentualen Verteilungen der NEPS-Aufgaben zu den

Subdimensionen nicht signifikant zu denen der TIMSS und Ländervergleichsverteilungen der Hauptuntersuchungen. Die Inhaltsbereiche der NEPS- und der TIMSS- bzw. Ländervergleichs-Rahmenkonzeptionen entsprechen sich nicht eins zu eins, aber weisen sehr große Ähnlichkeiten auf, da die Unterschiede nur tendenziell und nicht statistisch abzusichern sind.

Vergleich der Aufgabenmerkmale: Die Mathematikaufgaben in den Studien weisen auf formaler und sprachlicher Ebene viele Gemeinsamkeiten auf und es zeigen sich nur wenige signifikante Unterschiede. Die Mathematikaufgaben im NEPS nutzen beispielsweise keine Tabellen und weniger Aufgaben haben einen Stimulus im Vergleich zu den Aufgaben aus dem Ländervergleich. Der NEPS-Test hat keine offenen Antwortformate wie der Ländervergleich und dafür im Vergleich mehr geschlossene Aufgabenformate. Zudem verwendet der NEPS-Test signifikant mehr Wörter und auch mehr Sätze in den Aufgabenstellungen als der Ländervergleichs-Test. Zwischen NEPS und TIMSS ergab sich ebenfalls ein signifikanter Unterschied. In TIMSS werden in den Aufgaben häufiger Graphen, Grafiken oder Diagramme verwendet als im NEPS. Hinsichtlich der Komplexität der verwendeten Sprache ließen sich keine signifikanten Unterschiede aufzeigen.

(1e) Der methodische Vergleich der drei Studien hat viele Gemeinsamkeiten, aber auch einige Unterschiede aufgezeigt. Ein wichtiger Unterschied zwischen den drei Studien besteht in der Skalierung der Daten. TIMSS nimmt ein 3-PL-Modell mit einer mehrdimensionalen Struktur und einer Between-Item-Dimensionalität an, wohingegen NEPS ein eindimensionales 1-PL-Modell annimmt und der Ländervergleich ein 1-PL-Modell mit einer mehrdimensionalen Struktur. TIMSS und der Ländervergleich schätzen im Anschluss PVs als Personenparameter und NEPS WLEs.

Diskussion

Bezüglich der ersten Forschungsfrage lässt sich festhalten, dass eine inhaltliche Vergleichbarkeit der Tests in einem gewissen Maße vorhanden ist und damit eine gute Voraussetzung für ein anschließendes Linking gegeben ist. Es handelt sich bei den Testinstrumenten jedoch nicht um kongruente und damit austauschbare Instrumente. Die festgestellten Unterschiede können einen Einfluss auf die Exaktheit und Stabilität des Linking haben. Es ist beispielsweise anzunehmen, dass – bedingt durch die Unterschiede in der curricularen Validität der Tests – der NEPS- und der TIMSS-Mathematiktest eine etwas höhere Schwierigkeit aufweisen als der Ländervergleichs-Test. Hierbei bleibt zu berücksichtigen, dass

in dem Expertenreview die curriculare Validität des NEPS-Mathematiktests am Ende der Grundschulzeit bestimmt wurde, obwohl der NEPS-Test in der Hauptuntersuchung zu Beginn der fünften Jahrgangsstufe durchgeführt wird. Zur curricularen Validität des NEPS-Test bezogen auf die fünfte Jahrgangsstufe können dementsprechend keine Aussagen getroffen werden. Ceteris paribus sollten sich mögliche Unterschiede in der Schwierigkeit bei einem Linking ausgleichen lassen, da es sich hierbei um das ursprüngliche Ziel eines Equating handelt. Ungenauigkeiten beim Linking können jedoch dadurch entstehen, dass – bedingt durch die unterschiedlichen Schwierigkeiten der Tests – in den Randbereichen keine ausreichenden Fallzahlen vorliegen. Dieses Problem ergibt sich vor allem auch in der Linking-Studie, da die Teilnehmerzahlen hier geringer sind als in den Hauptuntersuchungen. Hierauf hat ebenfalls ein Einfluss, dass der NEPS-Test zwar für die fünfte Jahrgangsstufe entwickelt, in der Linking-Studie jedoch am Ende der vierten Jahrgangsstufe eingesetzt wurde. Dadurch können gerade in dem oberen Kompetenzbereich geringere Fallzahlen erwartet werden.

Weiterhin können die unterschiedlichen Aufgabenanzahlen und dadurch bedingt auch die unterschiedlichen Testzeiten in den Studien (Ländervergleich: 330 Aufgaben, TIMSS: 177 Aufgaben und NEPS: 25 Aufgaben) einen Einfluss auf die Exaktheit und Stabilität des Linking haben (Bos et al., 2012; Stanat et al., 2012; Duchhardt & Gerdes, 2012a). Es kann davon ausgegangen werden, dass der NEPS-Mathematiktest mit seinen 25 Aufgaben gerade in den Randbereichen nicht so differenziert messen kann wie beispielsweise der Ländervergleichs-Test mit seinen 330 Aufgaben (bzw. 277 Items in der Linking-Studie). Dies könnte dazu führen, dass das Linking in den Randbereichen weniger exakt ist.

Der Vergleich der Rahmenkonzeptionen (Bos et al., 2012; Stanat et al., 2012; Duchhardt & Gerdes, 2012a) hinsichtlich der hier beschriebenen Teildimensionen mathematischer Kompetenz zeigt große Überschneidungen zwischen den Studien auf. Jedoch kann die Frage nach dem Vorhandensein einer jingle oder jangle fallacy an dieser Stelle nicht hinreichend aufgeklärt werden, da die Rahmenkonzeptionen keine erschöpfenden Beschreibungen liefern. Um die Frage nach einer jingle oder jangle fallacy beantworten zu können, müssten weitere Analyse hinsichtlich der Aufgabeneigenschaften beispielsweise hinsichtlich der Komplexität der Aufgaben durchgeführt werden. Solche Analysen könnten weitere Unterschiede aufdecken (Neidorf et al., 2006; Wu, 2010).

Darüber hinaus wurden einige Unterschiede in den formalen und sprachlichen Aufgabenmerkmalen in den Studien deutlich. Diese können ebenfalls zu Ungenauigkeiten im Linking von NEPS mit TIMSS und dem LV führen. Zum einen zeigen die Ergebnisse aus dem Expertenreview, dass in den Test unterschiedlich viel Konstrukt irrelevante Varianz vorhanden ist. Beispielsweise nutzt der NEPS-Test deutlich mehr Wörter pro Aufgabe als der TIMSS- und der Ländervergleichs-Test, so dass vor allem im NEPS-Test zusätzlich zur mathematischen Kompetenz auch die Lesekompetenz der Schülerinnen und Schüler gefordert ist. Zum anderen zeigen die Ergebnisse des Expertenreviews auch mögliche Konstruktunterrepräsentationen auf. Beispielsweise gibt es im NEPS-Test keine Aufgabe, für deren Lösung das Lesen und Interpretieren einer Tabelle gefordert ist. Darüber hinaus gibt es im NEPS-Test auch keine offenen Antwortformate. Werden diese Kompetenzen jedoch zum Konstrukt mathematischer Kompetenz gezählt, definiert NEPS demnach im Gegensatz zu TIMSS und dem Ländervergleich ein eingeschränktes Konstrukt mathematischer Kompetenz. Damit ist die Annahme der Testfairness nur eingeschränkt gegeben, denn es kann einen Unterschied machen, ob ein Schüler Test X oder Test Y vorgelegt bekommt (vgl. Kapitel 1.1.1). Beispielsweise könnte eine niedrige Lesekompetenz eines Schülers dazu führen, dass dieser Schüler im NEPS-Test schlechter abschneidet als im TIMSS- oder Ländervergleichstest, weil der NEPS-Test mehr Lesekompetenz erfordert und nicht, weil er mathematisch nicht dazu in der Lage ist, die Aufgaben zu lösen.

Unterschiede ergaben sich auch bezüglich der Wahl der Skalierungsmodelle der Leistungsdaten in den Studien. Das von TIMSS angenommene 3-PL-Modell berücksichtigt die Ratewahrscheinlichkeit bei MC Aufgaben und die Trennschärfe, wodurch im Gegensatz zu einer 1-PL-Skalierung wie bei NEPS und dem Ländervergleich Unterschiede in den Ergebnissen erwartet werden können. Es ist anzunehmen, dass hierdurch Schülerinnen und Schüler im unteren Kompetenzbereich unter Annahme eines 1-PL-Modells etwas übervorteilt werden, weil die Ratewahrscheinlichkeit nicht berücksichtigt wird, in einem 3-PL-Modell hingegen nicht. Da im Rahmen der Verlinkung von NEPS und TIMSS bzw. Ländervergleich die Daten in Anlehnung an die Hauptuntersuchungen skaliert werden, könnte sich der Unterschied in den Skalierungsmodellen auf die Exaktheit und Stabilität des Linking auswirken. Brown et al. beispielsweise konnten in einer Studie zeigen, dass die Nutzung unterschiedlicher IRT Modelle hinsichtlich zentraler Tendenzen robust ist, nicht jedoch für die Verteilung der Ergebnisse. Insofern ist bezüglich des Linking zu erwarten, dass die zentralen Tendenzen reliabel zu

interpretieren sind, auf Individualebene jedoch gerade in den Randbereichen Unterschiede durch die verschiedenen Skalierungsmodelle auftreten können.

Weiterhin unterscheiden sich die Studien hinsichtlich der angenommenen Dimensionalität der Datenstruktur. Winkelmann und Robitzsch (2009) konnten in einer Analyse zeigen, dass die Korrelationsmuster zwischen den Teilbereichen unterschiedlich hoch bzw. niedrig ausfallen, je nachdem welche Dimensionalität angenommen wird. Wird eine Within-Item-Dimensionalität angenommen ergeben sich etwas niedrigere Korrelationen als unter Annahme einer Between-Item-Dimensionalität. Diese Befunde sind wichtig für die Interpretation des dimensionalen Vergleichs, der im folgenden Kapitel vorgenommen wird.

