LEUPHANA
UNIVERSITÄT LÜNEBURG

Ein Validitätsargument für den Mathematiktest der National Educational Panel Study für die neunte Klassenstufe

Von der Fakultät Bildung
der Leuphana Universität Lüneburg zur Erlangung des Grades
Doktorin der Philosophie
- Dr. phil. -

genehmigte Dissertation von

Ann-Katrin van den Ham
geboren am 26.04.1987 in Winsen (Luhe)

Eingereicht am: 30. November 2015

Mündliche Verteidigung (Disputation) am: 28. Juni 2016

Erstbetreuer und -gutachter:	Prof. Dr. Timo Ehmke
Zweitgutachter:	Prof. Dr. Alexander Freund
Drittgutachter:	Prof. Dr. Aiso Heinze

Danksagung

Ich möchte mich an dieser Stelle bei vielen Personen bedanken, die zum Gelingen dieser Dissertation beigetragen haben.
An erster Stelle gilt mein Dank meinem Doktorvater Prof. Dr. Timo Ehmke, für die intensive und vielseitige Unterstützung in allen Phasen dieser Dissertation und für die Möglichkeit, in diesem Projekt mitarbeiten zu dürfen. Prof. Dr. Aiso Heinze danke ich für das Interesse an meiner Arbeit sowie für die Unterstützung und die pragmatischen Anregungen. Außerdem Danke ich Prof. Dr. Alexander Freund für die Bereitschaft diese Arbeit zu begutachten.
Danken möchte ich außerdem meinen Kolleginnen und Kollegen für die unermüdlichen fachlichen Diskussionen, Ratschläge und Anregungen, aber auch für die nichtwissenschaftlichen und motivierenden Gespräche. Besonders danken möchte ich dabei Annika Nissen und Svenja Hammer für die wertvollen inhaltlichen Diskussionen und die moralische Unterstützung, sowie Marcus Pietsch für die methodischen Anregungen. Der Hilfskraft Maren Preuss möchte ich vor allem für ihr großes Engagement, die tolle Arbeit im Projekt sowie die umfangreiche LaTeX-Recherche danken.
Während dieser mehrjährigen Phase habe ich große Unterstützung durch meine Familie erfahren, der ich daher besonders danken möchte. Mein Mann Wieger van den Ham hat jede Hoch- und Tiefphase miterlebt und mir immer unterstützend zur Seite gestanden. Meiner Schwester möchte ich für die inhaltlichen Gespräche und für ihre Hilfe in der letzten Korrekturphase danken. Meinen Eltern danke ich für ihre Unterstützung, die Bereitstellung sonniger Arbeitsplätze sowie die leibliche Versorgung und meinem Bruder für die technische Unterstützung. Meinen Schwiegereltern danke ich für die vielen lieben Worte während der Erarbeitung dieser Dissertation.
Darüber hinaus geht ein großes Dankeschön an alle Schülerinnen und Schüler, die an diesem Projekt teilgenommen haben und ohne die diese Studie gar nicht möglich gewesen wäre.

Zusammenfassung

Das Nationale Bildungspanel (NEPS) untersucht wie sich Kompetenzen, unter anderem die mathematische Kompetenz, über die Lebensspanne entwickeln, wie diese die Bildungskarriere beeinflussen und inwiefern die Kompetenzentwicklung von Lerngelegenheiten abhängig ist. Erhoben werden die Daten mit Hilfe eines Multi-Kohorten-Sequenz-Designs, wobei in allen Bildungsetappen verschiedene Instrumente zum Einsatz kommen (Blossfeld & von Maurice, 2011). Um diese Daten sinnvoll nutzen zu können, müssen sie valide Testwertinterpretationen zulassen. Diese Arbeit leistet einen Beitrag zur Validierung der National Educational Panel Study (NEPS)-Testwertinterpretationen, indem der Mathematiktest für die neunte Klassenstufe (K9) umfassend analysiert wurde. Dabei wurde geprüft, ob die Unterschiede in der beobachteten Testleistung die Unterschiede in der mathematischen Fähigkeit wiedergeben, wie diese im NEPS-Rahmenkonzept definiert sind. Für die Validierung wurde eine Argumentationskette für den NEPS-K9-Mathematiktest, angelehnt an den Argument Based Approach von Kane (2013), entwickelt. Für die Auswertung wurde der *scientific use file* der NEPS-Haupterhebung 2010 verwendet. Die Stichprobe dieser Studie bestand aus 15239 Schülerinnen und Schülern der neunten Jahrgangsstufe. Für die Untersuchung der Zusammenhänge des NEPS-Mathematiktests mit anderen Messungen mathematischer und nicht mathematischer Kompetenz wurde zusätzlich eine Validierungsstudie durchgeführt. Zur Erfassung der mathematischen und naturwissenschaftlichen Kompetenz wurden in dieser Studie neben den NEPS-Testinstrumenten Aufgaben aus dem Programme for International Student Assessment (PISA) und dem IQB-Ländervergleich (LV) eingesetzt. Zusätzlich wurde ein kognitiver Fähigkeitstest verwendet. Insgesamt wurden 80 Schulen mit 1965 Schülerinnen und Schülern der neunten Klassenstufe für die Teilnahme an der Studie ausgewählt. Zur Vorbereitung des Validitätsargumentes wurden Literaturreviews bereits veröffentlichter Analysen und ein Expertenreview zu den Merkmalen der NEPS-Mathematikaufgaben durchgeführt. Des Weiteren wurden die Kompetenzdaten mit Hilfe von Modellen der Item Response Theorie skaliert und dimensional analysiert. Außerdem wurden korrelati-

ve Zusammenhänge mit Kriterien mathematischer Kompetenz berechnet. Die Ergebnisse liefern viele Hinweise bezüglich der Validität der Testwertinterpretationen, jedoch wurden auch einige Einschränkungen und weiterer Forschungsbedarf aufgezeigt. Beispielsweise ließ sich die Annahme der Eindimensionalität des NEPS-Mathematiktests nicht bestätigen. Außerdem wären weitere Analysen bezüglich des Zusammenhanges mit geeigneten Kriterien für die Zieldomäne Mathematische Kompetenz für eine Erweiterung der Testwertinterpretationen wünschenswert. Die gefundenen Evidenzen stützen jedoch die Testwertinterpretation „die Unterschiede in der Rangfolge der beobachteten Testleistung deuten auf Unterschiede in der Rangfolge der Ausprägung der mathematischen Kompetenz, wie diese im Rahmenkonzept definiert wird“. So konnten für die Zieldomäne relevante Teilkompetenzen in den NEPS-Aufgaben identifiziert werden, waren die Bewertungskriterien angemessen und wurden erwartungsgemäße Zusammenhänge mit den Mathematiktests aus dem LV und PISA gefunden. Auf diese Weise konnte eine vollständige Argumentationskette für den NEPS-K9-Mathematiktest entwickelt, sowie Empfehlungen für die Stärkung und Erweiterung der Testwertinterpretation empfohlen werden. Der Argument Based Approach von Kane erwies sich dabei als hilfreicher Ansatz, in welchem die Vorgaben der Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 2014) berücksichtigt werden konnten.

Inhaltsverzeichnis

Tabellenverzeichnis

Abbildungsverzeichnis

Abkürzungsverzeichnis

1PL Einparameter-Logistisch
2PL Zweiparameter-Logistisch
3PL Dreiparameter-Logistisch
AERA American Educational Research Association
AIC Akaike Information Criterion
APA American Psychology Association
BEFKI Berliner Test zur Erfassung fluider und kristalliner Intelligenz
BIC Bayesian Information Criterion
BLK Bund-Länder-Kommission für Bildungsplanung und Forschungsförderung
CAIC Consistent Aiken Information Criterion
CATI computergestütztes Telefoninterview
CEL Coefficient of effective length
CMC Complex Multiple Choice
DIF Differentielles Item Funktionieren
DPC Data Processing and Research Center
FCAT Florida Comprehensive Assessment Test
ICC Item Characteristic Curve
ICT Informationstechnologien
IEA International Association for the Evaluation of Educational Achievement
IQB Institut zur Qualitätsentwicklung im Bildungswesen
IRT Item Response Theorie
ISEI International Socio-Economic Index of Occupational Status
IUA Interpretation/Use Argument
K5 Klassenstufe 5
KMK Ständige Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland
LSA lokale stochastische Abhängigkeit

LSU lokale stochastische Unabhängigkeit
MEG Mathematics Expert Group
LV IQB-Ländervergleich
MC Multiple Choice
MCMLM Mixed Coefficient Multinomial Logit Model
MDTP Mathematics Diagnostic Testing Project
MML Weighted Maximum Likelihood Estimate
MNSQ Mean Square
MPT-G General Mathematics Placement Test
MSCEIT Mayer-Salovey-Caruso Emotional Intelligence Test
NaWi Naturwissenschaften
NCME National Council on Measurement in Education
NEPS National Educational Panel Study
OECD Organisation for Economic Cooperation and Development
PCM Partial-Credit-Model
PISA Programme for International Student Assessment
PÜ prozentuale Übereinstimmung
PV Plausible Value
SCR Short Constructed Response
SINUS Steigerung der Effizienz des mathematisch-naturwissenschaftlichen Unterrichts
Standards Standards for Educational and Psychological Testing
TIF Testinformationsfunktion
TILT Test zur Erfassung technologischer und informationsbezogener Literacy
TIMSS Trends in International Mathematics and Science Study
TOEFL Test of English as a Foreign Language
VERA Vergleichsarbeiten
WLE Weighted Maximum Likelihood Estimate
WMNSQ Weighted Mean Squares
YITS Youth in Transition Survey

Einleitung

In den letzten Jahrzehnten ist ein steigendes Interesse an Large Scale Schulleistungsstudien zu beobachten. Obwohl es schon immer Diskussionen über die Entwicklung des deutschen Schulsystems gab, stand über viele Jahrzehnte hinweg vorwiegend die Schulorganisation im Mittelpunkt (Input-Orientierung), die Ergebnisse beziehungsweise der Output der Bildungsprozesse wurden jedoch nicht systematisch überprüft. Als 1997 die erste deutschsprachige Publikation von Trends in International Mathematics and Science Study (TIMSS) veröffentlicht wurde, war der Schock über die nur mittelmäßigen Leistungen deutscher Schülerinnen und Schüler groß. Die Aufmerksamkeit der Öffentlichkeit und der Bildungspolitik wurde neu strukturiert und die Ergebnisse des Bildungsprozesses traten in den Vordergrund der Diskussionen (Output-Orientierung; S. Weinert et al., 2011). In diesem Zusammenhang hat die Ständige Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (KMK) 1997 mit den Konstanzer Beschlüssen die Qualitätssicherung im deutschen Schulwesen zu ihrem zentralen Thema gemacht und ist darin übereingekommen, konsequent an nationalen und internationalen Schulleistungsstudien teilzunehmen. Seitdem nimmt Deutschland regelmäßig an internationalen Schulleistungsstudien wie PISA und TIMSS teil und führt periodisch nationale Vergleichsstudien wie den LV und Vergleichsarbeiten (VERA) durch (Sekretariat der ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland, 02.06.2006). International ist ein ähnlicher Trend in der vermehrten Teilnahme an Large Scale Assessment zu erkennen. So stieg die Anzahl der Teilnehmerstaaten der PISA-Studie von 43 im Jahr 2000 auf 65 im Jahr 2009 (Klieme et al., 2010). Auch die Anzahl der Länder, die an der TIMS-Studie für die vierte und/oder achte Klasse teilgenommen haben, erhöhte sich von 43 teilnehmenden Staaten im Jahr 2003 auf 77 Staaten im Jahr 2011 (Mullis, Martin, Foy & Arora, 2012). Ein Großteil der Large Scale Assessments, die in Deutschland durchgeführt werden, hat ein querschnittliches Untersuchungsdesign (Blossfeld & von Maurice, 2011) und verfolgt in erster Linie die Überwachung und Überprüfung des Bildungswesens (Systemmonitoring), das Ver-

gleichen der Qualität von Bildungssystemen, die Bestimmung des nationalen Standorts (Benchmarking), das Testen von kriteriumsorientierten, normorientierten und ipsativen Standards und das Untersuchen von kausalen Beziehungen zwischen Bildungsvoraussetzungen und Bildungsergebnissen (Seidl & Prenzel, 2008; Bos & Schwippert, 2002). Es ist jedoch auch eine große Notwendigkeit für längschnittliche Bildungsforschung entstanden. Bildung hat sich in der modernen Gesellschaft zu einem lebenslangen Prozess entwickelt und gilt als eine Voraussetzung für die aktive Partizipation verantwortungsvoller Bürgerinnen und Bürger in einer demokratischen Gesellschaft. Bisher ist allerdings noch wenig über die kumulative Entwicklung von Kompetenzen innerhalb einer Lebensspanne bekannt (Blossfeld & von Maurice, 2011). Um Einsicht in den Bildungsprozess und die Kompetenzentwicklung zu erhalten, wurde in Deutschland die National Educational Panel Study (NEPS) entwickelt. Diese Studie wird im deutschen Sprachraum auch als das Nationale Bildungspanel bezeichnet. Das NEPS untersucht, wie sich Kompetenzen im Laufe des Lebens entwickeln, wie diese Kompetenzen die Bildungskarriere beeinflussen und inwiefern die Kompetenzentwicklung von Lerngelegenheiten beeinflusst wird. Die längsschnittliche Integration von Bildungsstufen in NEPS wird durch die Orientierung an fünf Dimensionen (Säulen) gewährleistet (Abbildung 1). Die erste Säule bezieht sich auf die *Kompetenzentwicklung*, die zweite Säule auf die *Lernumwelten*, die dritte Säule auf die *Bildungsentscheidungen*, die vierte Säule auf den *Migrationshintergrund* und die fünfte Säule auf *Bildungsrenditen*. Die Messung von Kompetenzen (Säule 1) gehört zu den zentralen Voraussetzungen für die zu beantwortenden Fragestellungen. Hierzu zählen unter anderem mathematische Kompetenz, Lesekompetenz und naturwissenschaftliche Kompetenz. Das NEPS differenziert die Bildungskarriere angelehnt an das deutsche Bildungssystem in 8 Etappen. Die erste Etappe beinhaltet *Neugeborene und den Übergang in die frühkindliche Betreuung* und die letzte Etappe *Erwachsenenbildung und lebenslanges Lernen*. Eine wichtige Etappe ist beispielsweise die *Sekundarstufe I und der Übergang in die Sekundarstufe II beziehungsweise in das Berufsleben* (Etappe vier). In dieser Etappe kann erstmalig der Einstieg in das Berufsleben stattfinden. Erhoben werden die Daten mit Hilfe eines Multi-Kohorten-Sequenz-Designs, wobei in allen Bildungsetappen verschiedene Instrumente zum Einsatz kommen (Blossfeld & von Maurice, 2011). So gibt es beispielsweise unterschiedliche Tests zur Erfassung der mathematischen Kompetenzen in der fünften und neunten Klasse. Die Ergebnisse und Daten der Studie werden der Wissenschaft in Form von *scientific use files* zu Forschungszwecken zur Verfügung gestellt. Das NEPS soll mit diesem Design repräsentative Daten für wichtige Bereiche des Bildungssystems erzeugen, welche eine Exploration der Kompetenzentwick-

lung und eine Untersuchung ihrer Rolle in der Bildungskarriere ermöglichen (Ehmke et al., 2009, S. 314).

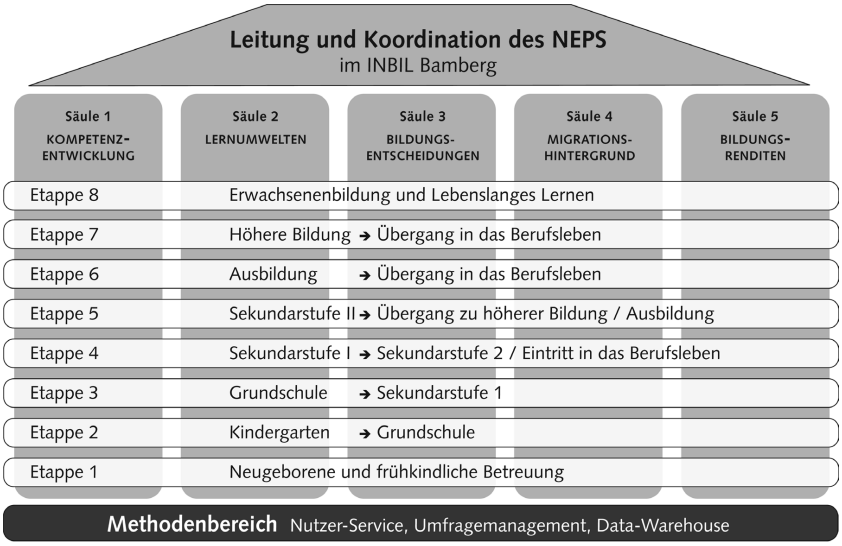


Abbildung 1: Forschungsbereiche (Säulen) und Altersstufen (Etappen) aus Ehmke et al. (2009)

Large Scale Assessments im Bereich der Leistungsmessung wie NEPS, PISA und LV untersuchen in erster Instanz die Leistungen der Testpersonen. Die hierbei gesammelten Rohwerte sind, als bestimmte Antworten auf bestimmte Aufgaben in spezifischen Testsituationen, in ihrer Reinform nicht sehr bedeutend. Unabhängig von den Testaufgaben, dem Testformat, der Testbewertung etc., gewinnen die Daten erst als Indikator für Fähigkeiten an Bedeutsamkeit. Ob die Ziele von Large Scale Assessments tatsächlich erreicht werden, ist daher auch davon abhängig, inwieweit die Studien verlässliche Ergebnisse zu den Leistungen und Kompetenzen der getesteten Personen sowie verlässliche Schlussfolgerungen zur Qualität des Bildungssystems beziehungsweise zur Kompetenzentwicklung zulassen. Die Validität von Large Scale Assessments ist eine zentrale Voraussetzung für die Legitimation der beabsichtigten Testwertinterpretationen und ist daher von großer Bedeutung. Für viele Tests wurden die beabsichtigten Testwertinterpretationen jedoch nicht vollständig evaluiert (Brennan, 2013). Dies gilt auch für die neu entwickelten Tests des NEPS. Zwar wurden einzelne Analysen bezüglich einiger Testwertinterpretationen veröffentlicht (vgl. Duchhardt & Gerdes, 2012, 2013; Senkbeil, Ihme & Wittwer, 2013; Pohl, Grafe & Rose, 2014), jedoch wurde bisher keine vollständige Auswertung aller beabsichtigten Testwertinterpretationen eines Tests durchgeführt. Diese

Arbeit soll daher einen Beitrag zur Validierung der NEPS-Testwertinterpretationen leisten, indem der Mathematiktest für die neunte Klassenstufe (K9) umfassend analysiert wird. Dabei wurde der NEPS-K9-Mathematiktest aufgrund seiner wichtigen Stellung in NEPS ausgewählt. Zum einen wird die mathematische Kompetenz in NEPS als eine wesentliche Basiskompetenz betrachtet, die besonders wichtig für den Bildungsverlauf und gesellschaftlichen Erfolg ist (Ehmke et al., 2009). Zum anderen wird dieser Test in der bedeutenden Etappe vier eingesetzt (s.o.). Bevor jedoch die Testwertinterpretationen des NEPS-K9-Mathematiktests validiert werden, sollen zunächst die theoretischen Grundlagen für die Validierung dargelegt werden und ebenfalls soll auf mathematische Kompetenztests eingegangen werden. Anschließend soll ein Schema für die Validierung entwickelt und der Test anhand dieses Schemas ausgewertet werden. Die Arbeit soll mit einer Diskussion der Validierung abschließen.

1 Theoretischer Hintergrund

In diesem Kapitel wird zuerst auf die Entwicklung des Validitätskonzeptes eingegangen. Es wird beschrieben, wie dieses Konzept zuerst zergliedert und anschließend wieder zusammengeführt wurde. Anschließend wird auf die argumentationsbasierten Ansätze eingegangen, die sich mit der Zusammenführung des Validitätskonzeptes entwickelten. Darauf folgt ein Einblick in die kontroversen Diskussionen, die bezüglich des zusammengeführten Validitätskonzeptes entstanden und eine Darlegung der Diskrepanz zwischen dem Validitätskonzept und der Validierungspraxis. In einem zweiten Unterkapitel wird auf den Kompetenzbegriff eingegangen und ein Überblick über mathematische Kompetenzmessungen in Deutschland gegeben. Diesbezüglich werden die mathematischen Rahmenkonzepte und bisherigen Evidenzen zur Validität der Testwertinterpretationen der Mathematiktests aus PISA 2012 und dem LV 2012 dargelegt. Außerdem wird das Rahmenkonzept des NEPS-K9-Mathematiktests beschrieben und mit den Rahmenkonzepten aus PISA 2012 und dem LV 2012 verglichen.

1.1 Validität

Die Komponenten des Validitäts-Konzeptes, wie die exakte Definition und die Art der notwendigen oder wünschenswerten Nachweise für haltbare und vertretbare Schlussfolgerungen, wurden in der Wissenschaft kontrovers diskutiert und modifiziert. Geisinger (1992) beschrieb, dass die Validitätstheorie während des letzten halben Jahrhunderts eine Metamorphose durchlaufen habe. Im Folgenden soll diese Entwicklung beschrieben werden.

1.1.1 Entwicklung und Zergliederung der Validität

Kriteriumsvalidität

Zwischen 1920 und 1950 wurde die Kriteriumsvalidität als der Goldene Standard der Validität angesehen (Shepard, 1993; Sireci, 2007). Validieren bedeutete zu untersuchen, wie gut die Testergebnisse das Testkriterium vorhersagten. Ein Test wurde für alle Kriterien als valide betrachtet, für welche er akkurate Vorhersagen schätzen konnte (Gulliksen, 1950). Die Kriteriumsvalidität wurde mittels Korrelationen zwischen Test und Kriterium berechnet. Es gab zwei Arten der Kriteriumsvalidität: die konkurrente Validität, bei der das Kriterium gleichzeitig mit den Testergebnissen erhoben wurde und die prädiktive Validität, bei der ein zukünftiges Kriterium nach der Erhebung der Testergebnisse gemessen wurde (Lissitz & Samuelsen, 2007). Vorteile der Kriteriumsvalidität waren die oft hohe Relevanz der Berechnung für die Plausibilität der Testinterpretationen und die Objektivität der Berechnungen (Kane, 2006). Ein Nachteil der Kriteriumsvalidität war die Schwierigkeit, ein geeignetes Kriterium zu finden (Thorndike, Bregman, Cobb & Woodyard, 1927).

Inhaltsvalidität

In diesem Zusammenhang diskutierten verschiedene Autoren über mögliche Rollen des Testinhaltes in der Validitätstheorie. So empfahl Kelley (1927), neben den Korrelationen zwischen Test und Kriterium auch Experteneinschätzungen ergänzend zur Evaluation hinzu zu ziehen. Auch Lennon wies auf die Möglichkeit des Gebrauchs von Inhaltsvalidität hin, wenn kein geeignetes Kriterium zur Verfügung steht oder wenn Korrelationen keine sinnvollen Indikatoren für die Validität darstellen. Die Validierung basiert dann auf Einschätzungen der Relevanz und Repräsentativität der Testitems (Lennon, 1956). Die Inhaltsvalidität tendierte daher allerdings zur Subjektivität und zum Confirmatory Bias (Kane, 2006).

Konstruktvalidität

Cronbach und Meehl (1955) stellten die Konstruktvalidität als Alternative zur prädiktiven, konkurrenten und Inhaltsvalidität vor. Die Autoren schlugen vor, Konstruktvalidität dann zu untersuchen, wenn das zu messende Merkmal nicht operationalisiert ist und weder ein zufriedenstellendes Kriterium noch eine adäquate inhaltliche Gesamtheit definiert werden kann. Das zu messende Konstrukt soll durch ein nomologisches Netzwerk definiert werden, aus welchem operationalisierbare Hypothesen abgeleitet werden können. Die 1966 und 1974 erschienenen Standards for Educational and Psychological Testing (Standards), welche in Zusammenarbeit von American Educational Research Association (AERA), American Psychology Association (APA) und National Council on Measurement in Education (NCME) entwickelt und seither periodisch überarbeitet wurden, übernahmen die Konstruktvalidität als vierte Art der Validität, neben der prädiktiven, der konkurrenten und der Inhaltsvalidität (*Standards for educational and psychological tests and manuals*, 1966, 1974).

1.1.2 Zusammenführung des Validitätskonzeptes

Ende der 1970er Jahre beobachteten und bemängelten Validitätsexperten den Trend, die Validität als Werkzeugkoffer zu betrachten, wobei unterschiedliche Methoden bei bestimmten Assessments eingesetzt wurden. So beanstandete Messick (1980), dass die Aufteilung der Validität in Kriteriums-, Konstrukt- und Inhaltsvalidität eine zu starke Vereinfachung sei. Die Konsequenz dieser Vereinfachung sei, dass sich Testanwender auf die Validitätskategorie konzentrieren, anstatt auf die Schlussfolgerungen, die sie aus den Testergebnissen ziehen wollen. Auch Guion (1978) kritisierte, dass die verschiedenen Aspekte der Validität viel zu oft als unterschiedliche Konzepte und somit als unabhängige Alternativen betrachtet würden. Anastasi (1986) beanstandete ebenfalls die Kategorisierung der Validität. Testentwicklerinnen und Testentwickler würden sich verpflichtet fühlen, die Arten der Validität checklistengleich abzuarbeiten, ungeachtet des Ziels und des Verwendungszweckes des Testes. Loevingers (1957) plädierte dafür, dass Validität ein ganzheitliches Konstrukt sei, da die anderen Formen der Validität (prädiktiv, konkurrent, Inhalt) ad hoc seien. Diese Validitätsperspektive nannte er Konstruktvalidität. In den neunzehnhundertachtziger Jahren wurde diese Sichtweise auf Validität schließlich weithin angenommen (Anastasi, 1986; Whitely, 1983; Messick, 1980, 1989a). So

wurde Validität in den Standards von 1985 als die holisistische Evaluation der Testergebnisinterpretationen definiert, wobei das Inhalts-, Konstrukt- und Kriteriummodell verschiedene Arten der Evidenz bieten können. Im Jahr 1988 kritisierte Messick (1988) jedoch die Standards von 1985. Die Aussage, dass verschiedene Validierungsbestrebungen unterschiedliche Arten der Evidenz benötigen, lässt laut Messick die Interpretation zu, dass unter bestimmten Umständen nur eine Art der Validitätsevidenz angemessen ist. Die Formulierung der Standards fördert daher die Stützung auf sehr begrenzter Validitätsevidenz. Im Folgenden soll die Konstruktvalidität nach Messick genauer beschrieben werden, da dieser einen maßgeblichen Einfluss auf die Denkweise dieser Zeit hatte und den Zeitgeist zwischen Mitte der neunzehnhundertsiebziger bis Ende der neunzehnhundertneunziger Jahre widerspiegelt (Newton & Shaw, 2014).

Konstruktvalidität nach Messick

Um die Validierung weniger komplex zu machen und Nuancen und Themen hervorzuheben, die durch die bisher mangelhafte Definition von Validität tendenziell übergangen oder vernachlässigt werden konnten, entwickelte Messick (1989a, 1994, 1995) sechs Aspekte der Validität, welche in Tabelle 1 dargestellt werden.

Messick identifizierte die Unterrepräsentanz des Konstruktes und die konstruktirrelevante Varianz als zwei mögliche Quellen der Invalidität. Die Unterrepräsentanz des Konstruktes ist die Folge eines zu begrenzten Assessments, welches nicht alle Komponenten des Konstruktes erfasst. Die konstruktirrelevante Varianz ist das Gegenteil hiervon und wird durch das Erfassen von mehr als nur dem zu messenden Konstrukt bedingt. Messick (1989a) forderte einen Prozess der Validierung, der den unangemessenen Fokus auf selektive Formen der Evidenz verhindert, der die wichtige, ergänzende Rolle von spezifischer inhalts- und kriteriumsbezogener Evidenz zur Konstruktvalidität hervorhebt und der Rücksicht auf die Wertimplikationen und sozialen Konsequenzen nimmt. Zu einem einheitlichen Validitätskonzept gehört nach diesem Ansatz die Unterscheidung zweier verbundener Facetten der Validität: die Testinterpretation als Legitimation des Testes, basierend auf einer Einschätzung der Bedeutung und der Konsequenzen der Testergebnisse sowie die Testnutzung als Funktion oder Ergebnis des Testes, in Form der Interpretation und der Anwendung (Abbildung 2).

Nach Messick bildet die Kontruktvalidität die Basis für eine angemessene Testinterpre-

Tabelle 1: Aspekte der Konstruktvalidität nach Messick (1995, S. 11-18)

Aspekte der Konstruktvalidität	Art der Evidenz
Inhalt	Analyse der inhaltlichen Relevanz und Repräsentativität
Theorie und Prozessmodelle	Theoretische Begründungen für die beobachtete Konsistenz in den Testwerten, einschließlich von Prozessmodellen der Testleistung und empirischer Evidenz über den tatsächlichen Gebrauch der theoretischen Prozesse bei der Aufgabenbearbeitung
Struktur	Beurteilung der Konsistenz der internen Struktur des Assessments mit der internen Struktur des zu erfassenden Konstrukts
Generalisierbarkeit	Abschätzung der Generalisierbarkeit und Grenzen der Bedeutung von Testergebnissen
externe Aspekte	Prüfung von konvergenten und divergenten Zusammenhängen mit externen Variablen
Konsequenzen	Analyse beabsichtigter und unbeabsichtigter Konsequenzen der Testwertinterpretation und -nutzung

	Testinterpretation	Testnutzung
Ebene der Nachweise	Konstruktvalidität (KV)	KV + Relevanz/Nützlichkeit (R/N)
Ebene der Konsequenzen	KV + Wertimplikationen (WI)	KV + R/U + soziale Konsequenzen

Abbildung 2: Progressive Matrix der Validitätsfacetten (Messick, 1989a)

tation auf der Ebene der Nachweise. Diese beinhaltet die Evaluation der Testwertinterpretationen durch unterschiedliche Arten empirischer Evidenz und logischer Analysen. Die Basis für eine angemessene Testnutzung auf der Ebene der Nachweise ist neben der Konstruktvalidität auch die Relevanz oder Nützlichkeit der Testwerte für den jeweiligen Verwendungszweck. Die Testinterpretation auf der Ebene der Konsequenzen umfasst neben der Konstruktvalidität auch die zugrundeliegenden Wertimplikationen der Testwerte, des Testkonstruktes und der Annahmen. Diese Wertimplikationen, welche sich oft schon in der Bezeichnung der zu erfassenden Eigenschaft befinden, müssten für das Konstrukt spezifiziert und geprüft werden. Die Testnutzung auf Ebene der Konsequenzen beinhaltet neben der Konstruktvalidität und der Relevanz/Nützlichkeit auch die Einschätzung der sozialen Konsequenzen der Testdurchführung. Die Relevanz/Nützlichkeit kann evaluiert werden, indem die Vorteile und Risiken der Testnutzung den Vor- und Nachteilen von Alternativen gegenübergestellt werden. Bezüglich der sozialen Konsequenzen soll evaluiert werden, wie diese entstanden sind und was diese bedingt hat. Nach Messick kann zusammenfassend geschlossen werden, dass Bedeutung und Wert sowie Testnutzung und Testinterpretation im Validitätsprozess miteinander verflochten sind und Validieren sowohl Wissenschaft als auch Ethik beinhaltet (Messick, 1989a).

Der Ansatz von Messick, welcher Validität als einheitliches Konzept definiert und Konsequenzen der Testnutzung integriert, beeinflusste und spiegelte den professionellen Konzens über Validierung zu dieser Zeit (Newton & Shaw, 2014). Im Laufe der Jahre zeigte jedoch auch dieser Ansatz Schwachstellen, die durch die folgende Kritik von Shepard (1993) zusammengefasst werden. Die Unterscheidung der Validierung in Facetten er-

weckt laut Shepard den Eindruck, dass die wissenschaftliche Evaluation und die Evaluation der Werteimplikationen von Testwerten voneinander getrennt seien. Des Weiteren bleibe aufgrund des Verortens der Konstruktvalidität auf der ersten Ebene und der Zufügung von zusätzlichen Aspekten auf den anderen Ebenen unklar, ob Konstruktvalidität ein Teilaspekt oder ein übergeordneter Begriff sei. Darüber hinaus verursache die Komplexität von Messicks Validitätskonzept Unklarheit über die Relevanz unterschiedlicher Validitätsnachweise. Aus diesen Gründen formulierte Shepard (1993) Implikationen für die neunzehnhundertneunziger Standards. Shepard plädierte für ein einheitliches Validitätskonzept in den neuen Standards, wobei der Prozess der Validierung von einem Konstrukt der theoretischen Beziehungen von Testwerten zu tatsächlichen Kompetenzen und durch die beabsichtigte Testnutzung gesteuert wird. Dabei sollten auch alternative Testwertbedeutungen und unbeabsichtigte Konsequenzen gründlich untersucht werden. Angelehnt an Cronbach (1989) kann Validierung hierbei als ein Prozess der evaluativen Argumentation betrachtet werden. Ferner forderte Shepard schlüssige Richtlinien für die Priorität von Validitätsnachweisen sowie Klarheit darüber, inwieweit kritische/wenig bewiesene Annahmen tragbar sind. En weiterer wichtiger Punkt war nach Shepard eine eindeutige Verteilung der Verantwortung für die Validität zwischen Testnutzerinnen und Testnutzer sowie Testentwicklerinnen und Testentwicklern. Testentwicklerinnen und Testentwickler sollten Nachweise für die ursprüngliche Testintention erbringen sowie konkurrierende Interpretationen und potentielle Nebenwirkungen der Testnutzung berücksichtigen. Testnutzerinnen und Testnutzer, die einen Test in einem anderen Kontext zu nutzen gedachten, müssten die Validitätsnachweise nachprüfen und neue, notwendig gewordene Analysen durchführen.