7.2 Dimensionale und skalenbezogene Vergleichbarkeit (F 2)

Die zweite Forschungsfrage bezieht sich auf die Höhe der empirischen Zusammenhänge zwischen den Mathematiktests der NEPS-K5-Studie und den Studien TIMSS und dem Ländervergleich in der Linking-Studie. Die Analyse der Äquivalenz der Studien erfolgte hinsichtlich der dimensionalen (2a) und skalenbezogenen (2b) Vergleichbarkeit.

Zusammenfassung der Ergebnisse

(2a) Hinsichtlich der Vergleichbarkeit der Studien auf dimensionaler Ebene wurden zwei Analysen durchgeführt. In einem ersten Schritt wurden die Korrelationen der Teildimensionen im NEPS-Mathematiktest den Korrelationen der Teildimensionen in den Mathematiktests denen von TIMSS und des Ländervergleichs gegenübergestellt und miteinander verglichen. Die Analysen haben ergeben, dass die latenten Korrelationen zwischen den Inhaltsbereichen in TIMSS und NEPS eher hoch ausfallen (TIMSS: $.83 < r < .85$; NEPS: $.88 < r < .93$), wohingegen der Ländervergleich niedrigere Korrelationen der Inhaltsbereiche aufweist ($.48 < r < .70$). Dies bedeutet, dass die Inhaltsbereiche in den jeweiligen Studien einen hohen Anteil gemeinsamer Varianz haben, die Inhaltsbereiche aber dennoch zusätzlich zu einer globalen mathematischen Kompetenz stoffgebietsspezifische Kompetenzen abbilden.

In einem zweiten Schritt wurde überprüft, ob hohe Korrelationen zwischen den Teildimensionen der Tests und damit hohe Überschneidungen in den Teildimensionen aufgezeigt werden können. Um dies zu überprüfen, wurde der NEPS-Test zum einen mit dem TIMSS-Test und zum anderen mit dem Ländervergleichs-Test gemeinsam skaliert. Der

inhaltliche Vergleich hat hohe Übereinstimmungen zwischen den inhaltsbezogenen Kompetenzen in den Studien aufgezeigt. Diese Annahme konnte durch die hohen latenten Korrelationen der Inhaltsbereiche zwischen dem NEPS- und dem TIMSS-Test bestätigt werden ($.71 < r < .91$). Die Korrelationen zwischen den Inhaltsbereichen von NEPS und dem Ländervergleich fallen dagegen etwas niedriger aus als zwischen von NEPS und TIMSS und sind auch nicht ausnahmslos erwartungskonform ($.42 < r < .80$). Beispielsweise korreliert der NEPS Inhaltsbereich Quantität am höchsten mit dem Ländervergleichs Inhaltsbereich Raum und Form ($r = .80$). Die höchsten Korrelationen wären jedoch gemäß dem inhaltlichen Vergleich mit den Inhaltsbereichen Zahlen und Operationen ($r = .65$) sowie Größen und Messen ($r = .69$) zu erwarten gewesen.

(2b) Hinsichtlich der skalenbezogenen Äquivalenz wurde die Vergleichbarkeit der Reliabilitäten der Tests untersucht, da dies eine Grundvoraussetzung für ein Equating ist (vgl. Kapitel 1.1.1; Dorans & Holland, 2000; Holland & Dorans, 2006; Dorans & Walker, 2007). Zudem sollten die Reliabilitäten möglichst adäquat sein (Dorans und Holland, 2000). Die Ergebnisse zeigen, dass die Reliabilitäten des NEPS- ($\alpha = .80$) und TIMSS-Mathematiktests ($\alpha = .81$) in einem angemessenen Bereich liegen und sich hinreichend ähnlich sind (Foy, Martin et al., 2012; Duchhardt & Gerdes, 2012b). Für den Ländervergleichs-Test lag zum Zeitpunkt der Arbeit kein Reliabilitätsmaß vor (vgl. Kapitel 5.2.1). Bezüglich des skalenbezogenen Vergleichs wurde weiterhin der Frage nachgegangen, ob es sich bei den drei Tests um unterschiedliche Konstrukte oder um ein gemeinsames Konstrukt mathematischer Kompetenz handelt. Hinweise hierzu können auch die Korrelationskoeffizienten zwischen den Tests liefern (Kolen & Brennan, 2010). Dafür wurden die Daten aus der Linking-Studie des NEPS-Tests zum einen mit den Daten von TIMSS und zum anderen mit den Daten des Ländervergleichs gemeinsam skaliert und dahingehend überprüft, ob eine Ein- oder eine Zweidimensionalität der Datenstruktur vorliegt. Die Analysen ergaben, dass obwohl der NEPS-Test hoch mit dem TIMSS- ($r = .90$) und mit dem Ländervergleichs-Test ($r = .92$) latent korreliert, dennoch bei Kontrolle der Modellgeltungstests eine zweidimensionale Struktur der Daten vorliegt.

Diskussion

Bezüglich der dimensionalen Vergleichbarkeit lässt sich festhalten, dass sich in den Tests die Inhaltsbereiche in einem gewissen Maße separieren lassen, jedoch auch ein hoher Anteil durch eine globale mathematische Kompetenz beschrieben werden kann. Die Tests weisen

eine sehr ähnliche, jedoch nicht dieselbe faktorielle Struktur auf. Wobei die Ähnlichkeit bzw. die Nähe vom NEPS- zum TIMSS-Test höher ist als zum Ländervergleichs-Test. Darüber hinaus konnten Überschneidungen der Inhaltsbereiche zwischen den Tests aufgezeigt werden, wenn auch nicht in allen Fällen zwischen den theoretisch erwarteten Inhaltsbereichen. Die hohen Korrelationen zwischen den Teildimensionen zwischen dem NEPS- und dem TIMSS-Test weisen darauf hin, dass ein ähnliches Konstrukt mathematischer Kompetenz erfassen wird. Im Vergleich hierzu fallen die Überschneidungen zum Ländervergleichs-Tests niedriger aus, was darauf hindeuten kann, dass die beiden Tests ein – zumindest in Teilen – unterschiedliches Konstrukt von mathematischer Kompetenz messen. Der Vergleich auf dimensionaler Ebene ist jedoch bedingt durch die unterschiedlichen Skalierungen schwierig. So kann beispielsweise die Annahme einer Between oder auch Within-Item-Dimensionalität bereits einen Einfluss auf die Höhe der Korrelation haben. Dies könnte die unterschiedlich hohen Korrelationen zwischen den Inhaltsbereichen innerhalb der Tests erklären. Der Ländervergleichs-Test weist niedrigere latente Korrelationen auf. Dies kann daran liegen, dass der Test unter Annahme einer Within-Item-Dimensionalität skaliert wurde. Winkelmann und Robitzsch (2009) konnten diesbezüglich in einer Untersuchung der Kompetenzdaten aus der Normierungs- und Pilotierungsstichprobe des Ländervergleichs zeigen, dass die latenten Korrelationen zwischen den inhaltsbezogenen Kompetenzbereichen unter Annahme einer Within-Item-Dimensionalität niedriger ausfallen als unter Annahme einer Between-Item-Dimensionalität. NEPS und TIMSS wurden unter der Annahme einer Between-Item-Dimensionalität skaliert und die latenten Korrelationen fallen demnach erwartungskonform höher aus. Das der NEPS-Test die höchsten Korrelationen aufweist könnte auch daran liegen, dass im NEPS-Test nur wenige Items pro Inhaltsbereiche (5 bis 8 Items pro Inhaltsbereich) vorliegen (Duchhardt & Gerdes, 2012b), so dass sich eventuell auch dadurch bedingt, die Inhaltsbereiche nicht so gut voneinander separieren lassen. Zudem wurde der NEPS-Test nicht dazu entwickelt, Ergebnisse auf Ebene der Inhaltsbereiche zu berichten, sondern soll eine eindimensionale globale mathematische Kompetenz erfassen. Durch solche und andere mögliche methodische Artefakte kann der Vergleich der dimensionalen Struktur daher höchstens Tendenzen abbilden.

Bezüglich der skalenbezogenen Vergleichbarkeit lässt sich feststellen, dass die Höhe der Reliabilitäten zwischen dem NEPS- und dem TIMSS-Test nahezu äquivalent sind, wobei zu berücksichtigen bleibt, dass die Studien ein unterschiedliches Reliabilitätsmaß angeben.

Jedoch lässt die Höhe der Reliabilitäten in den Tests auch auf ein gewisses Maß an Messungenauigkeit schließen. Diese Ungenauigkeit kann einen Einfluss auf die Exaktheit und Stabilität des Linking haben, beispielsweise hat sich bei Pietsch et al. (2009) gezeigt, dass die Höhe der Reliabilität einen Einfluss auf die Zuordnungsgenauigkeit zu den Kompetenzstufen hat. Einmal angenommen der eine Test hätte eine Reliabilität von $\alpha = 1.00$ und die Tests würden eine Korrelation von $r = .90$ aufweisen, dann könnte für eine Reliabilität von $\alpha = .80$ für den zweiten Test eine Zuordnungsgenauigkeit von .42 erwartet werden (Pietsch et al., 2009). Hierbei bleibt jedoch zu berücksichtigen, dass Pietsch et al. eine IRT Transformation vorgenommen haben. Da für den Ländervergleichs-Test zum jetzigen Zeitpunkt kein Reliabilitätsmaß vorlag, können hierzu keine dementsprechenden Aussagen getroffen werden.