1.1.3 Argumentationsbasierte Ansätze

Viele der von Shepard (1993) ausgesprochenen Forderungen wurden in neueren Validitätsansätzen berücksichtigt. Die sogenannten argumentationsbasierten Ansätze gründen auf der ganzheitlichen Perspektive der Konstruktvalidität, in welcher Validität als einheitliches Konzept verstanden wird, welche die Konsequenzen der Testnutzung mit einschließt und in der Sammeln von Evidenz bezüglich der beabsichtigten Testwertinterpretationen in Form von Argumenten geschieht. In 1999 wurde diese Sichtweise in den Standards aufgenommen und in 2014 noch weiter spezifiziert (American Educational Research Association, American Psychological Association & National Coun-

cil on Measurement in Education, 1999; American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 2014). Neuere Validitätsansätze, wie das „Evidence Centered Assessment Design“ von Mislevy (2007) und Mislevy, Steinberg und Almond (2003), das „Assessment Use Argument“ von Bachman (2005) und der „Arguments Based Approach“ von Kane, orientieren sich an dieser Sichtweise und schlagen konkrete Schemata für die Bildung von sogenannten *Validitätsargumenten* basierend dem Argumentationsschema von Toulmin (1958, 2003) vor. Im Folgenden werden zuerst die Standards von 1999 und 2014 beschrieben. Anschließend wird auf das Argumentationsschema von Toulmin (1958, 2003) eingegangen und es werden die argumentationsbasierten Ansätze von Mislevy et al. (2003) und Mislevy (2007), Bachman (2005) und Kane (2013) dargelegt.

Standards für Pädagogisches und Psychologisches Testen von 1999 und 2014

In den Standards von 1999 (American Educational Research Association et al., 1999) wurden viele der geforderten Aspekte von Shepard (1993) aufgenommen und auch in den überarbeiteten Standards von 2014 (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 2014) berücksichtigt. Validität wird als das Ausmaß definiert, in welchem die empirischen Beweise und die Theorien die beabsichtigten Interpretationen von Testwerten unterstützen, wobei die geplante Testnutzung berücksichtigt wird. Dabei wird hervorgehoben, dass nicht der Test selbst, sondern die Interpretation der Testwerte im Zusammenhang mit der geplanten Testnutzung evaluiert wird. Validierung kann als das Entwickeln eines wissenschaftlich soliden Argumentes zur Unterbauung der beabsichtigten Interpretation der Testwerte und der Relevanz für die geplante Testnutzung betrachtet werden. Welche Nachweise die beabsichtigten Interpretationen und die Testnutzung am besten stützen, muss durch Testentwicklerinnen und Testentwickler oder Testnutzerinnen und Testnutzer mit entsprechender Expertise entschieden werden. Außerdem wird die Berücksichtigung von rivalisierenden Hypothesen zur beabsichtigten Interpretation von unterschiedlichen Perspektiven verschiedener Parteien, von bestehenden Erfahrungen mit ähnlichen Tests und von zu erwartenden Konsequenzen der Testnutzung angeraten. Mögliche Invarianz durch Konstruktunterrepräsentation und konstruktirrelevanter Varianz müssen untersucht werden. Die Verantwortung für die Validität tragen Testentwicklerinnen und Testentwickler sowie Testnutzerinnen und Testnutzer gemeinsam. Testentwicklerin-

nen und Testentwickler sind vor allem für die Validierung der von ihnen beabsichtigten Testinterpretation und Testnutzung verantwortlich, während Testnutzerinnen und Testnutzer die Verantwortung für die Validierung der Interpretation und Konsequenzen im Rahmen ihrer Testnutzung tragen (American Educational Research Association et al., 1999; American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 2014). In den Standards von 2014 wird zusätzlich hervorgehoben, dass Validierung an sich ein nie endender Prozess ist, da es immer zusätzliche Informationen gibt, die helfen, einen Test und die beabsichtigten Schlussfolgerungen besser zu verstehen. Jedoch erlauben die gefundenen Evidenzen an einem bestimmten Punkt eine zusammenfassende Beurteilung über die beabsichtigte Testwertinterpretation (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 2014). Die Standards beinhalten statt der drei Arten der Validitätsnachweise von 1985 (Konstrukt, Kriterium und Inhalt) nun fünf Kategorien der Evidenz. Um sich von den ausdifferenzierten Validitätstypen früherer Validitätsperspektiven abzugrenzen und die ganzheitliche Definition von Validität hervorzuheben, werden für die Beschreibung der fünf Evidenzkategorien neue Fachausdrücke verwendet (American Educational Research Association et al., 1999; American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 2014). Diese werden in Tabelle 2 dargestellt.

Tabelle 2: Kategorien relevanter Evidenz nach den Standards von 1999 und 2014

Evidenzkategorien	Art der Evidenz
Testinhalt	Beurteilung der Passung zwischen Testinhalt (Inhalt, Formulierung und Format der Testitems/-aufgaben/-fragen, Richtlinien für Administration und Scoring) und zugrundeliegendem Konstrukt (z.B. durch Expertenratings, logische und empirische Analysen der Testaufgaben).
Antwortprozesse	Beurteilung der Passung zwischen Konstrukt und Testleistung durch theoretische und empirische Analysen der Antwortprozesse von Testpersonen (Methoden z.B. schriftliche Konzepte, elektronisch überwachte Verbesserungen, Reaktionszeit und Augenbewegungen).
Interne Struktur	Beurteilung, inwiefern die Beziehungen zwischen einzelnen Testitems/Testkomponenten dem Konstrukt entsprechen. Analysemethoden und Interpretationen sind abhängig von der Testnutzung (z.B. Itemhomogenität für eindimensionale Tests, Abwesenheit von Differential Item Functioning für geschlechterunabhängige Tests).
Zusammenhänge mit anderen Variablen	Beurteilung inwieweit Beziehungen zu externen Variablen dem Konstrukt entsprechen durch Konvergente Evidenz, diskriminante Evidenz, prädiktive Evidenz, konkurrente Evidenz und Validitätsgeneralisierung.
Testkonsequenzen	Beurteilung der beabsichtigten und unbeabsichtigten Testkonsequenzen sowie der Annahmen über die Testnutzung, welche nicht direkt auf den Testwertinterpretationen basieren. Nachweise über Konsequenzen auf Invalidität durch z.B. Konstruktunterrepräsentation oder konstruktirrelevante Validität sowie Nachweise über den Einfluss des Testprogramms.

Die Standards von 1999 und 2014 bilden unter anderem das Einvernehmen der Validitätsexperten darüber ab, dass der Prozess der Validierung das Sammeln von Evidenz für eine wissenschaftlich solide Basis der Testinterpretationen und Schlussfolgerungen beinhaltet. Validität ist in diesem Sinne graduell. Für unterschiedliche Tests werden unterschiedliche Arten der Evidenz benötigt, wobei die Validität ein einheitliches Konstrukt bleibt (Kane, 2006, 2013; Shaw & Crisp, 2012; American Educational Research Association et al., 1999; Messick, 1989a). Der Validität als wieder zusammengeführtes, ganzheitliches Konstrukt, wie es in den Standards beschrieben wird, fehlt es allerdings an deutlichen Richtlinien für die Praxis (Brennan, 2013; Chapelle, 2012; Chapelle, Enright & Jamieson, 2010; Moss, 2007). Chapelle et al. (2010) kamen bei der Validierung des Test of English as a Foreign Language (TOEFL)-Tests zu dem Schluss, dass die Standards explizitere Richtlinien bezüglich der Formulierung von beabsichtigten Interpretationen und Absichten für die Art der Validitätsbeweise benötigen würden. Ihrer Meinung nach werde in den Standards auch nicht deutlich, wie Gegenbeweise in Relation zu Beweisen interpretiert werden sollten. Des Weiteren seien die Standards konstruktbasiert, weshalb für die Validierung eine solide Theorie des Konstruktes benötigt werde, die nicht für alle Tests vorhanden sei. Brennan (2013) kritisierte vor allem die Fokussierung auf die unterschiedlichen Arten der Evidenz, ohne eine klare a priori Spezifizierung der beabsichtigten Interpretationen. Kane (2012) merkte an, dass das vereinigte Validitätsmodell der Konstruktvalidität zwar vielfältig und sinnvoll, jedoch nur schwer zu implementieren sei, da es weder einen logischen Startpunkt, noch einen Leitfaden oder Kriterien zum abschätzen des Fortschrittes in der Validierung gebe.

Argumentationsschema nach Toulmin

In Anlehnung an die ganzheitliche Perspektive auf Validität (Messick, 1989a; American Educational Research Association et al., 1999; American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 2014) wurden unterschiedliche argumentationsbasierte Ansätze entwickelt, die versuchen, den Prozess der Validierung stärker zu strukturieren (Bachman, 2005; Kane, 2013; Mislevy et al., 2003). Diese Ansätze nutzen das Argumentationsschema von Toulmin (1958, 2003) für präsumtives Argumentieren. In seinem Werk „The uses of argument“ (Toulmin, 1958), welches in 2003 als aktualisierte Ausgabe neu erschien (Toulmin, 2003), beschrieb Toulmin den Aufbau von präsumtiver Argumentation. In Toulmins

Schema zieht der Argumentierende basierend auf einer „Grundlage“ (data) und mit Hilfe eines „Argumentes“ (warrant) eine „Schlussfolgerung“ (claim). Dabei liefert eine Stützung (backing) Beweise für die Schlussfolgerung und eine Ausnahme (rebuttal) beschränkt oder verstärkt die Behauptung. Die Schlussregeln haben oft einen Qualifikator (qualifier), welche die Stärke der Behauptung zum Ausdruck bringt (Toulmin, 2003) (Abbildung 4).

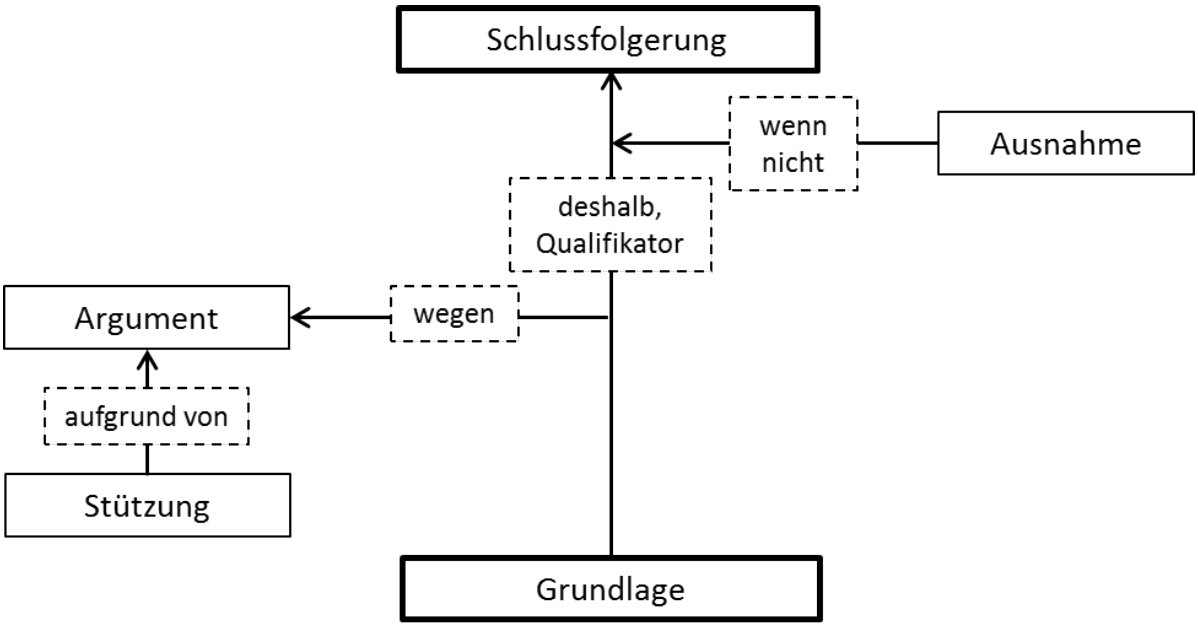


Abbildung 3: Modell für das Analysieren von präsumtiven Argumenten angelehnt an (Toulmin, 1958)

Auf ein Beispiel angewendet könnte das Argumentationsschema wie folgt aussehen. Aufgrund von Testergebnissen (Grundlagen) schließen wir darauf, dass eine Testperson eine Aufgabe, welche eine bestimmte Schwierigkeit hat, mit einer Wahrscheinlichkeit von 65% (Qualifikator) lösen kann (Schlussfolgerung). Das Argument dafür bildet die Passung des Item Response Theorie (IRT)-Modells und die Stützung dieses Arguments basiert auf der Auswertung der Passung des Modells zu den Daten. Eine Ausnahme kann ein differentielles Funktionieren der Aufgabe bezüglich des Geschlechts darstellen. Dies würde bedeuten, dass eine Aufgabe bei gleicher Fähigkeit beispielsweise für Mädchen schwieriger ist als für Jungen.

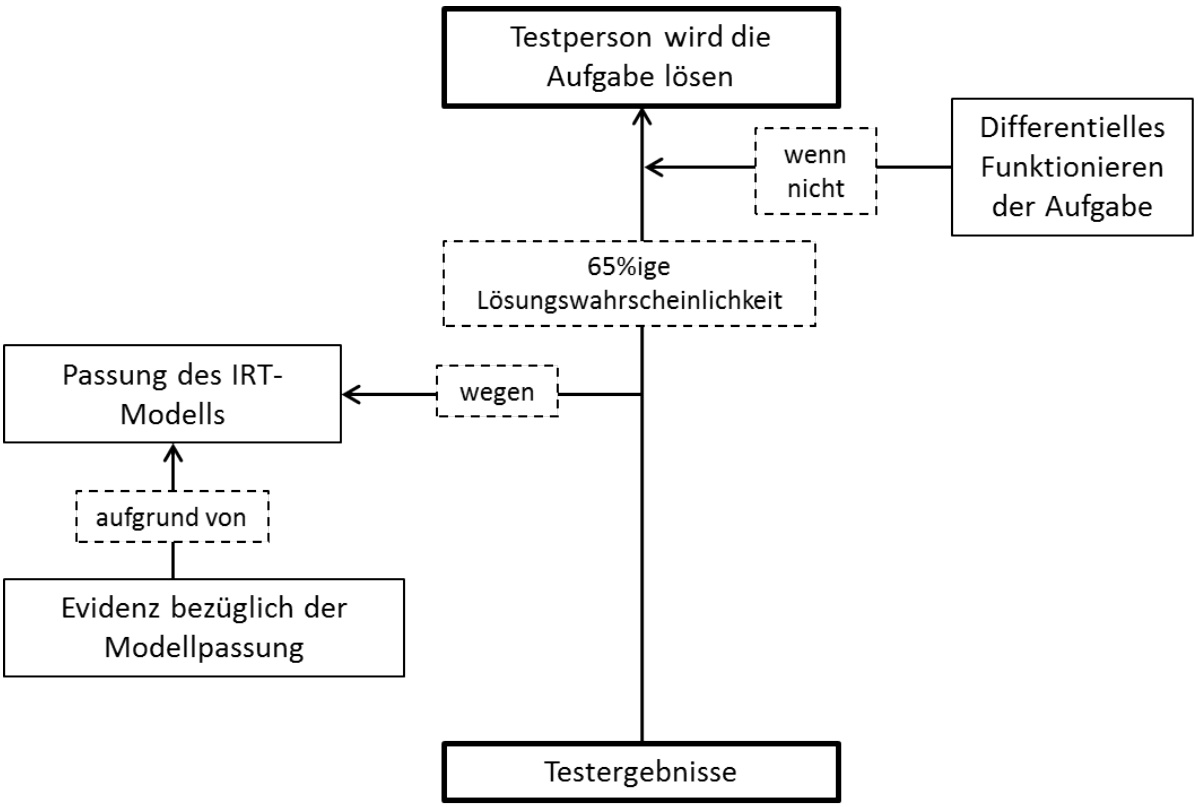


Abbildung 4: Beispiel für das Analysieren von präsumtiven Argumenten nach Toulmin (1958, 2003)

Evidence Centered Assessment Design

Mislevy (2007) entwickelte das *Evidence Centered Assessment Design*, in welches er die Validierung von Testwertinterpretationen in den Entwicklungsprozess des Tests einbindet. Basierend auf dem Argumentationsschema von Toulmin (1958) definierte Mislevy für die Validierung von Testwertinterpretationen ein *Design Argument* und ein *Use Argument*. Im Folgenden werden erst der Entwicklungsprozess und anschließend das *Design* und *Use Argument* beschrieben.

Mislevy et al. (2003) strukturierte die Entwicklung von Assessments (Assessment Design) in vier Stufen. Die erste Stufe ist die *Domänenanalyse* (Domain Analyses). Hier werden wesentliche Informationen über die Domäne organisiert. Dieser Schritt beinhaltet die Sammlung von Wissen aus unterschiedlichen Quellen, das Organisieren der Vorstellungen, Theorien und Studien, der Expertise, des Instruktionsmaterials, der Exemplare anderer Assessments etc.. Die zweite Stufe ist die *Domänenmodellierung* (Domain

Modelling). Hier werden die Informationen in Strukturen geordnet, die potentielle Annahmen organisieren. Sogenannte Leistungsparadigmen strukturieren Annahmen über die Testpersonen und Leistungsaspekte, die diese Annahmen widerspiegeln. Evidenzparadigmen strukturieren die Arten der Aussagen oder Handlungen der Testpersonen, welche Evidenz für die Leistungen darstellen. Aufgabenparadigmen strukturieren die Art der Situationen, welche die Sammlung der Evidenz ermöglichen. In dieser Stufe soll also eine Verbindung zwischen Eigenschaften der Testpersonen beziehungsweise deren Aussagen oder Handlungen sowie den Aufgaben und realen Situationen, in welchen die Testpersonen handeln, geschaffen werden. Somit wird in dieer Stufe die Struktur des Assessments skizziert. Die dritte Stufe ist das *konzeptionelle Rahmenkonzept des Assessment* (Conceptual Assessment Frameword, CAF). In dieser Stufe wird der Entwurf für die operationalen Elemente des Assessment expliziert. Dies betrifft technische Details für die Implementation wie Spezifikationen, operationale Voraussetzungen, statistische Modelle, Details für Bewertungsschemata etc.. Die letzte Stufe ist die *Operationalisierung des Assessment*. Auf dieser Stufe wird der Stimulus präsentiert, eine Interaktion geschieht und die Reaktion wird festgehalten. Diese Reaktion wird anschließend evaluiert, die Informationen werden zusammengefasst und anhand von Grundannahmen werden das Wissen und das Können der Testperson interpretiert. Anhand der Ergebnisse wird dann eine Entscheidung über zukünftige Handlungen getroffen. Dieser Prozess kann in einer Vielzahl von Formen stattfinden und sehr unterschiedlich organisiert sein.

Die Stufe der *Domänenmodellierung* ist für die Validierung besonders wichtig. Hier wird das sogenannte *Assessment Argument* gebildet und es werden wichtige Stränge des Validitätsargumentes expliziert. Bei der Bildung des Arguments stützt auch Mislevy sich auf das informelle Argumentieren von Toulmin (1958). Allerdings ist das Assessment Argument komplexer. Angelehnt an (Bachman, 2005) besteht das Assessment Argument von Mislevy aus dem *Design Argument*, welches wiederum die Basis für das *Use Argument* bildet (siehe Abbildung 6). Die Grundlage des *Design Argument* bilden die Handlungen der Testperson. Diese werden in Bezug auf die Leistung oder die Testsituation interpretiert. Die Interpretationen beruhen auf Annahmen über die Bewertung beziehungsweise das Aufgabendesign, wobei auch zusätzliche Informationen über die Beziehung der Testperson mit der Assessmentsituation berücksichtigt werden können. Die auf diese Weise interpretierten Daten sind wiederum die Grundlage für Schlussfolgerungen in Bezug auf die Testperson, die durch Annahmen hinsichtlich des Assesment gestützt werden. Diese Schlussfolgerungen sind gleichzeitig die Grundlage für das *Use Argument*. Zusammen

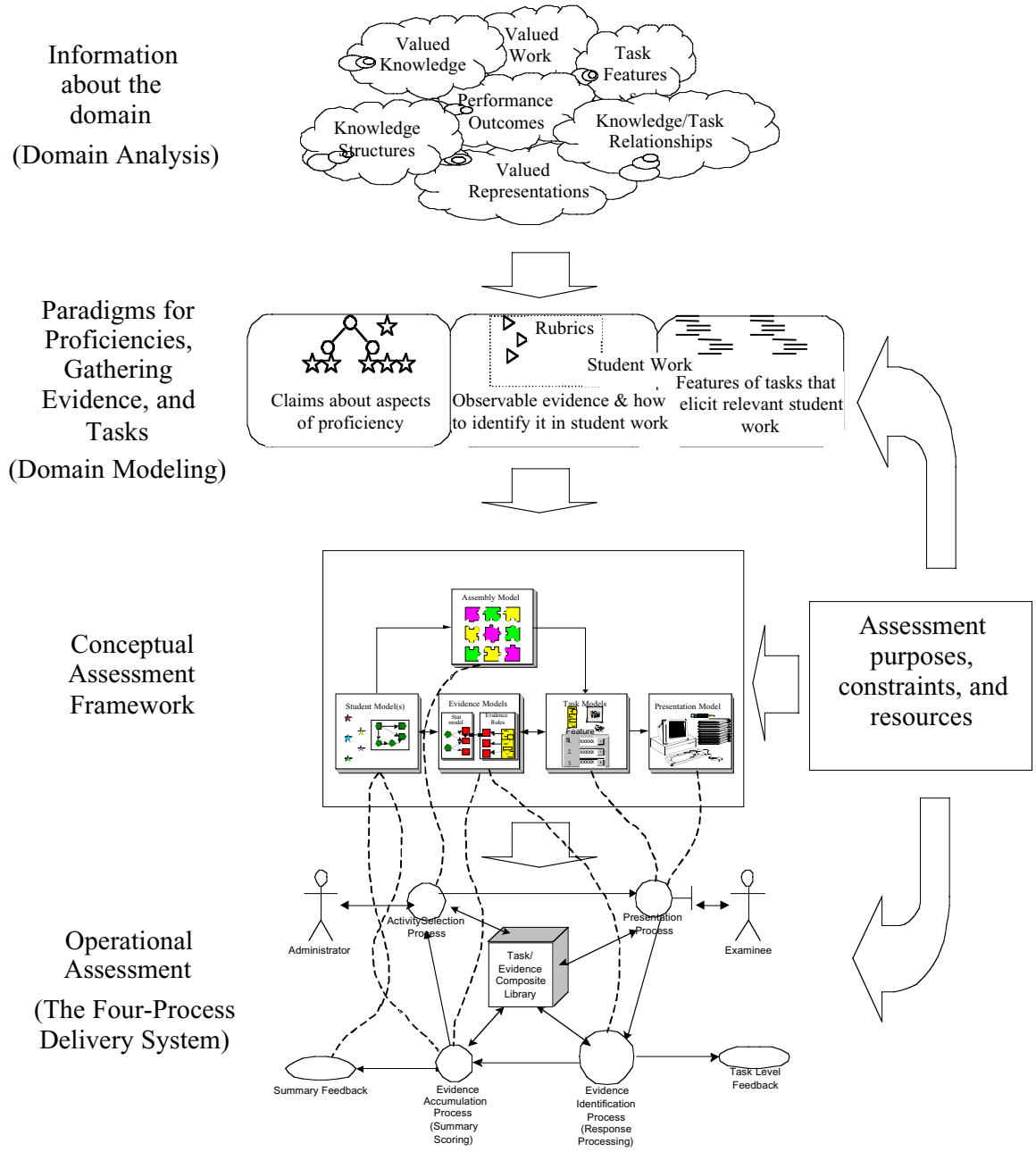


Abbildung 5: Stufen des Assessment Design nach Mislevy et al. (2003, S.5)

mit anderen Informationen hinsichtlich der Beziehung der Testperson mit der beabsichtigten Nutzung werden Interpretationen über die Testperson in der beabsichtigten

Nutzungssituation getätigt. Die Interpretationen werden auf Annahmen bezüglich dieser Situation basiert. Alle Annahmen und Interpretationen werden durch Analysen gestützt und alternative Erklärungen können in das *Assessment Argument* eingebaut werden.

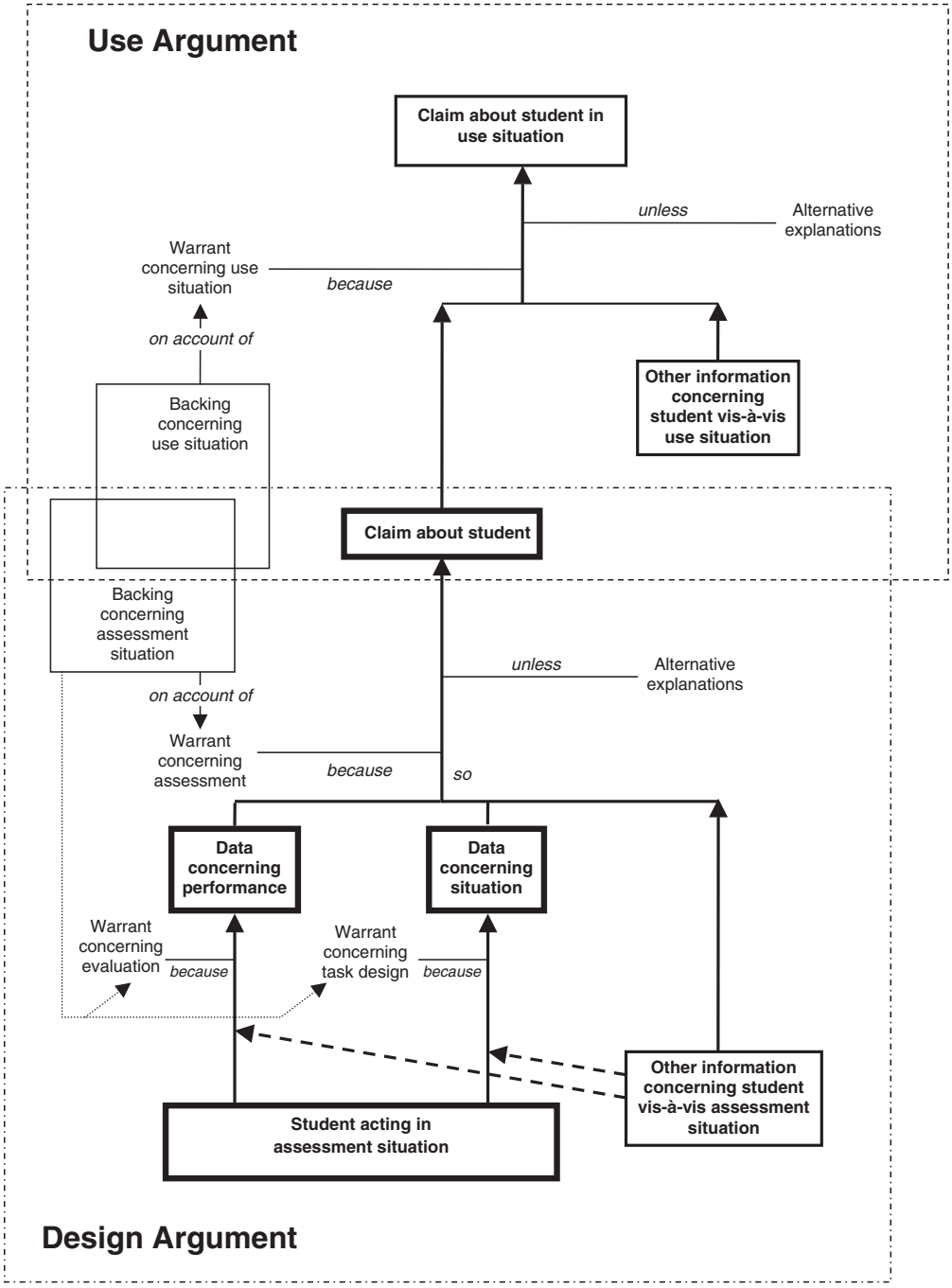


Abbildung 6: Assessment Design and Use Argument nach Mislevy (2007, S.465)

Assessment Use Argument

Auch der Ansatz von Bachman (2005) ist argumentationsbasiert und orientiert sich an dem Schema für informelles Argumentieren von Toulmin (1958). Das sogenannte *Assessment Use Argument* besteht aus einem *validity argument* und einem *utilization argument*. Das *validity argument* verbindet die Testleistung mit einer Interpretation und das *utilization argument* verbindet die Interpretation mit einer Entscheidung. Dabei benutzte Bachman den Begriff Validierung für das Sammeln von Evidenzen, welche das gesamte *Assessment Use Argument* stützen (siehe Abbildung 7). Bachman und Palmer (1996) definierten sechs Qualitäten, die einen adäquaten Test ausmachen. Die Konstruktvalidität (1) ist das Ausmaß, in dem die Testergebnisse als Indikator für zu messende Fähigkeit interpretiert werden können. Die Authentizität (2) ist der Grad der Übereinstimmung zwischen den Testeigenschaften und den Merkmalen der Zieldomäne. Die Interaktivität (3) ist das Ausmaß und die Art der Einbeziehung der individuellen Eigenschaften der Testperson in die Bearbeitung der Testaufgaben. Das Kriterium Einfluss (4) beinhaltet die Konsequenzen, welche die Testdurchführung für die individuellen Testpersonen aber auch für die Gesellschaft beziehungsweise das Bildungssystem haben. Die Reliabilität (5) ist der Grad, in dem die Testwerte über unterschiedliche Bedingungen der Testsituationen konsistent sind. Die Praktikabilität (6) ist das Ausmaß, in welchem die Voraussetzungen der Testspezifikationen mit den vorhandenen Mitteln erfüllt werden können. Das *validity argument* beinhaltet die Auswertung der Kriterien Zweckmäßigkeit, Reliabilität, Authentizität, Interaktivität und Konstruktvalidität (Bachman, 2005). Die Qualitätskriterien Einfluss, Authentizität und Interaktivität beinhalten ebenfalls Elemente, die das *utilization argument* stützen (Bachman, 2005). Das *utilization argument* bezieht sich vor allem auf die Auswertung von vier Aspekten. Diese Aspekte betreffen (1) die Relevanz bezüglich der Testwertinterpretationen für die zu treffende Entscheidung, (2) die Nützlichkeit der Testwertinterpretationen für die zu treffende Entscheidung, (3) den Vorteil der Konsequenzen, welche die Testnutzung und die Entscheidung für Individuen, die Institution, die Gesellschaft usw. haben und (4) die Angemessenheit der Informationen durch die Testung für das Treffen der Entscheidungen (Bachman & Palmer, 1996).

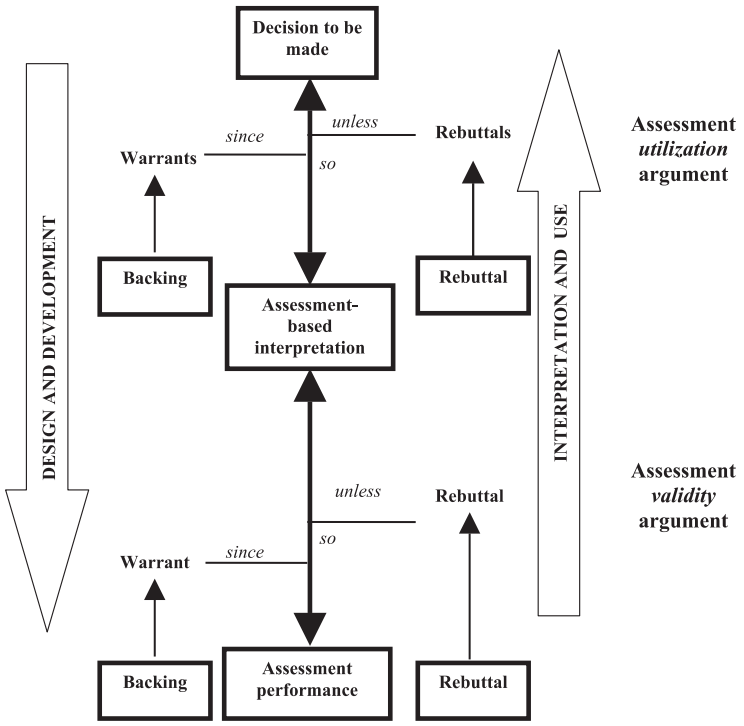


Abbildung 7: Assessment Use Argument nach Bachman (2005, S.25)

Argument Based Approach

Der Argument Based Approach von Kane (2013) spiegelt die generellen Prinzipien der Konstruktvalidität wider, ohne jedoch die soliden Konstrukt-Theorien der Konstruktvalidität unter anderem in den Standards zu erfordern. Diese Rolle übernimmt im Argument Based Approach das sogenannte Interpretation/Use Argument (IUA), welches die geplanten Interpretationen und Verwendungen der Testergebnisse spezifiziert. Anders als die argumentationsbasierten Ansätze der Standards aus (1999), von Mislevy et al. (2003) und von Bachman (2005) spezifiziert der Argument Based Approach die einzelnen Argumente, welche die beobachtete Testleistung mit der beabsichtigten Interpretation der Testwerte verbindet, in Form einer transparenten Argumentationskette. Laut Kane (2013) werden bei der Benutzung eines Tests mehrere Interpretationen getätigt. In der Regel sind nämlich die beobachteten Leistungen im Test an sich nicht von Interesse, sondern Schätzer für bestimmte Leistungen oder die Basis für einen Beschluss. So soll mit

einem Mathematiktest üblicherweise die mathematische Kompetenz der Testpersonen gemessen werden und die Antworten auf die einzelnen Aufgaben sind an sich nicht von Belang. Eine Besonderheit des Argument Based Approach nach Kane (2013) ist, dass die einzelnen Schritte, welche für das Schließen von den einzelnen Lösungen der Aufgaben auf einen Schätzer für mathematische Kompetenz nötig sind, transparent aufgegliedert und ausgewertet werden. Kane (2013) beschrieb vier dieser Schritte, welche er Schlussfolgerungen (inferences) nannte und welche in den meisten Assessments durchlaufen werden. Die erste Schlussfolgerung *Bewertung* (engl. Scoring) geschieht bei der Ergebnisermittlung. Die jeweiligen Antworten auf die Testaufgaben (Rohwerte) werden bewertet und in ein beobachtetes Testergebnis umgesetzt. Hierbei wird davon ausgegangen, dass die Bewertungsprozeduren angemessen sind, den Regeln entsprechend angewendet wurden und frei von offenkundigen Fehlern sind. Die zweite Schlussfolgerung bezieht sich auf die *Generalisierung* (engl. generalization) der Ergebnisse aus einer Testversion (mit bestimmten Aufgaben, an einem bestimmten Tag, zu einer bestimmten Zeit) auf andere, vergleichbare Testsituationen. Es wird in der Regel erwartet, dass die Schülerinnen und Schüler Aufgaben, die den im Test gestellten Aufgaben ähnlich sind, mit ähnlichem Erfolg lösen können und dass sie diese auch zu einem anderen Zeitpunkt, in einem anderen Raum etc. ähnlich gut lösen würden (Kane, 2013). Oft bleibt es nicht bei diesen Schlussfolgerungen und Testwerte werden nicht nur in Bezug auf die Testleistung interpretiert, sondern auch in Bezug auf weitere Leistungsdomänen wie zum Beispiel die Realsituation (Kane, 2013). In diesem Schritt, der *Extrapolation* (engl. extrapolation), wird von dem Testergebnis der Schülerinnen und Schüler auf ihre Kompetenz in der Zieldomäne geschlossen. Bei dieser Schlussfolgerung wird erwartet, dass das Testdesign (Aufgabenerstellung und –auswahl, Testadministration und –bewertung etc.) die Zieldomäne abdeckt. Diese drei Schlussfolgerungen ermöglichen eine deskriptive Interpretation der Testergebnisse. Die Einbeziehung von Testverwendungen und Beschlüssen auf Basis der Testergebnisse führt zu einer entscheidungsbasierten Interpretation (Kane, 2002). Bei der Schlussfolgerung *Entscheidung* (engl. decision) werden die Testergebnisse für das Treffen von Entscheidungen anhand von Entscheidungsregeln verwendet (Kane, 2013). Aus dem Schätzer für mathematische Kompetenz in einem *High Stake* Test kann beispielsweise abgeleitet werden, ob die Schülerinnen und Schüler genügend mathematische Kompetenzen für ein bestimmtes Studium besitzen. Alle Schlussfolgerungen zusammen bilden das IUA. Das IUA ist somit eine Argumentationskette, die das beobachtete Verhalten in der Testsituation mit den beabsichtigten Interpretationen der Testwerte verbindet.