Die beschriebenen Unterschiede hinsichtlich der Rahmenkonzeptionen und Test-spezifikationen können Auswirkungen auf die Vergleichbarkeit der Schülerinnen- und Schülerkompetenzen in den Tests haben. Beispielsweise kann ein Schüler zwar sehr gut in der Schulmathematik abschneiden, aber gleichzeitig bei der Übertragung der Schulmathematik zur Lösung von Alltagsproblemen Schwierigkeiten haben. Die Ergebnisse der latenten Korrelationen zwischen den Studien geben daher weitere Hinweise zur Vergleichbarkeit der Tests. Es zeigt sich, dass der NEPS-Test sowohl mit TIMSS ($r = .90$) als auch mit dem Ländervergleich ($r = .92$) hoch latent korreliert. Dabei ist die Höhe der Korrelationen mit Befunden aus anderen Studien vergleichbar. Böhme et al. (2014) konnten beispielsweise zwischen den Mathematiktests des Ländervergleichs 2001 und TIMSS 2011 eine Korrelation von $r = .91$ ermitteln. Die hohen Korrelationen zwischen den Tests spricht dafür, dass ein erheblicher Anteil eines gemeinsamen globalen Faktors existiert, jedoch das zweidimensionale Modell jeweils eine bessere Passung aufweist als das eindimensionale Modell. Wird den Empfehlungen von Rost (2004) gefolgt, sollte das zweidimensionale Modell bevorzugt werden. Somit sollten die Tests trotz hoher Übereinstimmungen nicht als gegenseitig austauschbar angesehen werden.

Die Ergebnisse aus dem dimensionalen und skalenbezogenen Vergleich zeigen, dass die in den Tests gemessenen Konstrukte zwar nicht vollständig identisch sind, jedoch eine zufriedenstellende Ähnlichkeit zwischen den in den Tests gemessenen Konstrukten besteht. Damit ist zwar kein Equating möglich, jedoch sind die Annahmen für eine Concordance

gegeben (vgl. Kapitel 1.1.1), sodass zwar keine präzisen Rankings möglich sind und die Linking-Ergebnisse nicht als tatsächliche TIMSS- bzw. Ländervergleich-Werte zu interpretieren sind, sich jedoch Aussagen auf Gruppenebene treffen lassen. Vor allem sprechen die hohen empirischen Zusammenhänge zwischen den Mathematiktests von NEPS und TIMSS bzw. Ländervergleich für eine adäquate Vergleichbarkeit. Dennoch ist davon auszugehen, dass die aufgedeckten Unterschiede einen Einfluss auf die Exaktheit und Stabilität des Linking, beispielsweise auf die Klassifikationskorrektheit zu den Kompetenzstufen, haben werden. Nach Kolen und Brennan (2010) sei eine Konstruktverschiedenheit jedoch kein Grund dafür, ein Linking nicht durchzuführen. Bei der Messung von menschlichen Fähigkeiten sei es nicht ungewöhnlich, dass zwei Tests nicht ein und dasselbe Konstrukt messen. Die aufgezeigten Unterschiede sind jedoch mit Limitationen bei der Interpretation der Testwerte, die aus dem Linking abgeleitet werden, zu berücksichtigen. Um diese detaillierter zu bestimmen wurde das Linking in einem nächsten Schritt auf seine Exaktheit und Stabilität hin untersucht.

7.3 Exaktheit und Stabilität des Linking (F 3)

Im Rahmen der dritten Forschungsfrage wurde untersucht, ob sich hinsichtlich des Linking des NEPS-Mathematiktests mit den beiden Mathematiktests von TIMSS und dem Ländervergleich in der Linking-Studie eine hohe statistische Exaktheit und Stabilität nachweisen lässt. Dies ist relevant, weil statistische Schätzungen aus verschiedenen Gründen messfehlerbehaftet sein können. Die Untersuchung der ersten beiden Forschungsfragen hat viele Gemeinsamkeiten zwischen den Studien ergeben, jedoch auch einige Unterschiede, welche Auswirkungen auf die Stabilität und Exaktheit des Linking haben können. Nach Kolen und Brennan (2010) ist es per se möglich, jeden Test mit einem anderen Test zu verknüpfen. Jedoch hat die Ähnlichkeit der Tests einen Einfluss auf die möglichen Interpretationen, die hinsichtlich der Linking-Ergebnisse getroffen werden können. Insofern wurden zum einen das Linking von NEPS und TIMSS und zum anderen das Linking von NEPS und dem Ländervergleich hinsichtlich der statistischen Exaktheit und Stabilität hin untersucht und die Ergebnisse werden unter Berücksichtigung der Unterschiede in den Rahmenkonzeptionen und Testspezifikationen diskutiert.

Zusammenfassung der Ergebnisse

Das Linking der Tests erfolgte unter Verwendung der Methode des equipercentilen Linking. Zunächst wurde der NEPS-Test mit dem TIMSS-Test verlinkt. Die Ergebnisse zeigen, dass sich die Verteilungen der äquivalenten Ergebniswerte und die Ergebnisse gemessen mit dem TIMSS-Test sehr ähnlich sind, jedoch Abweichungen in den Randbereichen, vor allem zwischen den Subgruppen differenziert nach Geschlecht, bestehen. Hier hat auch das Smoothing den größten Effekt, durch welches die Abweichungen etwas ausgeglichen werden. Die äquivalenten Ergebniswerte wurden darüber hinaus den Kompetenzstufen von TIMSS zugeordnet. Mit einem $\kappa = .34$ und einer prozentualen Übereinstimmung von 43.5 % zeigt sich eine zufriedenstellende Zuordnungsgenauigkeit im Vergleich der äquivalenten Ergebniswerte und den Ergebnissen gemessen mit dem TIMSS-Test. Es zeigte sich, dass die Abweichungen in den Zuordnungen vor allem in den unteren Kompetenzstufen zu finden sind. Hinsichtlich der Überprüfung der Stabilität über Subgruppen hat sich gezeigt, dass die Gesamtgruppe inkonsistent bezüglich der Subgruppe Geschlecht ist. Beispielsweise ergaben die Analysen, dass die Jungen bei einer Transformation der Gesamtgruppe im Gegensatz zu den Mädchen etwas übervorteilt werden. Der REMSD liegt unter dem Richtwert der DTM, jedoch leicht über dem Richtwert der standardisierten DTM, wobei die Gruppenunterschiede vor allem in den Randbereichen auftreten.

Bezüglich des Linking von NEPS mit dem Ländervergleich lässt sich festhalten, dass die Verteilung der äquivalenten Ergebniswerte des NEPS mit den Ergebnissen gemessen mit dem Ländervergleichs-Test im mittleren Kompetenzbereich annähernd linear ist. In den Randbereichen zeigen sich Abweichungen vor allem bezüglich der Subgruppen. Die Zuordnung zu den Kompetenzstufen liefert mit einem $\kappa = .31$ und einer Klassifikationskorrektheit von 47.5% ebenfalls akzeptable Werte. Die Analysen hinsichtlich der Stabilität über die Subgruppen des Geschlechts zeigen, dass Unterschiede zwischen den Gruppen bestehen. Wird die DTM zur Interpretation des REMSD herangezogen, sind die Unterschiede als nicht bedeutsam einzustufen. Wird jedoch die sDTM herangezogen ergibt sich ein gegensätzliches Bild und die Werte des REMSD liegen leicht über der sDTM.

Diskussion

Bezüglich der Exaktheit und Stabilität des Linking lässt sich festhalten, dass Abweichungen in den Randbereichen, vor allem im Vergleich der Subgruppen entstehen. Stabil sind die Ergebnisse im Bereich der Kompetenzwerte 314 und 624 (NEPS/TIMSS) bzw. 288 und 610 (NEPS/Ländervergleich). In diesem Bereich liegen die größten Gruppenunterschiede bei fünf Kompetenzpunkten. Ursache für die Abweichungen in den Randbereichen können die geringen prozentualen Anzahlen der Schülerinnen und Schüler sein. Zudem gibt es Schülerinnen und Schüler, die im NEPS-Test alle Aufgaben oder gar keine Aufgabe richtig gelöst haben. Werden die oberen und unteren 5 % der Stichprobe getrimmt, sind kaum noch Abweichungen in den Randbereichen sichtbar.

Die Sensibilität bezüglich Irregularitäten in den Verteilungen der Tests ist ein Nachteil des equipercentilen Linking. Es bietet sich daher an, die Daten vorher oder auch nachher zu glätten (Dorans, Moses & Eignor, 2011; Livingston, 2004; Yin, Brennan & Kolen, 2004). Die Ergebnisse der geglätteten äquivalenten Werte zeigen, dass die Abweichungen in den Randbereichen nahezu verschwinden. Daher ist es empfehlenswert bei der Interpretation die geglätteten äquivalenten Ergebniswerte herangezogen werden.

Hinsichtlich der Zuordnungsgenauigkeit zu den Kompetenzstufen hat sich gezeigt, dass mit einer korrekten Zuordnung von 43.5% (NEPS/TIMSS) bzw. 47.5% (NEPS/Ländervergleich) eine angemessene Genauigkeit erreicht werden konnte. Nach Pietsch et al. (2009) kann wie bereits aufgezeigt eine zu erwartende Zuordnungsgenauigkeit von 42% ermittelt werden. Da die Klassifikationskorrektheit in den vorliegenden beiden Fällen darüber liegt, kann die Genauigkeit unter den gegebenen Bedingungen (Reliabilität und Korrelation zwischen den Test) als angemessen eingestuft werden. Pietsch et al. erreichten bei ihrem Linking im Vergleich eine Klassifikationskorrektheit von 33%. Dennoch bleibt mit einer Fehlzuordnung in 56.5 % bzw. 52.5 % der Fälle eine zwar erwartete aber nicht unerhebliche Ungenauigkeit in der Zuordnung zu den Kompetenzstufen. Daher lässt die Zuordnung zu den Kompetenzstufen nur eine vorsichtige Interpretation der Ergebnisse zu und die Ergebnisse können nur Tendenzen aufzeigen.

Da das Linking damit einen gewissen Grad an Ungenauigkeit unterliegt, sollte bei der Interpretation der Ergebnisse berücksichtigt werden, dass nicht die gleichen

Schlussfolgerungen wie bei einem Equating erfolgen können, also, dass die Ergebniswerte nicht als austauschbar interpretiert werden können.

Bezüglich der Exaktheit und Stabilität des Linking lässt sich damit festhalten, dass die Ergebnisse hinsichtlich zentraler Tendenzen stabil sind und sich Aussagen auf Gruppenebene treffen lassen, jedoch auf Individualebene eine Interpretation der Ergebnisse vermieden werden sollte. Diese Schlussfolgerung geht konform mit den Konklusionen aus anderen Studien, beispielsweise Brown et al. (2005) und Pietsch et al. (2009). Da das Linking nicht umfassend invariant bezüglich der Subgruppe Geschlecht ist, bietet es sich zudem an, für eine Interpretation auf Gruppenebene die jeweiligen Konkordanz-Tabellen der Subgruppen zu nutzen.

7.4 Limitationen für die Übertragung der Ergebnisse auf die NEPS Hauptuntersuchung

Neben den genannten Unterschieden, die sich auf die Exaktheit und Stabilität und damit auch auf die Interpretation des Linking auswirken können, gibt es auch Aspekte, die bei einer Übertragung der Linking-Ergebnisse aus der Linking-Studie auf die NEPS Hauptuntersuchung beachtet werden müssen.

Eine erste Einschränkung ergibt sich durch die selektive Stichprobe der Linking-Studie. Es handelt sich bei den teilnehmenden Schulen um Schulen, die an dem SINUS-Programm teilnehmen (Demuth et al., 2011). Da mit SINUS das Ziel verfolgt wird, den mathematischen und naturwissenschaftlichen Unterricht weiterzuentwickeln, kann davon ausgegangen werden, dass diese Schülerinnen und Schüler tendenziell höhere Kompetenzen in diesen Bereichen entwickelt haben, als Schülerinnen und Schüler, deren Schulen nicht am SINUS-Programm teilnehmen. Zudem war die Teilnahme an der Linking-Studie freiwillig, so dass nicht davon ausgegangen werden kann, dass die Stichprobe repräsentativ für Deutschland ist. Bei der Übertragung der Linkingergebnisse aus der Linking-Studie auf die Hauptuntersuchung ist dies zu berücksichtigen. Vor allem bei der Interpretation auf Subgruppenebene sind Verzerrungen zu erwarten, wenn die Stichproben sich hinsichtlich bestimmter Merkmale unterscheiden.

Zudem wird der NEPS-Test der fünften Jahrgangsstufe in der Linking-Studie in der vierten Jahrgangsstufe eingesetzt. Da nicht alle Aufgaben laut Expertenmeinung für die vierte Jahrgangsstufe als curricular valide einzustufen sind (vgl. Kapitel 4.1: Anlage und Ziele), besteht die Möglichkeit, dass einige Aufgaben zu schwierig für Viertklässlerinnen und Viertklässler sein könnten.

Weiterhin ist es möglich, dass in der Linking-Studie Positionseffekte aufgetreten sind, da die Reihenfolge der Bearbeitung der Testhefte von TIMSS, dem Ländervergleich und NEPS nicht variiert wurde. Dies könnte einen Trainingseffekt oder aber auch eine geringere Motivation am zweiten Testtag zur Folge haben. Dadurch könnte ein systematischer Fehler (Systematic Equating Error) in der Linkingfunktion entstanden sein (Kolen & Brennan, 2010).

Wird die NEPS-Hauptuntersuchung mit der vorliegenden Konkordanz-Tabelle in einen nationalen oder internationalen Referenzmaßstab eingeordnet ist darüber hinaus zu bedenken, dass die Schülerinnen und Schüler, die an der NEPS-Hauptuntersuchung teilgenommen haben, in etwa ein halbes Schuljahr weiter sind als die Schülerinnen und Schüler, die an TIMSS bzw. am Ländervergleich teilgenommen haben. Nach Klieme et al. (2010) kann ein Kompetenzzuwachs von 30 Punkten pro Schuljahr angenommen werden. Diesbezüglich kann demnach davon ausgegangen werden, dass die NEPS-Teilnehmerinnen und Teilnehmer der Hauptuntersuchung, die ein halbes Jahr länger die Schule besucht haben, im Durchschnitt etwa 15 Kompetenzpunkte mehr erreichen als die Schülerinnen und Schüler, die an TIMSS bzw. am Ländervergleich teilgenommen haben.

Die Übertragbarkeit der vorliegenden Linkingergebnisse auf die NEPS-Hauptuntersuchung ist darüber hinaus insofern eingeschränkt, als dass das equipercentile Linking nur äquivalente Ergebniswerte schätzt, die auch in der Linking-Studie vorkommen. Das bedeutet, dass es beispielsweise für das Linking von NEPS und TIMSS nur äquivalente Ergebniswerte zwischen 145 und 900 ($-3.55 < \theta < 4.00$) Kompetenzpunkten im NEPS gibt. In der NEPS-Hauptuntersuchung gibt es jedoch einige Schülerinnen und Schüler die unter bzw. über dieser Kompetenzspanne liegen. Für diese Schülerinnen und Schüler können daher keine äquivalenten Ergebniswerte bestimmt werden. Zudem wurde im Rahmen der Linking-Studie mit WLEs gerechnet, wohingegen in den Hauptuntersuchungen von TIMSS und dem Ländervergleich PVs geschätzt werden. Da jedoch nicht alle Variablen aus den hier

verwendeten Hintergrundmodellen vorlagen und der aktuelle NEPS-Datensatz nur WLEs beinhaltet, wurde auf die Schätzung von PVs verzichtet.

Weiterhin bleibt zu beachten, dass die Ergebnisse für unterschiedliche Populationen und über die Zeit variieren können, so dass Verzerrungen in Gruppenvergleichen und Trends entstehen könnten.

7.5 Ausblick

Äquivalenz- und Linkingstudien sind in den meisten Fällen nicht hinreichend und nicht perfekt. Es kann immer mehr untersucht und verglichen werden und die Stichprobe lässt sich immer noch etwas exakter ziehen. Daher ist solche Art von Studien immer mit Limitationen verbunden. Jedoch liefern die Ergebnisse dieser Vergleichs- und Linkingstudien trotz der genannten Einschränkungen vielerlei brauchbare Informationen. Neben der Übertragbarkeit der Ergebnisse auf die NEPS-Hauptuntersuchung – die bereits hinreichen diskutiert wurde – können die Resultate aus dieser Studie auch wie folgt genutzt werden:

(1) Die Gegenüberstellung der Studien liefert Hinweise dazu, was der eine Test im Vergleich zum anderen Test anders bzw. detaillierter misst. Die Studien ergänzen sich durch ihre leicht unterschiedlichen Konzeptionen und liefern damit auch wichtige zusätzliche Informationen. Die aufgezeigten Unterschiede zwischen den Studien können dementsprechend auch bei der Interpretation der Ergebnisse aus den Hauptuntersuchungen helfen. Beispielsweise kann der NEPS-Test nur die passiven Prozesse messen, da keine offenen Aufgabenstellungen vorhanden sind. Hingegen werden mit dem TIMSS-Test auch aktive Prozesse gemessen. Unterschiede in den Konzeptionen können Unterschiede in den Ergebnissen der Studien aufklären. Der Vergleich liefert damit Hinweise, die bei der Interpretation der Ergebnisse der Hauptuntersuchungen berücksichtigt werden können und können dadurch den Entscheidungsträger und der Öffentlichkeit dabei helfen, ein vertiefendes Verständnis aufzubauen und ein umfassenderes Bild über die Kompetenzen der Schülerinnen und Schüler zu erhalten.

(2) Weiterhin können einige Ergebnisse aus der vorliegenden Studie für eine Validierung der Testwertinterpretation u. a. nach Kane (2013) genutzt werden. Beispielsweise wurden Analysen durchgeführt, die Hinweise zur Evidenz der Testwertinterpretation des NEPS-

Mathematiktests für die fünfte Jahrgangsstufe liefern. Der NEPS-Test basiert laut Rahmenkonzeption unter anderem auf der Rahmenkonzeption des Ländervergleichs. Im Rahmen dieser Studie wurde untersucht, ob sich die NEPS-Mathematikaufgaben der Rahmenkonzeption des Ländervergleichs zuordnen lassen und ob die Gewichtung der prozentualen Verteilung mit der Verteilung im Ländervergleichs-Test übereinstimmt. Die Ergebnisse zeigen, dass der NEPS-Test tatsächlich die in den Rahmenkonzeptionen des Ländervergleichs beschriebenen Teilkompetenzen in ähnlicher Gewichtung abdeckt. Darüber hinaus wurden konvergente Zusammenhänge mit einem national und einem international anerkannten Mathematiktest aufgezeigt. Diese Zusammenhänge bestätigen das Konstrukt mathematischer Kompetenz wie es in der NEPS-Rahmenkonzeption definiert wird.

(3) Da es im deutschsprachigen Raum noch nicht viele veröffentlichte Äquivalenz- bzw. Linkingstudien gibt und auch aus dem englischsprachigen Raum nur wenige so umfangreiche Untersuchungen vorliegen, die eine vergleichende Interpretation ermöglichen, kann die vorliegende Arbeit als Referenz für weitere Studien verwendet werden. Zudem wurde exemplarisch aufgezeigt, wie die Äquivalenz von verschiedenen Studien untersucht werden kann und wie sich zwei zum Teil unterschiedliche Tests miteinander verlinken lassen.