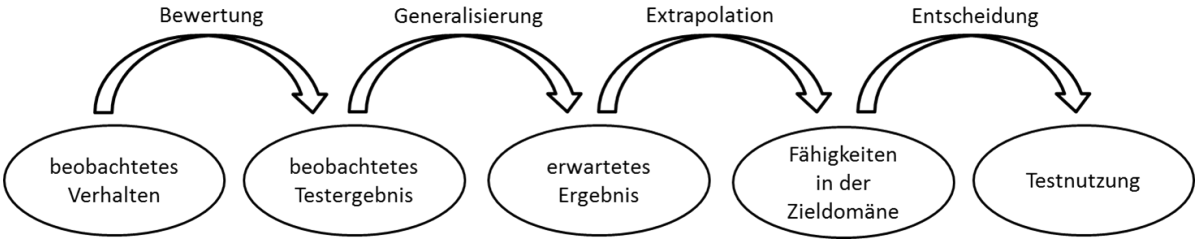


Abbildung 8: Argumentationskette des IUA nach Kane (2013)

Die vier von Kane beschriebenen Schlussfolgerungen *Bewertung*, *Generalisierung*, *Extrapolation* und *Entscheidung* sind Beispiele für mögliche Arten von Schlussfolgerungen und keinesfalls als Mustervorlage oder erschöpfende Auflistung von möglichen Schlussfolgerungen zu betrachten. Kane selbst schlug einige Erweiterungen für Argumentationsketten vor. Eine mögliche Erweiterung ist beispielsweise der *Konstruktbezug*. Ein IUA kann laut Kane zwar ohne diese Schlussfolgerung entwickelt werden, wenn jedoch ein Konstrukt für die Interpretation der Testergebnisse verwendet wird, sollte dieses auch spezifiziert werden (Kane, 2002). Andere Schlussfolgerungen, die laut Kane gebräuchlich sind, beziehen sich auf die *Skalierung* und die dazugehörigen Modelle, welche kriterienbezogene und normbezogene Interpretationen ermöglichen, und auf Hypothesen bezüglich bestimmter Merkmale. Ebenfalls ist es möglich die *Domänenbeschreibung* mit in die Argumentationskette aufzunehmen. Kane definierte diese zwar nicht als Schlussfolgerung, stellte aber die Wichtigkeit einer sorgfältigen Domänenbeschreibung und die Entwicklung von repräsentativen Aufgaben als Stützung für die Interpretation heraus (2004). Chapelle (2012) nutzt diese Möglichkeiten der Erweiterung und beschrieben die mögliche Anwendung bei der Validierung des TOEFL-Tests. Die meisten der Schlussfolgerungen in IUAs sind präsumtiv, können also eine Annahme zugunsten einer Schlussfolgerung herleiten, diese aber nicht definitiv beweisen. Für das Analysieren von Schlussfolgerungen bezog sich Kane (2013) auf das Werk „The uses of argument“ von Toulmin (1958, siehe Kapitel 1.1.3). Nach Kane (2013) können die Schlussfolgerungen der Argumentationskette des IUA mit Hilfe des Argumentationsschemas von Toulmin (1958) logisch evaluiert werden (siehe Abbildung 9). Eine solche Evaluation des IUAs nannte Kane (2013) „Validitätsargument“ (validity argument). Kane fügte jedoch zusätzlich *Annahmen* in das Argumentationsschema des IUA ein, auf dem die Argumente basieren. Diese Annahmen werden dann im Validierungsprozess durch Evidenzen gestützt. Auf das Beispiel aus Abschnitt 1.1.3 angewandt schließt man wie in Toulmins Schema von den Testergebnissen (Grundlagen) darauf, dass eine Testperson eine bestimmte Aufgabe lösen kann (Schluss-

folgerung). Das Argument dafür bildet die Passung des IRT-Modells. Annahmen für dieses Argument wären beispielsweise, dass die Aufgaben lokal stochastisch unabhängig sind, die Aufgaben rasch-homogen sind und dass die beobachtete Antwortwahrscheinlichkeit nicht von der vorhergesagten Antwortwahrscheinlichkeit abweicht. Die Stützung dieser Annahmen wiederum basiert auf der Auswertung der Annahmen. Eine Ausnahme kann ein differentielles Funktionieren der Aufgabe bezüglich des Geschlechts darstellen. Das IUA nach Kane enthält für jede Schlussfolgerung der Argumentationskette eine Grundlage, Argumente und Annahmen (Abbildung 9). Die Stützung und Ausnahmen werden durch die Evaluation des IUA im Validitätsargument gebildet. Abbildung 10 stellt das IUA und das Validitätsargument graphisch dar.

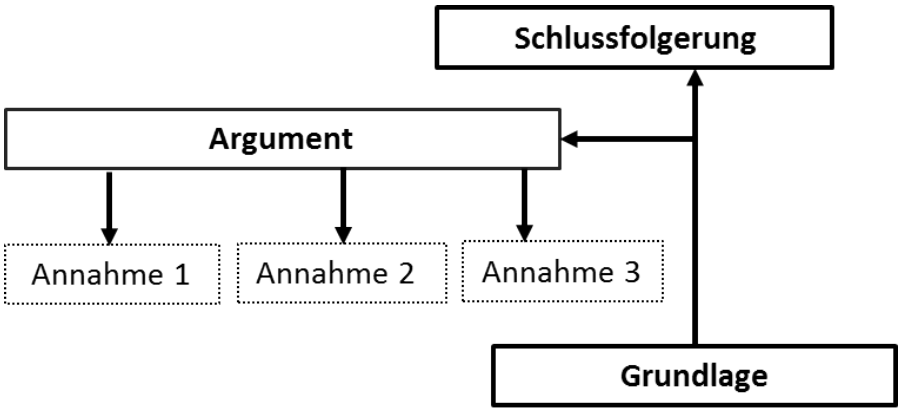


Abbildung 9: Argumentationsmodell nach Kane (2013)

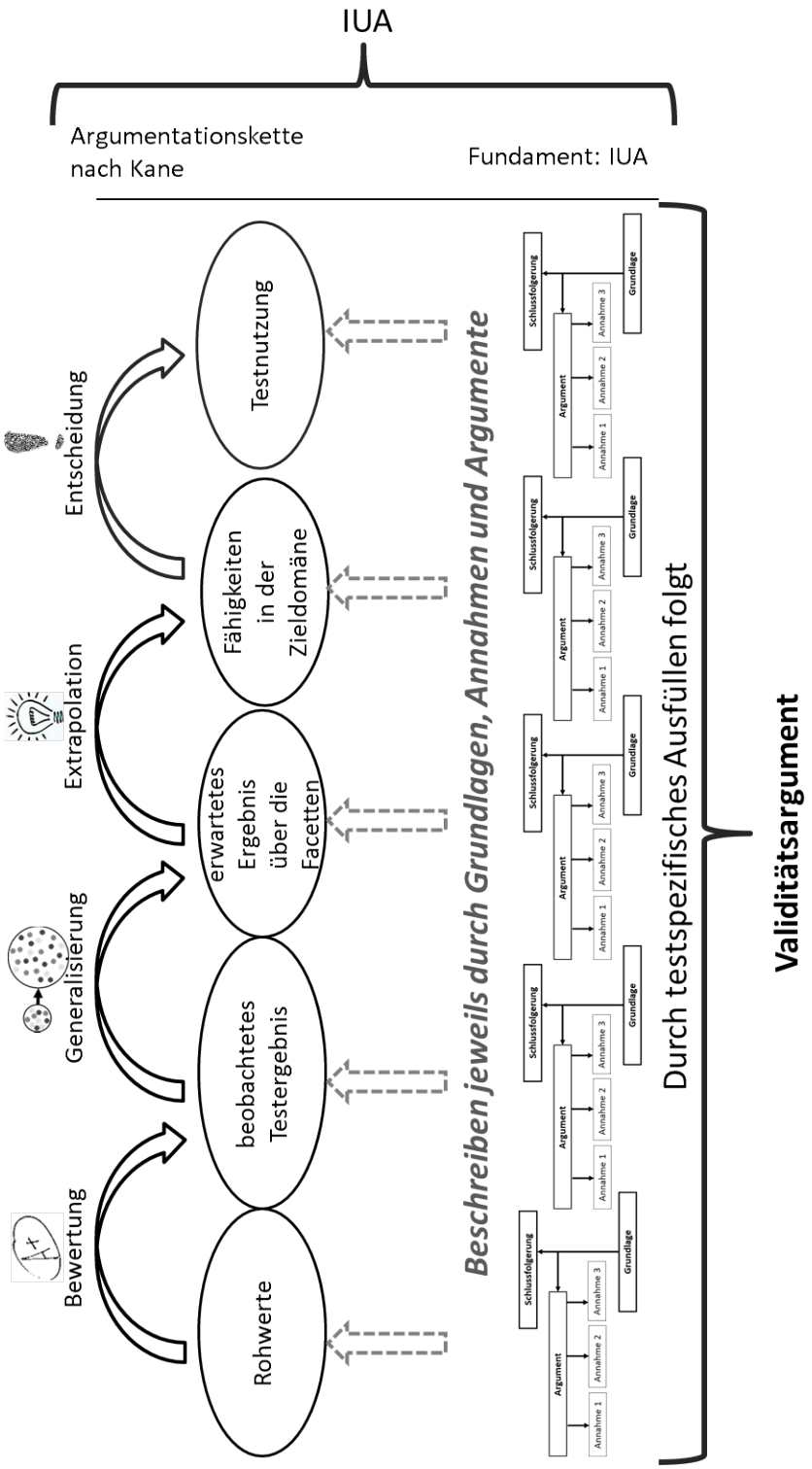


Abbildung 10: IUA und Validitätsargument nach Kane (2013)

Der erste Schritt zur Bildung des Validitätsargumentes ist die konzeptuelle Analyse des IUA. Diese sollte schlüssig sein und alle essentiellen Argumente, Annahmen und Schlussfolgerungen für die beabsichtigte Testwertinterpretation beinhalten. Das Validitätsargument evaluiert anschließend die Argumente, ebenso wie die Annahmen, auf welchen diese beruhen (Kane, 2013). Einige Annahmen können a priori akzeptiert werden, einige basieren auf Analysen und Prozeduren und einige können auf Grund von Erfahrungen akzeptiert werden. Die bedenklichsten Annahmen sollten im Validitätsargument die meiste Aufmerksamkeit bekommen. Unterschiedliche Argumente benötigen unterschiedliche Arten der Stützung (Kane, 2013). Bei der Validierung warnte Kane vor fünf möglichen Fehlern. Der „begging-the-question“ Fehler kann auftreten, wenn eine entscheidende Schlussfolgerung oder Annahme unterstellt wird (Walton, 1989). Beispielsweise wird bei Reliabilitätsstudien das Universum, über welches generalisiert werden soll, oft zu eng definiert. Die Items werden dann zwar in die Untersuchung eingeschlossen, aber unterschiedliche Itemformate, Testanlässe und Testkontexte nicht (Cronbach, Gleser, Nanda & Rajaratnam, 1972). Entgegengesetzt wirkt der „straw man“ Fehler. Dabei wird ein anspruchsvolleres IUA gewählt, als für die Interpretation und Nutzung notwendig ist. Ein Beispiel dafür ist das bieten von prädiktiver Evidenz für eine Lizenzprüfung (Kane, 2013). Wenn von einer observierten Gesetzmäßigkeit auf die Existenz eines Grundes für diese geschlossen wird, spricht man von einem „reification“ Fehler (H. V. Hansen & Pinto, 1995). Beispielweise scheint es logisch, eine Regelmäßigkeit in der Leistung über verschiedenen Aufgaben mit einem Merkmal zu assoziieren (beispielsweise analytische Fähigkeiten), welches die Leistungsmuster erklärt. Es ist jedoch riskant, anzunehmen, dass dieses Merkmal die Regularitäten erklärt oder verursacht, ohne dies zu beweisen oder zusätzliche Annahmen für dieses Merkmal zu spezifizieren (Kane, 2013). Ein weiterer Fehler ist der „gilding the lily“ Fehler. Hierbei werden zusätzliche Beweise für Annahmen angehäuft, die bereits angemessen bewiesen wurden. Ein Beispiel hierfür wäre die Bereitstellung mehrerer Reliabilitätsbeweise, wobei eine Art Beweis schon ausreichend gewesen wäre. Die Anhäufung solcher Beweise ist an sich zwar nicht schädlich, wird jedoch zu einem Fehler, wenn die Menge an Beweisen in einem Teil des IUAs fehlende Beweise in einem anderen Teil verdeckt (Kane, 2013). Der Fehler der „statistical surrogation“ beinhaltet den Ersatz eines statistischen Ansatzes für ein anspruchsvolleres Konzept (Scriven, 1987). Ein Beispiel hierfür ist die Annahme kausaler Inferenzen aufgrund von Korrelationen. Werden die Fehlerquellen bei der Validierung berücksichtigt, so bietet der Argument Based Approach einen logischen und nachvollziehbaren Rahmen für die Validierung von Testinterpretationen.

1.1.4 Differenzen bezüglich der Validität

Die von den Standards sowie von Kane (2013), Mislevy et al. (2003) und Bachman (2005) angenommene Konstruktvalidität sowie die Berücksichtigung von Testinterpretationen und Testkonsequenzen bei der Validierung spiegeln einen weit verbreiteten professionellen Konsens wider (Shaw & Crisp, 2012). Dennoch werden diese Kriterien fortlaufend kontrovers diskutiert und in der Praxis selten angewendet. Im Folgenden soll näher auf die Diskussionspunkte der Konstruktvalidität und die geringe Umsetzung der Konstruktvalidität in der Validierungspraxis eingegangen werden.

Differenzen bezüglich des Validitätskonzeptes

Das ganzheitliche Validitätskonzept der Konstruktvalidität und die Ansicht, dass Validität keine Eigenschaft des Testes selbst, sondern eine Eigenschaft von Schlussfolgerungen aus Testergebnissen ist, bilden die Grundlage für eine kontroverse Diskussion. Borsboom, Mellenbergh und van Heerden (2004) beispielsweise sahen keine Notwendigkeit für ein ganzheitliches Validitätskonzept, da es in ihre Augen nichts gebe, was ganzheitlich im Sinne einer Zusammenführung von Validitätsfacetten betrachtet werden könnte. Ein Test sei ihrer Meinung nach valide,

> „... if and only if (a) the attribute exists and (b) variations in the attribute causally produce variations in the outcomes of the measurement procedure.“ (Borsboom et al., 2004, S. 1061)

Bei der Validierung solle laut Borsboom et al. nicht die Theorie über eine Beziehung zwischen der gemessenen Eigenschaft und anderen Eigenschaften getestet werden, sondern eine Theorie über Antwortverhalten. Weitere Kritik kam von Lissitz und Samuelsen (2007). Die Autoren vertraten den Standpunkt, dass vor allem bei pädagogischen Assessments der Fokus nicht auf der Konstruktvalidität liegen solle. Ihrer Meinung nach seien die Testaufgaben eine operationelle Definition der Testdomäne, wodurch eine Definition von nomologischen Netzwerken nicht mehr benötigt werde. Lissitz und Samuelsen (2007) entwickelten daher einen eigenen Validitätsansatz, der sich vor allem auf die Testdefinition und -entwicklung, sowie auf die Teststabilität konzentriert. Validität sei in diesem Zusammenhang auch unabhängig von der Testnutzung.