Es gäbe eine Reihe weiterer Analysen, die hinsichtlich der Vergleichbarkeit und der Güte des Linking durchgeführt werden könnten. Im Folgenden werden exemplarisch weitere Analysemöglichkeiten aufgeführt:

(1) Die Unterschiede und die Größe der Stichproben können zu messfehlerbehafteten Linkingfunktionen führen. Daher sollten zusätzliche Maße zur Exaktheit des Linking berechnet werden. In der Literatur (u. a. Holland & Dorans, 2006; Kolen & Brennan, 2010) wird hier beispielsweise vorgeschlagen den sogenannten Standard Error of Equating (SEE) und den Standard Error of Equating Difference (SEED) zwischen zwei Linkingfunktionen zu berechnen. Eine größere Stichprobe in der Linkingstudie könnte zudem zu einer Verringerung der Messfehler führen (Kolen & Brennan, 2010).

(2) Weiterhin sollten bei der Planung künftiger Linkingstudien darauf geachtet werden, Positionseffekte auszugleichen. In der vorliegenden Studie haben alle Schülerinnen und Schüler den TIMSS-Test am ersten Testtag und den Ländervergleichs- und NEPS-Test am zweiten Testtag gemacht. Dadurch können Positionseffekte entstanden sein, die in systematischen Fehlern resultieren können.

(3) Zur Bestätigung der Dimensionalität der Tests könnten zudem konfirmatorische Faktorenanalysen berechnet werden. Bevor ein Vergleich auf dimensionaler Ebene erfolgt, sollte man überprüfen, ob sich die angenommenen Dimensionen und Subdimensionen in den Tests überhaupt bestätigen lassen.

(4) Weiterhin wäre es möglich, Strukturgleichungsmodelle zu rechnen, um die Vergleichbarkeit des mathematischen Konstrukts in den Tests näher zu untersuchen. So ließe sich beispielsweise der Zusammenhang mit dem im KFT bestimmten Intelligenzfaktor oder anderen Hintergrundmerkmalen vergleichen.

(5) Darüber hinaus könnten auch noch andere Methoden für das Linking verwendet werden, um zu überprüfen, ob eine andere Methode eine Linkingfunktion generiert, die die Datenstruktur besser abbilden und damit beispielsweise invariant bezüglich Gruppenunterschiede ist. Für das Linking von TIMSS und NEPS wurde dieser Vergleich exemplarisch vorgenommen (Nissen et al., 2015). Hierbei hat sich gezeigt, dass die verschiedenen Methoden unterschiedliche Vor- und Nachteile bieten und dass die Vorteile des equipercentilen Linking in diesem Zusammenhang leicht überwiegen. Jedoch müssten auch hierzu noch differenziertere Analysen gemacht werden, beispielsweise die Überprüfung und der Vergleich der Stabilität der Linkingmethoden. Einen Vergleich unterschiedlicher Linking Methoden kann u. a. bei Kolen und Brennan (2010) nachgelesen werden.

(6) Des Weiteren haben die Ergebnisse des Linking Ungenauigkeiten in den Randbereichen offenbart. Hier könnten in einem weiteren Schritt, die Ausreißer aus den Daten herausgenommen werden. Beispielsweise könnten die Schülerinnen und Schüler aus den Analysen ausgeschlossen werden, die mehr als drei Standardabweichungen vom Mittelwert abweichen, da es sich hierbei um Extremwerte handelt, die die Analysen stark beeinflussen können, obwohl die Vorkommenshäufigkeiten sehr gering sind. Zudem können unter Annahme eines Rasch-Modells Schülerinnen und Schüler, die alle Aufgaben richtig bzw. alle Aufgaben falsch gelöst haben, nur unzureichend exakt geschätzt werden (Rost, 2004).

(7) Es ließen sich darüber hinaus noch weitere Subgruppenunterschiede untersuchen. Hierzu können beispielsweise Subgruppenunterschiede hinsichtlich des Migrationshintergrundes, des HISEI (sozioökonomischer Status) oder der zu Hause gesprochenen Sprache analysiert werden, um die Einflüsse weiterer möglicher Invarianzen zu überprüfen.

Es lässt sich damit festhalten, dass weiterer Forschungsbedarf besteht. Auch die Übertragung einer solchen Untersuchung auf Tests aus anderen Klassenstufen oder anderer Studien wäre wünschenswert. Die Analysen der Äquivalenz wurde hierbei bereits an unterschiedlichen Stellen begonnen und auch publiziert (u. a. Böhme et al., 2014; van den Ham et al., 2016; Wagner et al., 2014), jedoch wäre auch ein Linking von diesen Studien wünschenswert, da die Ergebnisse einen umfassenderen Blick über die Messungen von Leistungen in verschiedenen Kontexten geben könnten. Dies könnte sowohl für die Wissenschaft als auch für die politischen Entscheidungen, die oftmals aus solchen Studien gezogen werden, entscheidende Informationen liefern.

Literaturverzeichnis

Ahrenholz, B. (Hrsg.). (2010). *Fachunterricht und Deutsch als Zweitsprache*. Tübingen: Narr.

American Educational Research Association, American Psychological Association & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Anger, C., Plünnecke, A. & Schmidt, J. (2010). *Bildungsrenditen in Deutschland: Einflussfaktoren, politische Optionen und ökonomische Effekte*. Köln: Institut der Deutschen Wirtschaft Köln Medien GmbH.

Angoff, W. H. (1957). The "Equating" of Non-Parallel Tests. *The Journal of Experimental Education, 25* (3), 241-247.

Angoff, W. H. (1971). Scales, Norms and Equivalent Scores. In R. L. Thorndike, W. H. Angoff, E. F. Lindquist & American Council on Education (Hrsg.), *Educational Measurement* (2. Aufl., S. 508-600). Washington: American Council on Education.

Apeltauer, E. Prof. Dr. (2003). Literalität und Spracherwerb. *Flensburger Papiere zur Mehrsprachigkeit und Kulturenvielfalt im Unterricht* (32), 5-35.

Barbu, Otilia C. & Beal, C. R. (2010). Effects of Linguistic Complexity and Math Difficulty on Word Problem Solving by English Learners. *International Journal of Education, 2* (2), 1-19.

Bauer, G. & Jacob, M. (2010). Fertilitätsentscheidungen im Partnerschaftskontext. Eine Analyse der Bedeutung der Bildungskonstellation von Paaren für die Familiengründung anhand des Mikrozensus 1996-2004. *Kölner Zeitschrift für Soziologie und Sozialpsychologie, 62* (1), 31-60.

Baumert, J. (1998). TIMSS - Mathematisch-Naturwissenschaftlicher Unterricht im internationalen Vergleich. Anlage der Studie und ausgewählte Befunde. In J. List (Hrsg.), *TIMSS: mathematisch-naturwissenschaftliche Kenntnisse deutscher Schüler auf dem Prüfstand* (S. 13-65). Köln: Deutscher Instituts-Verlag.

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

Beal, C. R., Adams, M. M., Cohen & Paul R. (2010). Reading Proficiency and Mathematics Problem Solving by High School English Language Learners. *Urban Education, 45* (1), 58-74.

Becker, T. (2005). Mündliche Vorstufen literaler Textentwicklung: vier Erzählformen im Vergleich. In H. Feilke & R. Schmidlin (Hrsg.), *Literale Textentwicklung. Untersuchungen zum Erwerb von Textkompetenz* (S. 19-42). Frankfurt a. M.: Lang.

Berendes, K., Fey, D., Linberg, T., Wenz, S. E., Roßbach, H.-G., Schneider, T. et al. (2011). Kindergarten and elementary school. In H.-P. Blossfeld, H.-G. Roßbach & J. von Maurice (Hrsg.), *Education as a Lifelong Process: The German National Educational Panel Study (NEPS)* (Zeitschrift für Erziehungswissenschaft - Sonderheft, Bd. 14, S. 203-216). Wiesbaden: VS-Verlag.

Blossfeld, H.-P., Doll, J. & Schneider, T. (2009). Die Nationale Bildungspanelstudie (NEPS). In W. Böttcher, J. N. Dicke & H. Ziegler (Hrsg.), *Evidenzbasierte Bildung: Wirkungsevaluation in Bildungspolitik und pädagogischer Praxis* (S. 59-68). Münster: Waxmann.

Blossfeld, H.-P., Maurice, J. von & Schneider, T. (2011). The National Educational Panel Study: need, main features, and research potential. In H.-P. Blossfeld, H.-G. Roßbach & J. von Maurice (Hrsg.), *Education as a Lifelong Process: The German National Educational Panel Study (NEPS)* (Zeitschrift für Erziehungswissenschaft - Sonderheft, Bd. 14, S. 5-17). Wiesbaden: VS-Verlag.

Blossfeld, H.-P., Roßbach, H.-G. & Maurice, J. von (Hrsg.). (2011). *Education as a Lifelong Process: The German National Educational Panel Study (NEPS)* (Zeitschrift für Erziehungswissenschaft - Sonderheft, Bd. 14). Wiesbaden: VS-Verlag.

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

Blum, W., Neubrand, M., Ehmke, T., Senkbeil, M., Jordan, A., Ulfig, F. et al. (2004). Mathematische Kompetenz. In PISA-Konsortium Deutschland (Hrsg.), *PISA 2003. Der Bildungsstand der Jugendlichen in Deutschland - Ergebnisse des zweiten internationalen Vergleichs* (S. 47-92). Münster: Waxmann.

Böhme, K., Richter, D., Stanat, P., Pant, H. A. & Köller, O. (2012). Die länderübergreifenden Bildungsstandards in Deutschland. In P. Stanat, H. A. Pant, K. Böhme & D. Richter (Hrsg.), *Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Deutsch und Mathematik. Ergebnisse des IQB-Ländervergleichs 2011* (S. 11-18). Münster: Waxmann.

Böhme, K., Richter, D., Weirich, S., Haag, N., Wendt, H., Bos, W., Pant, H.A. & Stanat, P. (2014). Messen wir dasselbe? Zur Vergleichbarkeit des IQB-Ländervergleichs 2011 mit den internationalen Studien IGLU und TIMSS 2011. *Unterrichtswissenschaft*, *42* (4), 342–365.