Der zweite Diskussionspunkt bezieht sich auf die Testkonsequenzen, welche in den Standards von 1999 erstmals offiziell als Validitätsnachweis aufgeführt wurden. Gegner wie Popham (1997) argumentierten, dass Testkonsequenzen zwar untersucht werden sollten, jedoch nicht zum Validitätskonzept gehören würden. So schriebt Mehrens (1997), dass die Konsequenzen der Testnutzung keine Information bezüglich der Konstruktbedeutung oder Adäquatheit eines bestimmten Tests und daher auch keine Validitäsbeweise liefern würden. Auch Lissitz und Samuelsen (2007) unterstützten zwar die Untersuchung der Testkonsequenzen, allerdings hätten ungewollte Testkonsequenzen in den Augen der Autoren keinen relevanten Einfluss auf die Validität. Shepard (1997) hielt dagegen, dass Konsequenzen Teil des nomologischen Netzwerkes seien und daher unbedingt Gegenstand von Validitätsuntersuchungen sein sollten. Linn (1997) argumentierte weiter, dass die Konsequenzen der Testnutzung und –interpretation einen zentralen Stellenwert in der Evaluation der Testnutzung und –interpretation hätten und daher in die Domäne der Validität gehören würden. Auch Kane (2001) befürwortete die Einbeziehung der Testkonsequenzen in die Validität, jedoch handele es sich hier seiner Meinung nach um ein vermutlich noch lange umstrittenes Thema ohne einfache Lösung.

Differenzen bezüglich der Validierung

Trotz der Heterogenität in der Validitätstheorie schrieb Brennan (2013), dass sich diese in einem guten Zustand befinde. Er kritisierte jedoch die nur unzureichende Validierung in der Praxis. Dieser Mangel wurde von empirischen Studien bestätigt. So untersuchten Hogan und Agnello (2004) 696 Testbeschreibungen aus der Directory of Unpublished Mental Measures der American Psychological Association (APA), Jahrgang 7. In 45% der Fälle wurden keine Validitätsbeweise berichtet, 52% berichteten einen Validitätsbeweis und 2% berichteten zwei Arten von Validitätsbeweisen. Die Art der gefundenen Beweise erwies sich als sehr beschränkt. So enthielten 90% der gefundenen Beweise Korrelationen mit anderen Variablen. Es wurden nur wenige Faktorenanalysen und keine inhaltlichen Beweise oder Beweise bezüglich der Antwortprozesse gefunden. Auch Cizek, Rosenberg und Koons (2008) untersuchten die Art der verrichteten Validitätsbweise. Sie führten ein Review der Validitätsuntersuchungen im Mental Measurements Yearbook durch und fanden heraus, dass das moderne Validitätskonzept der Konstruktvalidität in den Validitätsuntersuchungen keineswegs die Norm darstellt. So wird Validität mit 30% zum Beispiel häufiger als eine Eigenschaft des Testes betrachtet anstatt als eine Eigen-

schaft der Testwerte, Interpretationen oder Schlussfolgerungen (25%). In den meisten Fällen (42%) konnte keine eindeutige Perspektive erkannt werden. Auch das ganzheitliche Validitätskonzept wurde in den wenigsten Fällen (3%) spezifiziert. Cizek, Bowen und Church (2010) führten eine follow up Studie durch, die den Fokus vor allem auf die Untersuchung der Konsequenzen innerhalb der Validität legte. In einem Review der Ausgaben von acht Zeitschriften zwischen 1999 und 2008 sowie einem Review der Tagungsprogramme der APA, AERA und NCME in den Jahren 2007 und 2008 fanden die Autoren keine einzige Validitätsuntersuchung, welche die Testkonsequenzen einschloss.
Bezüglich Large Scale Schulleistungsstudien können viele Analysen und Studien gefunden werden, die sich auf einzelne Validitätsaspekte beziehen. Dabei bleibt jedoch oft unklar, welche Gewichtung die in den Studien hergeleiteten Beweise auf die Validität der Testinterpretationen haben. So wurden beispielsweise für die Mathematiktests aus dem LV und PISA viele Hinweise für die Validität bezüglich der *Domänenbeschreibung*, *Bewertung*, *Skalierung* und *Generalisierung* in den Berichten und technischen Berichten veröffentlicht (OECD, 2013, 2014; Pant, Stanat, Pöhlmann & Böhme, 2013) und einige externe Studien zu Testwertinterpretationen bezüglich des *Konstruktbezugs* oder der *Extrapolation* auf die Zieldomäne durchgeführt (J. Hansen, 2010; Bertschy, Cattaneo & Wolter, 2009; Prenzel, Walter & Frey, 2007; Hartig & Frey, 2012, vgl. Kapitel 1.2.2 und 1.2.3). Eine vollständige Bewertung der Validität ist jedoch in den meisten Fällen nicht gegeben.

Die Situation bezüglich der Validierung lässt sich in Brennans Worten wie folgt beschreiben:

> „[...] validation is still quite impoverished. Relatively few testing programs give validation the attention it deserves. There is a notable lack of a clear listings of the claims made with a corresponding evaluation of each of them. An uncoordinated discussion of a subset of validity matters does not qualify as validation; nor does the mere presence of a chapter or document labeled "Validity" or "Sources of Validity Evidence" constitute validation in a meaningful sense.“ (Brennan, 2013, S.81)

Es gibt jedoch auch einige wenige Beispiele für Studien, welche für die Untersuchung der Validität der Testwertinterpretationen eine logische Argumentationskette gebildet und diese vollständig ausgewertet haben. Chapelle et al. (2010) beschrieben zum Beispiel die Validierung der Testwertinterpretationen des „Test of English as a Foreign Language“.

Sie entwickelten eine Argumentationskette mit sechs Schlussfolgerungen für den *High Stake* Test. Neben den von Kane (2006) vorgeschlagenen Argumenten der *Bewertung*, *Generalisierung*, *Extrapolation* und *Entscheidung*, fügten sie noch die Argumente *Domänenbeschreibung* und *Konstruktbezug* hinzu. Für diese Schlussfolgerungen entwickelte das Team Argumente und Annahmen, welche im Anschluss umfassend ausgewertet wurden (Chapelle, Enright & Jamieson, 2009). Eine genau Darstellung der Schlussfolgerungen und der Evaluation kann bei Chapelle et al. (2010) nachgelesen werden. Die Autoren kamen im Zuge der Validierung zu dem Schluss, dass die Anwendung des Argument Based Approach nach Kane (2006) viele Vorteile gegenüber einer Orientierung an den Standards (1999) habe. So biete der Argument Based Approach unter anderem mehr Richtlinien für den Prozess der Validierung. Des Weiteren könne das Konstrukt in die Validierung eingebunden werden, ohne dabei als Ausgangspunkt für die Validierung vorausgesetzt zu werden (Chapelle et al., 2009). Shaw und Crisp (2012) bildeten ein Validitätsargument für den *High Stake* Test „A level Physics“. Auch sie orientierten sich an dem Argument Based Approach von Kane. Das Validitätsargument für den Test beinhaltete fünf Schlussfolgerungen: *Konstruktrepräsentation*, *Bewertung*, *Generalisierung*, *Extrapolation* und *Entscheidung*. Für diese Schlussfolgerungen entwickelten die Autoren Argumente und Annahmen, die sie anschließend evaluierten. Auf diese Weise schufen sie ein vollständiges Bild über die Hinweise und Gefährdungen der Validität der beabsichtigten Testwerteinterpretation. Für eine genaue Beschreibung der konkreten Schlussfolgerungen und ihrer Evaluation wird an dieser Stelle auf die Literatur von Shaw und Crisp (2012) verwiesen. Das Ziel der Studie von Shaw und Crisp (2012) war nicht nur die Validierung der Testwertinterpretationen, sondern auch das Bieten von Orientierung für zukünftige Validierungsstudien.

Eine sorgfältige Validierung ist für *High Stake* Tests aufgrund der Testkonsequenzen für die Testpersonen zwar besonders wichtig, jedoch auch für die Rechtfertigung von Testwertinterpretationen von *Low Stake* Schulleistungsstudien und Kompetenztests sehr relevant. Aus diesem Grund soll diese Arbeit ein praxisbezogenes Beispiel für die umfassende Validierung eines *Low Stake* Schulleistungstests anhand des ganzheitlichen Ansatzes der Konstruktvalidität erzeugen und die Vorgehensweise der Validierung reflektieren. Für den Prozess der Validierung scheint der Argument Based Approach von Kane (2013) besonders geeignet, da dieser Ansatz zum einen deutlichere Richtlinien für den Aufbau einer logischen und transparenten Argumentationskette bereitstellt als die übrigen

in diesem Kapitel beschriebenen Validierungsansätze. Zum anderen bietet das Modell trotz der Richtlinien Flexibilität für das Einfügen von testspezifischen Erweiterungen und setzt kein präzise ausformuliertes Konstrukt voraus.

1.2 Mathematische Kompetenz in Large Scale Assessments

Bevor das IUA für den NEPS-K9-Mathematiktest nach Kane (2013) gebildet und evaluiert wird, soll im Folgenden dargestellt werden, was unter Kompetenz verstanden wird und wie die mathematische Kompetenz am Ende der Sekundarstufe zur Zeit durch nationale und internationale *Large Scale* Schulleistungsstudien in Deutschland gemessen wird. Dafür wird zuerst der Kompetenzbegriff erläutert und anschließend werden die Rahmenkonzepte des PISA- und LV-Mathematiktest aus 2012 für die neunte Klassenstufe, an welchen sich das NEPS-K9-Mathematikrahmenkonzept orientiert, beschrieben. Außerdem werden bereits dargelegte Hinweise für die Validität der Testwertinterpretationen dieser beiden Studien zusammengetragen. Letzteres soll einen Einblick in die verfügbare Validitätsevidenz schaffen, jedoch keinesfalls ein Validitätsargument für diese Studien bilden oder eine Einschätzung des Grades der Validität der beabsichtigten Testwertinterpretationen dieser Studien darstellen. Anschließend wird das mathematische Rahmenkonzept des Nationalen Bildungspanels (NEPS) beschrieben und den Konzepten aus PISA und dem LV gegenübergestellt.

1.2.1 Der Kompetenzbegriff

Der Begriff Kompetenz spielt in der Bildungswissenschaft eine große Rolle. Bildungsziele, welche in Bildungssystemen erreicht werden sollen, werden in Form von Kompetenzen charakterisiert (Hartig & Klieme, 2006). Nationale und internationale Schulleistungsstudien wie der LV und PISA messen das Erreichen dieser Kompetenzen im Rahmen des Bildungsmonitoring (Prenzel, Sälzer, Klieme & Köller, 2013; Pant, Stanat, Pöhlmann & Böhme, 2013). In der Literatur lässt sich eine Vielzahl unterschiedlicher Kompetenzdefinitionen finden (Blömeke, Gustafsson & Shavelson, 2015). Weinert unterschied beispielsweise sieben unterschiedliche Kompetenzdefinitionen (F. E. Weinert,

2001). Die erste Definition bezieht sich auf (1) allgemeine kognitive Fähigkeiten und Fertigkeiten. Diese beinhalten alle geistigen Ressourcen eines Individuums, welche für das Meistern anspruchsvoller Aufgaben in unterschiedlichen inhaltlichen Bereichen, für das Erlangen des nötigen deklarativen und prozeduralen Wissens und für das Erreichen guter Leistungen benötigt werden. Die zweite Definition geht von (2) fachbezogenen, kognitiven Kompetenzen aus. Die Kompetenzen beziehen sich auf Klassen kognitiver Voraussetzungen, die für die Leistung in einem bestimmten Kontext notwendig sind. Das Kompetenz – Leistung Modell entwickelte sich aus dieser Definition und geht davon aus, dass die Beziehung zwischen Kompetenz und Leistung durch andere Variablen, wie beispielsweise Gedächtniskapazität, Bekanntheit mit der Aufgabe, etc. moderiert wird. In der Entwicklungspsychologie wurde Kompetenz in drei Komponenten aufgeteilt (konzeptuelle, prozedurale und leistungsbezogene Kompetenz) (F. E. Weinert, 2001). Eine weitere Definition bezieht sich auf die (3) kognitiven Kompetenzen und motivationalen Handlungsorientierungen. Nach dieser Definition ist Kompetenz ein motivationales Konzept, in welchem das Selbstkonzept, Erfolgsmotive und persönliche Kontrolleinstellungen eine Rolle spielen. Kompetenzen können außerdem in (4) objektive und subjektive Kompetenzen differenziert werden. Objektive Kompetenzen beziehen sich dabei auf Leistungen und Leistungsdispositionen, die mit standardisierten Skalen und Tests gemessen werden können. Die subjektiven Kompetenzen beziehen sich auf subjektive Messungen von leistungsrelevanten Fähigkeiten und Fertigkeiten, welche für das Bewältigen von Aufgaben und Lösen von Problemen benötigt werden. Dieses Konzept kann außerdem weiter in heuristische Kompetenzen, epistemologische Kompetenzen und verwirklichte Kompetenzen differenziert werden. Die fünfte Definition bezieht sich auf (5) Handlungskompetenzen. Diese beinhalten alle kognitiven, sozialen und motivationalen Voraussetzungen, die für erfolgreiches Lernen und Handeln nötig und / oder vorhanden sind. Außerdem werden sogenannte (6) Schlüsselkompetenzen definiert. Diese schließen multifunktionale und transdisziplinäre Kompetenzen ein, die für das Erreichen von vielen wichtigen Zielen, das Meistern von unterschiedlichen Aufgaben und das Handeln in unbekannten Situationen genutzt werden. Die letzte Definition nimmt Bezug auf (7) Metakompetenzen, die die Fähigkeit beinhalten, die Verfügbarkeit, den Nutzen und die Erlernbarkeit von personengebundenen Kompetenzen einzuschätzen und anzuwenden. Aufgrund der Vielfalt an Kompetenzdefinitionen plädierte F. E. Weinert (2001) für eine Eingrenzung des Kompetenzbegriffs und er formulierte fünf Schlussfolgerungen für eine pragmatische Nutzung des Kompetenzkonzeptes: Zum einen bezieht sich das Konzept nach F. E. Weinert (2001) auf notwendige Voraussetzungen, die einem Individuum oder

einer Gruppe von Individuen für das erfolgreiche Bestreiten komplexer Ansprüche zur Verfügung stehen. Die (psychologische) Struktur der Kompetenz basiert auf den logischen und psychologischen Strukturen der Anforderungen. Die notwendigen Voraussetzungen von Kompetenzen beinhalten im Kompetenzkonzept nach F. E. Weinert kognitive und motivationale, ethische, volitionale und / oder soziale Komponenten. Das Konzept impliziert ferner, dass eine hinreichende Komplexität für die Erfüllung der Anforderungen vorausgesetzt wird. Lernprozesse sind eine notwendige Bedingung für die Aneignung der Grundvoraussetzungen zur Bewältigung der komplexen Anforderungen. Die Begriffe Schlüsselkompetenz und Metakompetenz sind nach F. E. Weinert (2001) konzeptuell zu unterscheiden. Der Begriff Schlüsselkompetenz sollte nur dann verwendet werden, wenn die Kompetenz für die Bewältigung vieler unterschiedlicher und gleichermaßen wichtiger Anforderungen benötigt wird. Der Begriff Metakompetenz sollte genutzt werden, um deklaratives oder prozedurales Wissen über die eigenen Kompetenzen zu beschreiben.
Blömeke et al. (2015) dagegen unterschieden bezüglich Kompetenz grundsätzlich zwei Perspektiven. Die erste Perspektive basiert auf der Sicht von McClelland (1973), der Kompetenz als Leistung in Situationen aus dem wirklichen Leben definiert. Die zweite Perspektive hebt die dispositionelle und im besonderen die kognitive Natur von Kompetenzen hervor. Hierunter fallen generische Kompetenzen, welche oft auch als Intelligenz oder Informationsverarbeitungsfähigkeiten bezeichnet werden oder domänenspezifische Kompetenzen. Nach Blömeke et al. (2015) trete jedoch in beiden Ansätzen die Frage auf, welche Prozesse Kognition und Volition-Affekt-Motivation auf der einen Seite und Leistung auf der anderen Seite verbinden. Blömeke et al. (2015) schlugen daher vor, dass Kompetenz als Prozess bzw. ein Kontinuum betrachtet werden sollte (vgl. Abbildung 11), anstatt auf eine unproduktive, dichotome Perspektive zu bestehen. Das Kompetenzkontinuum habe nach Blömeke et al. in Bezug auf Leistungsniveaus und Entwicklungsstufen eine vertikale Komponente. In diesem Zusammenhang sollten Trait-Ansätze die Notwendigkeit erkennen, behavioristisch zu messen, und behavioristische Ansätze sollten die Rolle kognitiver, affektiver und konativer Ressourcen berücksichtigen.