Bonsen, M., Lintorf, K., Bos, W. & Frey, K. A. (2008). TIMSS 2007 Grundschule. Eine Einführung in die Studie. In W. Bos, M. Bonsen, J. Baumert, M. Prenzel, C. Selter & G. Walther (Hrsg.), *TIMSS 2007. Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* (S. 19-48). Münster: Waxmann.

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

Brennan, R. L. (2004). *Manual for LEGS. Version 2.0*. Iowa City, IA: Center for Advanced Studies in Measurement and Assessment (CASMA).

Brown, G., Micklewright, J., Schnepf, S. V. & Waldmann, R. (2005). Cross- National Surverys of Learning Achievement: How Robust are the Findings? *IZA Discussion Paper* (1652), 1-34.

Bundesministerium für Bildung und Forschung. *Rahmenprogramm Empirische Bildungsforschung.* Zugriff am 05.03.2014. Verfügbar unter http://www.dlr.de/pt/desktopdefault.aspx/tabid-8612/14832_read-36922/

Cartwright, F. (2012). *Linking the British Columbia English Examination to the OECD Combined Reading Scale. Prepared for the British Columbia Ministry of Education.* Zugriff am

05.03.2014. Verfügbar unter http://www.bced.gov.bc.ca/assessment/linking_bc_eng_pisa.pdf

Cartwright, F., Lalancette, D., Mussio, J. & Xing, D. (2003). *Linking provincial student assessments with national and international assessments* (Education, skills and learning, research papers). Ottawa: British Columbia Ministry of Education. Zugriff am 05.03.2014. Verfügbar unter http://www.publications.gc.ca/Collection/Statcan/81-595-MIE/81-595-MIE2003005.pdf

Demuth, R., Walther, G. & Prenzel, M. (2011). *Unterricht entwickeln mit SINUS. 10 Module für den Mathematik- und Sachunterricht in der Grundschule*. Seelze: Klett-Kallmeyer.

Dorans, N. J. (2004). Equating, Concordance, and Expectation. *Applied Psychological Measurement, 28,* 227-246. Zugriff am 05.03.2014. Verfügbar unter http://apm.sagepub.com/content/28/4/227.full.pdf+html

Dorans, N. J. & Holland, P. W. (2000). Population Invariance and the Equatability of Tests: Basic Theory and The Linear Case. *Journal of Educational Measurement, 37* (4), 281-306.

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

Dorans, N. J., Moses, T. P. & Eignor, D. R. (2011). Equating Test Scores: Toward Best Practices. In A. A. von Davier (Hrsg.), *Statistical models for test equating, scaling, and linking* (Statistics for social and behavioral sciences, S. 21-42). New York: Springer.

Dorans, N. J. & Walker, M. E. (2007). Sizing Up Linkages. In N. J. Dorans, M. Pommerich & P. W. Holland (Hrsg.), *Linking and Aligning Scores and Scales* (Statistics for social and behavioral sciences, S. 179-198). New York: Springer.

Duchhardt, C. & Gerdes, A. (2012a). *NEPS Technical Report for Mathematics - Scaling Results of Starting Cohort 3 in Fifth Grade* (NEPS Working Paper Nr. 19). Bamberg: Otto-Friedrich-Universität, Nationales Bildungspanel. Verfügbar unter https://www.neps-data.de/Portals/0/Working%20Papers/WP_XIX.pdf

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

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

Edmondson, W. J. & House, J. (1993). *Einführung in die Sprachlehrforschung* (UTB, Bd. 1697, 1. Aufl.). Tübingen: Francke.

Ehmke, T., Duchhardt, C., Geiser, H., Grüßing, M., Heinze, A. & Marschick, F. (2009). Kompetenzentwicklung über die Lebensspanne. Erhebung von mathematischer Kompetenz im Nationalen Bildungspanel. In A. Heinze (Hrsg.), *Mathematiklernen vom Kindergarten bis zum Studium. Kontinuität und Kohärenz als Herausforderung für den Mathematikunterricht* (S. 313-327). Münster: Waxmann.

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

Einsiedler, W. (2003). Unterricht in der Grundschule. In K. S. Cortina, J. Baumert, A. Leschinsky, K. U. Mayer & L. Trommer (Hrsg.), *Das Bildungswesen in der Bundesrepublik Deutschland.*

Strukturen und Entwicklungen im Überblick (rororo Sachbuch, Bd. 61122, Vollst. überarb. und erw. Neuausg., S. 285-341). Reinbek bei Hamburg: Rowohlt Taschenbuch Verlag.

Fend, H. (2009). *Neue Theorie der Schule. Einführung in das Verstehen von Bildungssystemen* (2. Aufl.). Wiesbaden: VS, Verl. für Sozialwiss.

Feuer, M. J., Holland, P. W., Green, B. F., Bertenthal, M. W. & Hemphill, F. C. (1999). *Uncommon Measures. Equivalence and Linkage Among Educational Tests*. Washington: National Academy Press.

Fischer, C., Rieck, K. & Dedekind, B. (2009). SINUS-Transfer Grundschule. Lehrkräfte verändern ihren Mathematikunterricht und ihren naturwissenschaftlichen Sachunterricht an Grundschulen- (wie) geht das? *Das Journal für den frühen mathematischen und naturwissenschaftlichen Unterricht, 1* (2), 44-49.

Fischer, C., Rieck, K., Döring, B. & Köller, O. (Hrsg.). (2014). *Zusammenwirken- zusammen wirken. Unterrichtsentwicklung anstoßen, umsetzen und sichern*. Seelze: Kallmeyer/ Klett.

Flanagan, J. C. (1951). Units, Scores, and Norms. In E. F. Lindquist (Hrsg.), *Educational Measurement* (S. 695-763). Washington, D.C: American Council on Education.

Foy, P., Brossmann, B. & Galia (2012). Scaling TIMSS and PIRLS 2011 achievement data. In M. O. Martin & I. V. Mullis (Hrsg.), *TIMSS &PIRLS- Methods and Procedures in TIMSS and PIRLS 2011.* Chestnut Hill, Mass: International Study Center, Lynch School of Education, 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. V. Mullis (Hrsg.), *TIMSS &PIRLS- Methods and Procedures in TIMSS and PIRLS 2011* (S. 1-27). Chestnut Hill, Mass: International Study Center, Lynch School of Education, Boston College.

Granzer, D. (2009). Von Bildungsstandards zu ihrer Überprüfung: Grundlagen der Item- und Testentwicklung. In D. Granzer, O. Köller, A. Bremerich-Vos, M. van den Heuvel-Panhuizen, K. Reiss & G. Walther (Hrsg.), *Bildungsstandards Deutsch und Mathematik. Leistungsmessung in der Grundschule* (1. Aufl., S. 21-30). Weinheim: Beltz.

Grimm, H. & Weinert, S. (1998). Kapitel 15. Sprachentwicklung. In R. Oerter & L. Montada (Hrsg.), *Entwicklungspsychologie* (Lehrbuch, S. 517-550). Weinheim [u.a.]: Beltz PVU.

Grønmo, L. S. & Olsen, R. V. (2007). TIMSS versus PISA: The case of pure and applied mathematics. In The International Association for the Evaluation of Educational Achievement (Hrsg.), *The second IEA International research conference. Proceedings of the IRC-2006* (S. 201-214). The International Association for the Evaluation of Educational Achievement.

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.

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

Hartig, J. & Frey, A. (2012). Validität des Tests zur Überprüfung des Erreichens der Bildungsstandards in Mathematik. Zusammenhänge mit den bei PISA gemessenen Kompetenzen und Varianz zwischen Schulen und Schulformen. *Diagnostica, 58* (1), 3-14.

He, J. & van de Vijver, F. J. R. (2012). Bias and Equivalence in Cross-Cultural Research. *Online Readings in Psychology and Culture, 2* (2), 1-19.

Heinze, A., Herwartz-Emden, L. & Reiss, K. (2007). Mathematikkenntnisse und sprachliche Kompetenz bei Kindern mit Migrationshintergrund zu Beginn der Grundschulzeit. *Zeitschrift für Pädagogik, 53* (4), 562-581.

Heller, K. A. & Perleth, C. (2000). *Kognitiver Fähigkeitstest für 4. bis 12. Klassen, Revision*. Göttingen: Beltz Test.

Holland, P. W. (2007). A Framework and History for Score Linking. In N. J. Dorans, M. Pommerich & P. W. Holland (Hrsg.), *Linking and Aligning Scores and Scales* (Statistics for social and behavioral sciences, S. 5-30). New York: Springer.

Holland, P. W. & Dorans, N. J. (2006). Linking and Equating. In R. L. Brennan (Hrsg.), *Educational Measurement* (American Council on Education/ Oryx Press Series on Higher Education, 4. Aufl., S. 187-220). Westport, CT: Praeger Publishers.

Hutchison, D. & Schagen, I. (2006). *Comparisons Between PISA and TIMSS – Are We the Man with Two Watches?* National Foundation for Educational Research.

Institut zur Qualitätsentwicklung im Bildungswesen. (2006). *Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring* (Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland, Hrsg.), München. Verfügbar unter http://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2006/2006_08_01-Gesamtstrategie-Bildungsmonitoring.pdf

IQB - Institut zur Qualitätsentwicklung im Bildungswesen. (2007). *Perspektiven und Visionen. Tätigkeitsbericht 2005/06*. Berlin: Humboldt-Universität.

Johnson, E., Cohen, J., Chen, W-H., Jiang, T. & Zhang, Y. (2003). *2000 NAEP – 1999 TIMSS Linking Report*. Working Paper of the American Institutes for Research. The Educational Testing Service.