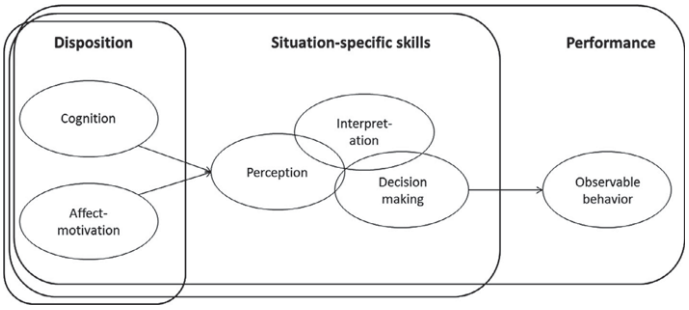


Abbildung 11: Kompetenzmodellierung als Kontinuum nach Blömeke et al. (2015, S.7)

Ebenso wie F. E. Weinert (2001) versuchten auch Blömeke et al. (2015) nicht eine allumfassende Definition für Kompetenz zu finden, sondern das „chaotische“ Konstrukt der Kompetenz zu strukturieren. In Schulleistungsstudien werden Kompetenzen in der Regel mit Hilfe von standardisierten Tests erfasst, die auf Basis von Kompetenzmodellen entwickelt werden (Hartig & Klieme, 2006). Dabei kann zwischen Kompetenzstrukturmodellen, welche die Dimensionalität der Kompetenzen beschreiben, und Niveaumodellen, welche die empirisch erfassten Kompetenzen kriteriumsorientiert beschreiben, unterschieden werden. Im Folgenden werden die Rahmenkonzepte und die darin beschriebenen mathematischen Kompetenzstrukturmodelle der in Deutschland am Ende der Sekundarstufe I eingesetzten Schulleistungstests PISA und LV beschrieben und es werden Validitätsevidenzen bezüglich dieser Studien zusammengetragen. Außerdem wird das Rahmenkonzept des in dieser Arbeit zu untersuchenden NEPS-K9-Mathemathiktests vorgestellt und den anderen beiden Studien gegenübergestellt.

1.2.2 Mathematisches Kompetenzstrukturmodell aus PISA 2012

Das Programme for International Student Assessment (PISA) wird durch die Organisation for Economic Cooperation and Development (OECD) koordiniert. PISA erfasst weltweit Schülerleistungen in regelmäßigen Abständen von drei Jahren und vergleicht diese international. Die Leitfrage der PISA-Studie ist dabei, inwiefern Schülerinnen und Schüler in den verschiedenen Staaten auf das Erwachsenenleben, lebenslanges Lernen und auf die Anforderungen einer Teilhabe an der Gesellschaft vorbereitet werden. Das Testalter beträgt 15 Jahre, da sich Schülerinnen und Schüler dieses Alters in vielen Ländern am Ende der Pflichtschulzeit befinden. PISA unterscheidet die drei Domänen Lesen, Mathematik und Naturwissenschaften, wobei zu jedem Messzeitpunkt einer der Domänen alternierend tiefergehend getestet wird. In der jeweiligen Hauptdomäne werden umfassendere Tests durchgeführt und die Testkonzeptionen und Aufgabenpools werden überprüft und gegebenenfalls überarbeitet. Die letzte PISA-Erhebung im Jahr 2012 hatte die Domäne Mathematik als Schwerpunkt (Sälzer & Prenzel, 2013). Die mathematische Kompetenz wird in der Rahmenkonzeption im Sinne von Mathematical Literacy (mathematische Grundbildung) verstanden. Letztere wird in der Rahmenkonzeption von PISA 2012 als wichtige Voraussetzung für das Leben als mündige Bürgerin beziehungsweise als mündiger Bürger verstanden und wie folgt definiert:

> „Mathematical literacy is an individual's capacity to formulate, employ, and interpret mathematics in a variety of contexts. It includes reasoning mathematically and using mathematical concepts, procedures, facts and tools to describe, explain and predict phenomena. It assists individuals to recognise the role that mathematics plays in the world and to make the well-founded judgments and decisions needed by constructive, engaged and reflective citizens." (OECD, 2013, S. 25)

Das auf der mathematischen Grundbildung basierende Kompetenzstrukturmodell hat vier Elemente als Grundlage: Inhaltsbereiche, Prozesse, fundamentale mathematische Fähigkeiten und Kontexte. Diese Elemente werden im Folgenden näher beschrieben. Für eine einheitliche Bezeichnung aller Teilbereiche in dieser Dissertation werden die Prozesse im Folgenden als Anforderungsbereiche beschrieben und die fundamentalen mathematischen Fähigkeiten als kognitive Prozesse.

Für die mathematische Kompetenz in PISA lassen sich vier Inhaltsbereiche definieren. Der Inhaltsbereich „Veränderung und Beziehungen" beinhaltet die Beziehungen zwischen Variablen und das Verständnis, wie diese repräsentiert werden können (Gleichungen eingeschlossen). Der Inhaltsbereich „Raum und Form" schließt Geometrie, Raum und Eigenschaften von Objekten ein. Beziehungen und Muster, die Zahlen beinhalten, fallen unter den Bereich „Quantität". Der letzte Inhaltsbereich „Unsicherheit und Daten" umfasst Wahrscheinlichkeiten und Statistik. Jede Aufgabe lässt sich in einen Inhaltsbereich einordnen (OECD, 2013, S. 24-37). Im mathematischen Kompetenzstrukturmodell von PISA wird kein separater Inhaltsbereich „Messen" definiert. Diese Inhalte sind Bestandteil verschiedener Inhaltsbereiche, wie zum Beispiel „Raum und Form" oder bezüglich Messfehler „Unsicherheit und Daten" (Sälzer & Prenzel, 2013).

Die kognitiven Prozesse umfassen sieben Arbeitsweisen, die zum Lösen mathematischer Aufgaben benötigt werden. Bei dem ersten Prozess handelt es sich um „Kommunizieren", welcher es dem Individuum ermöglicht, Fragen, Aussagen, Aufgaben oder Objekte zu lesen, zu entschlüsseln und zu interpretieren sowie Lösungen, Begründungen und Rechtfertigungen zu präsentieren. Für das Transformieren von Problemen in der realen Welt in eine mathematische Form sowie für das Interpretieren oder Evaluieren von mathematischen Ergebnissen oder Modellen in Relation zu dem originalen Problem wird der Prozess „Mathematisieren" benötigt. Der Prozess „Argumentieren" beinhaltet logische Denkprozesse die es einem Individuum ermöglichen, Problemelemente zu un-

tersuchen und zu verlinken, Schlussfolgerungen daraus zu ziehen und Rechtfertigungen für Problemlösungen und Aussagen zu treffen oder zu überprüfen. „Repräsentieren“ ist ein Prozess, der das Selektieren, Interpretieren und Übersetzen zwischen und Benutzen von unterschiedlichen Repräsentationen, das Interagieren mit einem Problem und das Präsentieren eigener Arbeit beinhaltet. Die Fähigkeit, einen Plan oder eine Strategie für eine Problemlösung zu selektieren oder zu entwickeln sowie diese zu implementieren, fällt unter den Prozess „Problemlösestrategien entwickeln“. „Mit Mathematik symbolisch, formal und technisch umgehen“ ist ein Prozess, der das Verstehen, Interpretieren, Manipulieren und Benutzen von symbolischen Ausdrücken sowie das Verstehen und Benutzen von formellen Konstrukten beinhaltet, die auf Definitionen, Regeln und formellen Systemen basieren, ebenso wie das Benutzen von Algorithmen mit diesen Entitäten. Das Anwenden verschiedener Hilfsmittel und die Kenntnis der Einschränkungen dieser Mittel gehört zu dem Prozess „mathematische Hilfsmittel verwenden“. Bei der Lösung einer mathematischen Aufgabe können mehrere Prozesse benötigt werden. Die Prozesse lassen sich nicht eindeutig voneinander abgrenzen und sind miteinander verbunden. So wird z.B. der Prozess „Kommunizieren“ oft gebraucht, um andere Kompetenzen sichtbar zu machen. Für die Entwicklung der Mathematikaufgaben werden die Prozesse mit den Anforderungsbereichen verknüpft. Dabei wird jeder Prozess für jeden der drei Anforderungsbereiche definiert (OECD, 2013, S. 24-37).

Die Inhaltsbereiche sind mit drei mathematischen Anforderungsbereichen verknüpft, die für das mathematische Problemlösen benötigt werden. Das Erkennen und Identifizieren von Möglichkeiten, in denen Mathematik angewendet werden kann und das anschließende Anwenden einer mathematischen Struktur beim Problemlösen in einem Kontext bilden den Anforderungsbereich „Situationen mathematisch formulieren“. Das „Anwenden mathematischer Konzepte, Fakten, Prozeduren und Schlussfolgerungen“ gehört zum zweiten Anforderungsbereich. Mittels dieses Prozesses werden mathematisch formulierte Probleme gelöst und mathematische Schlussfolgerungen gezogen. Der dritte Anforderungsbereich, das „Interpretieren, Anwenden und Bewerten von mathematischen Ergebnissen“, bezieht sich auf die Fähigkeit des Reflektierens von mathematischen Lösungen oder Resultaten und auf das Interpretieren dieser in Problemkontexten (OECD, 2013, S. 24-37).

Das Kompetenzstrukturmodell wird vor allem für die Entwicklung von Aufgaben verwendet. Für die Auswertung und den Vergleich der Kompetenzen der Jugendlichen zwischen den Staaten findet eine globale Skala Anwendung. Ein differenzierter Blick auf die

Stärken und Schwächen der Jugendlichen in den verschiedenen Staaten wird durch das Verwenden der Subskalen der Inhaltsbereiche und der Anforderungsbereiche ermöglicht. Das Kompetenzstrukturmodell wurde in PISA 2012 durch 110 Mathematikaufgaben mit offenen, halb-offenen und geschlossenen Aufgabenformaten abgebildet. Diese Aufgaben verteilten sich so über die Teilbereiche, dass sie das Kompetenzstrukturmodell repräsentierten. Aufgrund der eingeschränkten Testzeit erhielten die Schülerinnen und Schüler jeweils nur eine Teilmenge der Aufgaben. Durch die Anwendung eines Multi-Matrix-Designs konnten die Testaufgaben systematisch über die Testpersonen verteilt werden (Sälzer & Prenzel, 2013).

Hinweise auf Validität

Im technischen Bericht für PISA 2012 (OECD, 2014) werden Maßnahmen bei der Testentwicklung, -durchführung und -auswertung beschrieben, welche die Validität der PISA-Testwertinterpretationen gewährleisten sollen. Diese Maßnahmen würden nach Kane (2013) in die Schlussfolgerungen *Domänenbeschreibung*, *Bewertung* und *Generalisierung* fallen.

In der PISA-Studie wurde versucht, sicherzustellen, dass die Aufgaben die mathematische Domäne repräsentieren. Die Entwicklung der Aufgaben basierte auf dem mathematischen Rahmenkonzept, das die Domäne definiert, den Kompetenzbereich beschreibt und die Struktur des Tests, einschließlich der Aufgabenformate und der Verteilung der Aufgaben, spezifiziert. Das Rahmenkonzept wurde erstmals 2000 entwickelt und 2003 spezifiziert. Diese Entwicklung wurde sehr sorgfältig von internationalen Experten durchgeführt. Anschließend wurde das Rahmenkonzept von 170 Mathematikexperten begutachtet (OECD, 2014). Für die Erhebung im Jahr 2012 wurde das mathematische PISA-Rahmenkonzept aus 2003 revidiert. Eine Befragung von leistungsstarken OECD-Ländern und von über 80 Experten aus 34 unterschiedlichen Ländern bildete die Basis für die Revision des Rahmenkonzeptes. Durchgeführt wurde diese Befragung sowie die Entwicklungsarbeit von der Mathematics Expert Group (MEG) (OECD, 2014). Die neuen Mathematikaufgaben für PISA 2012 wurden in neun Testentwicklungszentren aus kulturell vielfältigen Institutionen entwickelt. Die Aufgabenentwicklung wurde in zwei Phasen geplant. Die Testentwicklungsteams erstellten neue Aufgaben, sowohl in ihrer Muttersprache als auch in englischer Sprache. Danach prüften sie jede neu entwickelte

Aufgabe erst intern und anschließend in *cognitive interviews*. Die Aufgaben wurden angepasst und in einer lokalen Pilotstudie getestet. Anschließend wurde jede Aufgaben-Unit von mindestens einem Testentwicklungsteam, welches diese Unit nicht entwickelt hat, begutachtet. Hiernach wurden die Units in einer Serie von internationalen Pilotstudien getestet. Die Aufgaben wurden in einem anschließenden Feldtest in allen PISA-Ländern getestet und in einem letzten nationalen Review durch die Nationalen Center aufgrund der Erfahrungen im Feldtest begutachtet. Bei der Auswahl der Aufgaben, die in der Haupterhebung verwendet werden sollten, wurde die Aufnahme von Aufgaben vermieden, die durch die Nationalen Center zuvor als nicht geeignet eingestuft worden waren (OECD, 2014). Die Übersetzung der Aufgaben wurde sowohl aus dem Englischen als auch aus dem Französischen vorgenommen. Dabei wurden deutliche Richtlinien für die Übersetzung entwickelt und ein Training für die nationalen Verantwortlichen durchgeführt. Die übersetzten Aufgaben wurden anschließend von unabhängigen Gutachterinnen und Gutachtern überprüft (OECD, 2014).
Insgesamt können also Hinweise für eine sorgfältige Aufgabenentwicklung und Übersetzung gefunden werden, welche auf für die Domäne repräsentative Aufgaben hindeuten.

Die Eingabe der Daten erfolgte über das Programm KeyQuest, welches Gültigkeitsprüfungen bei der Eingabe durchführt. Außerdem wurden die Daten in den jeweiligen nationalen Centern anhand eines Data Management Manuals standardisiert geprüft und es wurden Berichte über das Cleaning verfasst. Die Bewertung der Aufgaben wurde bereits bei der Aufgabenentwicklung entworfen und im Entwicklungsprozess und den Feldtests angepasst und optimiert. Für die Haupterhebung im Jahr 2012 wurden Kodierrichtlinien für alle manuell zu bewertenden Aufgaben in englischer und französischer Sprache vorbereitet. Die Materialien für das Kodiertraining wurden nach dem Feldtest noch einmal überarbeitet. Aufgaben, die im Feldtest Kodierprobleme verursachten, wurden nicht in die Haupterhebung aufgenommen. Alle Kodiererinnen und Kodierer der Haupterhebung mussten an einem Kodiertraining teilnehmen. Außerdem wurden Konsistenzanalysen bezüglich der Raterübereinstimmung innerhalb der Länder für Aufgaben durchgeführt, die eine Beurteilung benötigten, ebenso wie ein Kodierer Review, welches mögliche länderbezogene Fehler aufdecken sollte. Für die Mathematikaufgaben wurden keine Inkonsistenzen in den zu beurteilenden Aufgaben gefunden (OECD, 2014).
Aufgaben, die im Feldtest schlechte psychometrischen Itemeigenschaften zeigten, wurden nicht für die Haupterhebung ausgewählt. In der Haupterhebung wurden die psychometrische Itemeigenschaften der Aufgaben erneut berechnet und für jedes Land analysiert.

Problematische Aufgaben wurden aus den gesamten oder länderspezifischen Analysen entfernt.
Zusammenfassend können Evidenzen für eine hohe Standardisierung der Eingabe und Bewertung der Aufgaben sowie für die psychometrische Qualität der Aufgaben gefunden werden.

Für die Skalierung der Daten wurde das mixed coefficients multinomial logit model (Adams, Wilson & Wang, 1997) verwendet. Bei der Auswertung der Modellpassung wurden für jedes teilnehmende Land die MNSQ-Werte für jeweils alle Aufgaben berechnet. Unpassende Aufgaben wurden für die gesamte Stichprobe oder für bestimmte Länder ausgeschlossen (OECD, 2014). Das Testmanual enthielt präzise und deutliche Anweisungen zur Testdurchführung und zur Testumgebung, die von den Testleiterinnen und Testleitern einzuhalten waren sowie mögliche Anpassungen der Testprozesses. Instruktionen waren vorgegeben und mussten von den Testleiterinnen und Testleitern vorgelesen werden. Einige Testungen wurden von externen und geschulten Gutachterinnen und Gutachtern beobachtet. Alle Abweichungen von dem standardisierten Vorgehen wurden im technischen Bericht beschrieben (OECD, 2014).
Für die Schätzung der Personenfähigkeiten wurden PVs verwendet, da diese konsistente Schätzungen auf Gruppenebene erlauben. Ein Hintergrundmodell sollte die Genauigkeit der Schätzungen noch steigern. Für die Generalisierung der Daten auf die Population wurden Gewichte berechnet. Außerdem wurden Testhefteffekte bei den Brechungen berücksichtigt. Die WLE-Reliabilität der Mathematikskala beträgt bei separater Skalierung 0.82 mit einem Measurement Error Design Effect von 1.22. Die PV-Reliabilität ist mit 0.85 etwas höher und hat einen etwas niedrigeren Measurement Error Design Effect von 1.18. Die Reliabilität der internationalen Mathematikskala liegt bei 0.91, die der internationalen mathematischen Subskalen der Inhaltsbereiche und Anforderungsbereiche zwischen 0.89 und 0.91. Zusammenfassend kann geschlossen werden, dass Evidenzen für eine hohe Standardisierung der Umstände der Messung gefunden werden konnten. Die Testwerte scheinen nicht durch unterschiedliche Durchführungsbedingungen beeinflusst. Die Realiabilität des Personenschätzer und die Skalenreliabilität, sowie die Berechnung von Gewichten weisen darauf hin, dass die PISA-Ergebnisse über die konkrete Testsituation hinaus generalisiert werden können.