Jordan, A., Ross, N., Krauss, S., Baumert, J., Blum, W., Neubrand, M. et al. (2006). *Klassifikationsschema für Mathematikaufgaben. Dokumentation der Aufgabenkategorisierung im COACTIV-Projekt* (Materialien aus der Bildungsforschung, Bd. 81). Berlin: Max-Planck-Inst. für Bildungsforschung.

Junk-Deppenmeier, A. & Schäfer, J. (2010). Lesekompetenz als Voraussetzung für das Lernen im Fachunterricht. In B. Ahrenholz (Hrsg.), *Fachunterricht und Deutsch als Zweitsprache* (S. 69-86). Tübingen: Narr.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50,* 1-73.

Klann-Delius, G. (2008). *Spracherwerb* (Sammlung Metzler, Bd. 321, 2. aktualisierte und erweiterte Aufl). Weimar: Verlag J.B. Metzler.

Klieme, E., Avenarius, H., Blum, W., Döbrich, P., Gruber, H., Prenzel, M. et al. (2003). *Zur Entwicklung nationaler Bildungsstandards. Eine Expertise.* (Bundesministerium für Bildung und Forschung, Hrsg.) (Zur Entwicklung nationaler Bildungsstandards Bildungsforschung Band 1), Bonn, Berlin. Zugriff am 24.03.2014. Verfügbar unter http://www.bmbf.de/pub/zur_entwicklung_nationaler_bildungsstandards.pdf

Klieme, Eckhard; Artelt, Cordula; Hartig, Johannes; Jude, Nina; Köller, Olaf; Prenzel, Manfred et al. (Hg.) (2010*): PISA 2009. Bilanz nach einem Jahrzehnt.* Münster: Waxmann.

KMK. (2005). *Beschlüsse der Kultusministerkonferenz: Bildungsstandards im Fach Mathematik für den Primarbereich. Beschluss vom 15.10.2004*. Neuwied: Luchterhand.

KMK & IQB - Institut zur Qualitätsentwicklung im Bildungswesen. (2012, 15. Februar). *Kompetenzstufenmodell zu den Bildungsstandards für den Hauptschulabschluss und den Mittleren Schulabschluss im Fach Mathematik*. Verfügbar unter https://www.iqb.hu-berlin.de/bista/ksm/Kompetenzstufenm.pdf.

Kolen, M. J. (2004). Linking Assessments: Concept and History. *Applied Psychological Measurement, 28* (4), 219-226.

Kolen, M. J. & Brennan, R. L. (2010). *Test Equating, Scaling, and Linking. Methods and Practices* (2. Auflage). New York: Springer Science + Business Media, Inc.

Köller, O. (2008). Bildungsstandards - Verfahren und Kriterien bei der Entwicklung von Messinstrumenten. *Zeitschrift für Pädagogik, 54* (2), 163-173. Verfügbar unter http://nbn-resolving.de/urn:nbn:de:0111-opus-43418

Leuze, K. (2008). *Bildungswege besser verstehen: das Nationale Bildungspanel* (Nr. 02). Verfügbar unter http://www.wzb.eu/sites/default/files/publikationen/wzbrief/wzbrief-bildung200802_leuze.pdf

Lindquist, E. F. (1964). Equating Scores on Non-Parallel Tests. *Journal of Educational Measurement, 1* (1), 5-9.

Linke, A., Nussbaumer, M. & Portmann-Tselikas, P. R. (2001). *Studienbuch Linguistik. Ergänzt um ein Kapitel "Phonetik und Phonologie" von Urs Willi* (4. Auflage). Tübingen: Niemeyer.

Linn, R. L. (1993). Linking Results of Distinct Assessments. *Applied Measurement in Education, 6* (1), 83-102.

Livingston, S. A. (2004). *Equating Test Scores. (without IRT)* (ETS Educational Testing Service, Hrsg.), Princeton, NJ.

Lord, F. M. & Wingersky, M. S. (1984). Comparison of IRT True-Score and Equipercentile Observed-Score "Equatings". *Applied Psychological Measurement, 8* (4), 453-461.

Maehler, D. & Schmidt-Denter, U. (2013). *Migrationsforschung in Deutschland: Leitfaden und Messinstrumente zur Erfassung psychologischer Konstrukte*. Wiesbaden: Springer.

Marco, G. L., Abdel-Fattah, A. A. & Baron, P. A. (1992). *Methods Used to Establish Score Comparability on the Enhanced ACT Assessment and the SAT* (College Board Publications, Hrsg.), New York. Verfügbar unter http://research.collegeboard.org/sites/default/files/publications/2012/7/researchreport-1992-3-methods-establish-score-comparability-enhanced-act-sat.pdf

Martin, M. O. & Mullis, I. V. (2012a). *Methods and procedures in TIMSS and PIRLS 2011*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center Lynch School of Education Boston College.

Martin, M. O. & Mullis, I. V. (Hrsg.). (2012b). *TIMSS &PIRLS- Methods and Procedures in TIMSS and PIRLS 2011*. Chestnut Hill, Mass: International Study Center, Lynch School of Education, Boston College.

Merkens, H. (2006). Bildungsforschung und Erziehungswissenschaft. In H. Merkens (Hrsg.), *Erziehungswissenschaft und Bildungsforschung* (1. Aufl., S. 9-20). Wiesbaden: VS Verlag für Sozialwissenschaften. Verfügbar unter http://download.springer.com/static/pdf/554/bok%253A978-3-531-90089-6.pdf?auth66=1396436611_28594159df98f037a96840b9bad1283b&ext=.pdf

Mislevy, R. J. (1992). *Linking Educational Assessments: Concepts, Issues, Methods, and Prospects* (ETS Educational Testing Service, Hrsg.), Princeton, NJ.

Mullis, I. V., Martin, M. O., Foy, P. & Arora, A. (2012). *TIMSS 2011 International Results in Mathematics*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center Lynch School of Education Boston College.

Mullis, I. V., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y. & Preuschoff, C. (2009). *TIMSS 2011. Assessment Framework*. Amsterdam: International Association for the Evaluation of Educational Achievement.

Muraki, E. & Bock, R. D. (1997). *Parscale: IRT item analysis and test scoring for rating scale data* (Scientific Software International, Hrsg.). Skokie, Illinois: Pennsylvania State University.

Muraki, E., Hobo, C. M. & Lee, Y.-W. (2000). Equating and Linking of Performance Assessments. *Applied Psychological Measurement, 24* (4), 325-337.

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

National Council of Teachers of Mathematics. (2005). *Principles and standards for school mathematics* (4. print). Reston, Va: National Council of Teachers of Mathematics.

Nationales Bildungspanel. (2011). *Startkohorte 3 Haupterhebung 2010/2011 (A28) Schüler/innen , Klasse 5 in Regelschulen. Informationen zum Kompetenztest*. Bamberg: Universität Bamberg.

Neidorf, T. S., 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. Technical Report*. Washington, D.C.: U.S. Department of Education. Zugriff am 30.07.2013. Verfügbar unter http://files.eric.ed.gov/fulltext/ED491692.pdf

Nohara, D. (2001). *A Comparison of the National Assessment of Educational Progress (NAEP), the Third International Mathematics and Science Study Repeat (TIMSS-R), and the Programme for International Student Assessment (PISA)*. Washington, D.C.: U.S. Department of Education; Office of Educational Research and Improvement; National Center for Education Statistics.

Nissen, A., Ehmke, T., Köller, O. & Duchhardt, C. (2015). Comparing apples with oranges? An approach to link TIMSS and the National Educational Panel Study in Germany via equipercentile and IRT methods. *Studies in Educational Evaluation, 47*(1), 58–67.

OECD. (2004). *The PISA 2003 Assessment Framework. Mathematics, Reading, Science and Problem Solving Knowledge and Skills* (SourceOECD). Paris: OECD Publishing. Verfügbar unter http://www.sourceoecd.org/926410173X

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, M. van den Heuvel-Panhuizen, K. Reiss & G. Walther (Hrsg.), *Bildungsstandards Deutsch und Mathematik. Leistungsmessung in der Grundschule* (S. 393-428). Weinheim: Beltz.

Pohl, S. & Carstensen, C. H. (2012). *NEPS Technical Report – Scaling the Data of the Competence Tests (NEPS Working Paper No. 14)* (NEPS Working Paper Nr. 14). Bamberg: Otto-Friedrich-Universität, Nationales Bildungspanel. Verfügbar unter https://www.neps-data.de/Portals/0/Working%20Papers/WP_XIV.pdf

Pöhlmann, C., Neumann, D., Tesch, B. & Köller, O. (2010). Die Bildungsstandards im allgemeinbildenden Schulsystem. In O. Köller, M. Knigge & B. Tesch (Hrsg.), *Sprachliche Kompetenzen im Ländervergleich* (S. 13-18). Münster, New York: Waxmann.

Pommerich, M. (2007). Concordance: The Good, the Bad, and the Ugly. In N. J. Dorans, M. Pommerich & P. W. Holland (Hrsg.), *Linking and Aligning Scores and Scales* (Statistics for social and behavioral sciences, S. 199-216). New York: Springer.

Pommerich, M. & Dorans, N. J. (2004). Linking Scores Via Concordance: Introduction to the Special Issue. *Applied Psychological Measurement, 28* (4), 216-218.

Pommerich, M., Hanson, B. A., Harris, D. J. & Sconing, J. A. (2004). Issues in Conducting Linkages between Distinct Tests. *Applied Psychological Measurement, 28* (4), 247-273.

Prenzel, M. SINUS-Transfer Grundschule. Weiterentwicklung des mathematischen und naturwissenschaftlichen Unterrichts an Grundschulen. Gutachten des Leibnitz-Instituts für die Pädagogik der Naturwissenschaften (IPN) Kiel. *Materialien zu Bildungsplanung und Forschungsförderung* (112). Verfügbar unter http://www.blk-bonn.de/papers/heft112.pdf

Prenzel, M., Häußler, P., Rost, J. & Senkbeil, M. (2002). Der PISA-Naturwissenschaftstest: Lassen sich die Aufgabenschwierigkeiten vorhersagen? *Unterrichtswissenschaft, 30* (2), 120-135.

Rabe, T. (2012). Vorwort des Präsidenten der Kultusministerkonferenz. In P. Stanat, H. A. Pant, K. Böhme & D. Richter (Hrsg.), *Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Deutsch und Mathematik. Ergebnisse des IQB-Ländervergleichs 2011* (S. 9-10). Münster: Waxmann.

Reusser, K. (1997). Erwerb mathematischer Kompetenzen. In F. Weinert & A. Helmke (Hrsg.), *Entwicklung im Grundschulalter* (S. 141-155). Weinheim: Beltz/Psychologie Verlags Union.

Richter, D., Engelbert, M., Böhme, K., Haag, N., Hannighofer, J., Reimers, H. et al. (2012). Anlage und Durchführung des Ländervergleichs. In P. Stanat, H. A. Pant, K. Böhme & D. Richter (Hrsg.), *Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Deutsch und Mathematik. Ergebnisse des IQB-Ländervergleichs 2011* (S. 85-102). Münster: Waxmann.

Robitzsch, A. (2009). Methodische Herausforderungen bei der Kalibrierung von Leistungstests. In D. Granzer, O. Köller, A. Bremerich-Vos, M. van den Heuvel-Panhuizen, K. Reiss & G. Walther (Hrsg.), *Bildungsstandards Deutsch und Mathematik. Leistungsmessung in der Grundschule* (S. 42-106). Weinheim: Beltz.

Roppelt, A. & Reiss, K. (2012). Beschreibung der im Fach Mathematik untersuchten Kompetenzen. In P. Stanat, H. A. Pant, K. Böhme & D. Richter (Hrsg.), *Kompetenzen von*

Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Deutsch und Mathematik. Ergebnisse des IQB-Ländervergleichs 2011 (S. 34-48). Münster: Waxmann.

Rosebrock, C. & Nix, D. (2011). *Grundlagen der Lesedidaktik. Und der systematischen schulischen Leseförderung* (4. Auflage). Baltmannsweiler: Schneider Hohengehren.

Rost, J. (1996). *Lehrbuch Testtheorie, Testkonstruktion* (1. Auflage). Bern: Huber.

Rost, J. (2004). *Lehrbuch Testtheorie - Testkonstruktion* (2. vollst. überarb. und erw. Auflage). Bern: Huber.

Ryan, J. & Brockmann, F. (2009). *A Practioner's Introduction to Equating with Primers on Classical Test Theory and Item Response Theory* (Council of Chief State School Officers, Hrsg.). Washington, DC: State Collaborative on Assessment and Student Standards - Technical Issues in Large-Scale Assessment. Verfügbar unter http://files.eric.ed.gov/fulltext/ED544690.pdf

Sälzer, C., Reiss, K., Schiepe-Tiska, A., Prenzel, M. & Heinze, A. (2013). Zwischen Grundlagenwissen und Anwendungsbezug: Mathematische Kompetenz im internationalen Vergleich. In M. Prenzel, C. Sälzer, E. Klieme & O. Köller (Hrsg.), *PISA 2012. Fortschritte und Herausforderungen in Deutschland* (S. 47-97). Münster: Waxmann.

Schnotz, W. (2005). An integrated model of text and picture comprehension. In R. E. Mayer (Hrsg.), *The Cambridge handbook of multimedia learning* (S. 49-69). Cambridge, U.K.: Cambridge University Press.

Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland. (2005). *Bildungsstandards der Kultusministerkonferenz. Erläuterungen zur Konzeption und Entwicklung*. München: Luchterhand.

Selter, C., Walther, G., Wessel, J. & Wendt, H. (2012). Mathematische Kompetenz im internationalen Vergleich: Testkonzeption und Ergebnisse. In W. Bos, H. Wendt, O. Köller & C. Selter (Hrsg.), *TIMSS 2011. Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* (S. 69-122). Münster: Waxmann.

Shaftel, J., Belton-Kocher, E., Glasnapp, D. & Poggio, J. (2006). The Impact of Language Characteristics in Mathematics Test Items on the Performance of Englisch Language Learners and Students with Disabilities. *Education Assessment, 11* (2), 105-126.

Sheehan, K. M. (1985). *M-Group: Estimation of group effects in multivaritate models [Computer program]* (Educational Testing Service, Hrsg.). N.J: Boston College.

Skopek, J., Pink, S. & Bela, D. (2012). *Data Manual. Starting Cohort 3- From Lower to Upper Secondary School. NEPS SC3.1.0.0. NEPS Research Data Paper*. Bamberg: Universität Bamberg.

Stanat, P., Pant, H. A., Böhme, K. & Richter, D. (Hrsg.). (2012). *Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Deutsch und Mathematik. Ergebnisse des IQB-Ländervergleichs 2011*. Münster: Waxmann.

Stone, C. A., Weissman, A. & Lane, S. (2005). The Consistency of Student Proficiency Classifications under Competing IRT Models. *Educational Assessment, 10* (2), 125-146.

van de Vijver, F. J. R. (1998). Towards a Theory of Bias and Equivalence. In J. A. Harkness (Hrsg.), *Cross-cultural survey equivalence* (Bd. 3, S. 41-65). ZUMA-Nachrichten Spezial.

Mannheim: Zentrum für Umfragen, Methoden und Analysen (ZUMA). Verfügbar unter http://arno.uvt.nl/show.cgi?fid=46374

van den Ham, A.-K., Ehmke, T., Nissen, A., Roppelt, A. (2016). Assessments verbinden, Interpretationen erweitern? Lassen sich die mathematischen Kompetenzskalen im Nationalen Bildungspanel und im IQB-Ländervergleich 2012 verbinden? *Zeitschrift für Erziehungswissenschaft.* 1-23.

von Davier, A. A. (2011). A Statistical Perspective on Equating Test Scores. In A. A. von Davier (Hrsg.), *Statistical models for test equating, scaling, and linking* (Statistics for social and behavioral sciences, S. 1-17). New York: Springer.

von Davier, A. A., Carstensen, C. H. & von Davier, M. (2008). Linking Competencies in Horizontal, Vertical, and Longitudinal Settings and Measuring Growth. In J. Hartig, E. Klieme & D. Leutner (Hrsg.), *Assessment of Competencies in Educational Contexts* (S. 121-150). Toronto: Hogrefe & Huber Publishers.

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.

Warm, T. (1989). Weighted Likelihood Estimation of Ability in Item Response Theory. *Psychometrika, 54* (3), 427-450.

Weinert, F. E. (2002). Vergleichende Leistungsmessung in Schulen - eine umstrittene Selbstverständlichkeit. In F. E. Weinert (Hrsg.), *Leistungsmessungen in Schulen* (2. Auflage, S. 17-31). Weinheim: Beltz-Verl.

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

Weirich, S., Haag, N. & Roppelt, A. (2012). Testdesign und Auswertung des Ländervergleichs: technische Grundlagen. In P. Stanat, H. A. Pant, K. Böhme & D. Richter (Hrsg.), *Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Deutsch und Mathematik. Ergebnisse des IQB-Ländervergleichs 2011* (S. 277-290). Münster: Waxmann.

Wendt, H., Tarelli, I., Bos, W., Frey, K. & Vennemann, M. (2012). Ziele, Anlage und Durchführung der Trends in International Mathematics and Science Study (TIMSS 2011). In W. Bos, H. Wendt, O. Köller & C. Selter (Hrsg.), *TIMSS 2011. Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* (S. 27-68). Münster: Waxmann.

Westhoff, G. (1997). *Fertigkeit Lesen* (Fernstudienprojekt zur Fort- und Weiterbildung im Bereich Germanistik und Deutsch als Fremdsprache, Fernstudieneinheit 17). Kassel: Langenscheidt.

Winkelmann, H. & Robitzsch, A. (2009). Modelle mathematischer Kompetenzen: Empirische Befunde zur Dimensionalität. In D. Granzer, O. Köller, A. Bremerich-Vos, M. van den Heuvel-Panhuizen, K. Reiss & G. Walther (Hrsg.), *Bildungsstandards Deutsch und Mathematik. Leistungsmessung in der Grundschule* (1. Aufl., S. 169). Weinheim: Beltz.

Winter, H. (1995). Mathematikunterricht und Allgemeinbildung. *Mitteilungen der Gesellschaft für Didaktik der Mathematik, 61,* 37-46.

Wirtz, M. & Caspar, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität. Methoden zur Bestimmung und Verbesserung der Zuverlässigkeit von Einschätzungen mittels Kategoriensystemen und Ratingskalen*. Göttingen: Hogrefe.

Wolf, M. K. & Leon, S. (2009). An Investigation of the Language Demands in Content Assessments for English Language Learners. *Education Assessment, 14* (3-4), 139-159.

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

Wu, M., Adams, R. J. & Wilson, M. (1998). *ACER ConQuest: generalised item response modelling software* (ACER Press, Hrsg.), Melbourne.

Wu, Margaret L.; Adams, R. J.; Wilson; Haldane, S. A. (2007): ACER ConQuest (Version 2.0). In: *Camberwell, Victoria, Australia: ACER Press, Australian Council for Educational Research.*

Yin, P., Brennan, R. L. & Kolen, M. J. (2004). Concordance Between ACT and ITED Scores From Different Populations. *Applied Psychological Measurement, 28* (4), 274-289.

Zinn, S. (2013). *Replication weights for the cohort samples of students in Grade 5 and 9 in the National Educational Panel Study (NEPS Working Paper No. 27* (NEPS Working Paper Nr. 27). Bamberg: Otto-Friedrich-Universität, Nationales Bildungspanel. Verfügbar unter https://www.neps-data.de/Portals/0/Working%20Papers/WP_XXVII.pdf