Außerhalb des technischen Berichtes wurden auch Studien veröffentlicht, die das Konstrukt mathematischer Kompetenz aus PISA und die Extrapolation auf die Zieldomäne untersuchten (Prenzel et al., 2007; Bertschy et al., 2009; J. Hansen, 2010). Diese Studien

bezogen sich allerdings auf ältere Versionen des PISA-Mathematiktests.
Für die Interpretation der Testergebnisse im Sinne des Konstruktes beziehungsweise der Zieldomäne wurden einige wenige Studien durchgeführt. Prenzel et al. (2007) untersuchten unter anderem die Abgrenzbarkeit der in PISA 2003 gemessenen Kompetenzbereiche von den anderen in PISA gemessenen Kompetenzen sowie von der kognitiven Fähigkeit. Auf Basis der deutschen Stichprobe von PISA 2003 (N = 4660) berechneten die Autoren ein eindimensionales Rasch-Modell, in dem die vier PISA-Tests (Mathematik, Lesen, Naturwissenschaften und Problemlösen) und die „Figuralen Analogien“ des in PISA 2003 eingesetzten KFT-Untertest in einer gemeinsamen Dimension laden, und ein fünfdimensionales Modell, in welchem die Tests in einer jeweils separaten Dimension laden. Die Ergebnisse zeigen bessere Modellgütekriterien (BIC, CAIC, Likelihooddifferenz) für das fünfdimensionale Modell (Prenzel et al., 2007).
Bertschy et al. (2009) untersuchten den Zusammenhang zwischen den aus PISA 2000 gewonnenen Ergebnissen von Schülerinnen und Schülern aus der Schweiz auf deren Übergang von Schule in den Arbeitsmarkt in der Transitions from Education to Employment (TREE) Studie. Die Basis für diese Analysen war die Stichprobe der Schülerinnen und Schüler, die im Jahr 2000 an PISA teilgenommen hatten. Diese Testpersonen wurden fünf Jahre lang längsschnittlich befragt. Dabei wurden nur die Testpersonen für die längsschnittlichen Analysen ausgewählt, die nach ihrem Schulabschluss direkt eine dreijährige Ausbildung begonnen hatten. Für 642 Testpersonen lagen längsschnittliche Daten bis 1.75 Jahre nach Abschluss der Ausbildung vor. Die Ergebnisse zeigten, dass die individuellen PISA-Ergebnisse keine direkten Effekte auf den Übergang in den Arbeitsmarkt fünf Jahre nach der PISA-Erhebung hatten. Jedoch schien es einen indirekten Effekt zu geben. Die PISA-Ergebnisse hingen mit der Art der Berufsausbildung und einem möglichen Schulabbruch zusammen. Höhere PISA-Testergebnisse waren mit einer intellektuell anspruchsvolleren Berufsausbildung verbunden. Schülerinnen und Schüler, die wiederum eine intellektuell anspruchsvollere Berufausbildung abgeschlossen hatten, hatten eine höhere Wahrscheinlichkeit, eine adäquate Arbeitsstelle zu finden.
In einer Studie von J. Hansen (2010) wurden unter anderem Aspekte der Exploration von PISA durch die Schätzung von Effekten der mit PISA gemessenen Mathematik- und Lese-Kompetenzen auf die Übergänge zwischen Schule, Ausbildung und Arbeit untersucht. Die Analysen basierten auf einer Stichprobe der längsschnittlichen, kanadischen Youth in Transition Survey (YITS). Die wichtigsten Ergebnisse dieser Studie waren, dass Schülerinnen und Schüler mit höheren Ergebnissen in den PISA Mathematik- und Lese-Tests eine signifikant längere Beschulung erreichten als Schülerinnen und Schüler mit

niedrigeren Ergebnissen. Selbst nach der Kontrolle für Charakteristika wie Einkommen, Ausbildung der Eltern etc. blieben diese Effekte bestehen. Darüber hinaus war nicht nur die Wahrscheinlichkeit, die Schule früher zu verlassen, für Schülerinnen und Schüler mit niedrigen Lese-Ergebnissen größer als die von Schülerinnen und Schülern mit höheren Lese-Ergebnissen. Auch hatten sie eine kleinere Chance, die Schulbildung später wieder aufzunehmen. J. Hansen interpretierte die Ergebnisse als Hinweise für prädiktive Validität der PISA-Mathematik- und der PISA-Leseergebnisse.
Insgesamt konnten einige Hinweise dafür gefunden werden, dass die Testergebnisse das PISA-Kompetenzkonstrukt widerspiegeln.

1.2.3 Mathematisches Kompetenzsstrukturmodell aus dem LV 2012

Die Ständige Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (KMK) erteilte den Auftrag, Bildungsstandards zu entwickeln, um eine gemeinsame Basis aller Länder für Bildungsmonitoring und Qualitätsentwicklung zu schaffen. Diese Standards wurden in den Jahren 2003 und 2004 verabschiedet. Das Institut zur Qualitätsentwicklung im Bildungswesen (IQB) wurde wiederum von der KMK beauftragt, die Einhaltung der Standards regelmäßig und flächendeckend zu überprüfen, die Standards weiterzuentwickeln und Kompetenzstufenmodelle zu entwerfen. Am Ende der Sekundarstufe I findet alle drei Jahre der LV des IQB statt, alternierend für die Bereiche Deutsch und erste Fremdsprache sowie Mathematik und Naturwissenschaften. Die letzte Erhebung im Jahr 2012 hatte den Bereich Mathematik und Naturwissenschaften zum Schwerpunkt (Pant, Stanat, Schroeders et al., 2013). Die bildungstheoretische Grundlage des Kompetenzmodells der Bildungsstandards ist der Grundbildungsauftrag des Unterrichtsfaches Mathematik, welcher durch Winter (2003) beschrieben wurde. Nach Winter sollen Schülerinnen und Schüler die folgenden drei Grunderfahrungen im Mathematikunterricht machen: (1) Erscheinungen der Umwelt mithilfe der Mathematik in spezifischer Weise wahrnehmen und verstehen, (2) mathematische Gegenstände im Sinne geistiger Schöpfungen als eigenständige Welt kennenlernen und verstehen und (3) aufgrund der Beschäftigung mit Mathematik heuristische Fähigkeiten erwerben, die über die Mathematik hinausgehen (Winter, 2003). Die Bildungsstandards im Bereich Mathematik schreiben vor, welche mathematischen Kompetenzen Schülerinnen und Schüler am Ende des Hauptschulabschlusses beziehungsweise am Ende des Mittleren Bildungs-

abschlusses erreicht haben sollen (Roppelt, Blum & Pöhlmann, 2013).

Das mathematische Kompetenzstrukturmodell basiert auf drei Elementen: Leitideen, allgemeine mathematische Kompetenzen und Anforderungsbereiche. Für eine einheitliche Benennung in dieser Dissertation werden im Nachfolgenden die Leitideen als Inhaltsbereiche und die allgemeinen mathematischen Kompetenzen als kognitive Prozesse bezeichnet. Alle Mathematikaufgaben aus dem LV lassen sich in diesen Teilbereichen verorten (Roppelt et al., 2013).

Die Inhaltsbereiche sind phänomenologisch und haben enge Beziehungen mit den Stoffgebieten der Schulmathematik. Der Inhaltsbereich „Zahl“ beinhaltet alle Aspekte, die mit der Verwendung von Zahlen zur Beschreibung und Organisation von Situationen zu tun haben. Der Umgang mit Größen, insbesondere mit Längen, Winkeln, Flächeninhalten, Volumina, Geldwerten, Zeitspannen und Massen, fällt unter den Inhaltsbereich „Messen“. Der Inhaltsbereich „Raum und Form“ beinhaltet Eigenschaften und Beziehungen aller Arten räumlicher Konfigurationen, Gestalten oder Muster. Funktionale und relationale Beziehungen zwischen mathematischen Objekten, einschließlich deren Darstellung und Eigenschaften, gehören zum Inhaltsbereich „funktionaler Zusammenhang“. Der Inhaltsbereich „Daten und Zufall“ umfasst den Umgang mit statistischen Daten, ebenso wie den Umgang mit Situationen, in denen Zufall und Wahrscheinlichkeit eine Rolle spielen. Jede Aufgabe ist jeweils einem dieser Inhaltsbereiche zugeordnet. Die unterschiedlichen Inhaltsbereiche lassen sich nicht immer eindeutig voneinander abgrenzen. Das Berechnen von Inhalten oder Umfängen geometrischer Strukturen beispielsweise erfordert häufig auch das Einbeziehen spezieller Eigenschaften, sodass Zugänge sowohl aus dem Inhaltsbereich „Messen“ als auch aus dem Inhaltsbereich „Raum und Form“ benötigt werden (Roppelt et al., 2013).

Im Rahmenkonzept werden sechs unterschiedliche Prozesse definiert, die das Spektrum des mathematischen Arbeitens erfassen. Der Prozess „mathematisches Argumentieren“ beinhaltet das Entwickeln situationsadäquater mathematischer Argumentationen und das Verstehen oder Bewerten gegebener Argumentationen. Das Finden und Anwenden geeigneter Lösungsstrategien fällt dem Prozess „Problemlösestrategien entwickeln“ zu. Wechseln zwischen außermathematischen Realsitutationen und mathematischen Begriffen, Resultaten oder Methoden wird als „mathematisch Modellieren“ definiert. „Mathematische Darstellungen verwenden“ umfasst das Auswählen oder Erzeugen mathematischer Darstellungen und den Umgang mit gegebenen Darstellungen. „Mit Mathematik

symbolisch, formal, technisch umgehen“ ist ein Prozess, der das Ausführen von Operationen mit Zahlen, Größen, Variablen und Termen oder mit geometrischen Objekten abbildet. Das Entnehmen von Informationen aus schriftlichen Texten, mündlichen Äußerungen oder sonstigen Quellen sowie das Darlegen von Überlegungen und Resultaten unter Verwendung einer angemessenen Fachsprache werden den Prozess „mathematisch Kommunizieren“ zugeschrieben. Für die Bearbeitung mathematischer Probleme werden meist mehrere Prozesse gleichzeitig benötigt. Auch die Prozesse können daher nicht klar voneinander abgetrennt werden. Für die erfolgreiche Lösung einer typischen Modellierungsaufgabe werden beispielsweise neben symbolischen, formalen und technischen Fertigkeiten oft auch Problemlösekompetenzen benötigt.

Die Prozesse können auf unterschiedlichen Anforderungsniveaus gefordert werden. Aus diesem Grund werden die drei Anforderungsniveaus „Reproduzieren“, „Zusammenhänge Herstellen“ und „Verallgemeinern und Reflektieren“ in einer ausführlicheren Darstellung für jeden Prozess einzeln beschrieben. Beim Anforderungsbereich „Reproduzieren“ handelt es sich oft um einfache Strategien und Routineverfahren. Der Anforderungsbereich „Verallgemeinern und Reflektieren“ verlangt dagegen meist komplexe Lösungsstrategien und Umgang mit unvertrautem Material.

Das hier beschriebene, komplexe Kompetenzstrukturmodell wird vor allem für die Entwicklung von Aufgaben genutzt. Für die Beschreibung der individuellen Unterschiede zwischen Schülerinnen und Schülern ist die Betrachtung einer globalen Skala oft ausreichend, da die unterschiedlichen Kompetenzen eng miteinander verbunden sind. Für differenzierte Aussagen bezüglich der Inhaltsbereiche können auch getrennte Skalen genutzt werden. Im Ländervergleich 2012 wurden 374 Aufgaben mit offenen, geschlossenen und halboffenen Aufgabenformaten eingesetzt, um das Kompetenzstrukturmodell repräsentativ abzubilden. Jede Testperson bearbeitete nur eine Auswahl dieser Aufgaben (Roppelt et al., 2013). Mit Hilfe eines Multi-Matrix-Designs wurden die Aufgaben gruppiert und über die Stichprobe verteilt (Siegle, Schroeders & Roppelt, 2013).

Hinweise auf Validität

Im Bericht des LV 2012 (Pant, Stanat, Pöhlmann & Böhme, 2013) werden Maßnahmen bezüglich der Testentwicklung, -durchführung und -auswertung beschrieben, welche die Validität der LV-Testwertinterpretationen sichern sollen. Diese Maßnahmen können nach

dem Argument Based Approach (Kane, 2013) unter die Schlussfolgerungen *Domänenbeschreibung*, *Bewertung* und *Generalisierung* fallen.

Bei der Testkonstruktion unter Federführung des IQB für die im LV untersuchten Kompetenzen wurde versucht, sicherzustellen, dass die Domänen sorgfältig beschrieben werden und dass die entwickelten Aufgaben die Domäne repräsentieren. Dafür wurde in einem ersten Schritt in Kooperation mit fachdidaktischen Expertinnen und Experten ausgearbeitet, welche Kompetenzaspekte im LV operationalisiert werden können und wie diese zu spezifizieren sind (Pant, Stanat, Pöhlmann & Böhme, 2013). Anschließend wurden Richtlinien und Trainingsmaterial zur Konstruktion von Testaufgaben entwickelt, anhand derer Lehrkräfte aller 16 Bundesländer geschult wurden. Unter fachdidaktischer Anleitung generierten diese Lehrkräfte Testaufgaben, die sie selbst in ausgewählten Klassen im Rahmen einer Präpilotierung erprobten und daraufhin erneut optimierten. In einem nächsten Schritt wurden die Aufgaben von Expertinnen und Experten aus der Bildungsforschung und den Fachdidaktiken begutachtet. Darauf basierend wurden die Aufgaben überarbeitet.
Zusammengefassen kann geschlossen werden, dass Hinweise für eine sorgfältige Aufgabenentwicklung gefunden werden konnten, die auf für die Domäne repräsentative Aufgaben hindeuten.

Die Testhefte wurden durch das Data Processing and Research Center (DPC) eingescannt. Multiple Choice Aufgaben wurden maschinell ausgewertet, offene und halboffene Aufgaben wurden von geschulten Kodiererinnen und Kodierern bewertet. Dabei stand den Kodiererinnen und Kodierern eine fachdidaktisch geprüfte und empirisch bewährte Kodieranleitung mit deutlichen Bewertungskriterien zur Verfügung. Die Auswertungsanleitung von offenen Aufgaben wurde anhand der Schülerantworten erstellt und optimiert (Pant, Stanat, Pöhlmann & Böhme, 2013; Köller, 2010). Die nachfolgende psychometrische Prüfung der Aufgaben ermöglichte eine Identifikation fehlerhaft kodierter Items. Diese konnten anschließend berichtigt werden. Die psychometrische Eignung aller Aufgaben wurde anhand der Itemschwierigkeiten und Trennschärfen analysiert. Problematische Aufgaben wurden identifiziert und überarbeitet oder entfernt.
Es konnten mit diesem Verfahren also Hinweise für eine fehlerfreie und konsistente Bewertung der Aufgaben sowie für angemessene psychometrische Aufgabeneigenschaften gefunden werden.

Die Modellpassung der Aufgaben wurde anhand des WMNSQ berechnet. Insgesamt wur-

den 38 Aufgaben, die eine schlechte Modellpassung (Weighted Fit > 1.15) aufwiesen, aus weiteren Analysen entfernt. Insgesamt konnten 192 Mathematikaufgaben und 236 Naturwissenschaften (NaWi)-Aufgaben in den Tests eingesetzt werden (Pant, Stanat, Pöhlmann & Böhme, 2013). Die Erhebung fand unter hoch standardisierten Testbedingungen statt. Die Testungen wurden durch das DPC der International Association for the Evaluation of Educational Achievement (IEA) in Hamburg organisiert und durchgeführt, welche auf die Durchführung solcher großen Schulleistungsstudien spezialisiert ist. Die Testleiterinnen und Testleiter wurden vom DPC ausgewählt und geschult, sodass sie mit dem Testmaterial vertraut gemacht wurden. Der Ablauf der Testung wurde präzise in Testskripts festgelegt. Instruktionen und Erklärungen wurden im Vorhinein festgelegt und von den Testleiterinnen und Testleitern nur vorgelesen. Bearbeitungszeiten, Störungen und andere Vorkommnisse wurden in Testleiterprotokollen festgehalten.
Das Testdesign der LV-Tests wurde so gewählt, dass Effekte der Testumstände, wie beispielsweise Ermüdungseffekte und Motivationsverlust während der Testung, die Ergebnisse nicht beeinflussten. Es wurde ein sogenanntes Youden-Square-Design eingesetzt, in welchem alle Aufgabenblöcke so verlinkt wurden, dass jeder Aufgabenblock mit allen übrigen Aufgabenblöcken in mindestens einem anderen Testheft zusammen vorkam. Außerdem wurde die Vorkommenshäufigkeit der Aufgabenblöcke konstant gehalten und jeder Block wurde an jeder Position im Testheft gleich oft eingesetzt.
Für die Skalierung der Kompetenzdaten wurde das Rasch-Modell eingesetzt. Die Itemparameter wurden unter Einbeziehung aller Testdaten pro Fach und Kompetenzbereich in eindimensionalen Modellen geschätzt. Anschließend wurden die Personenfähigkeiten unter Berücksichtigung von Hintergrundmerkmalen als Plausible Values berechnet. Dieses Vorgehen sollte gewährleisten, dass die Zusammenhänge sowohl zwischen den Individualmerkmalen und den Leistungswerten als auch den Populationsvarianzen unverzerrt abgebildet wurden. Um aus den Daten dieser Stichprobe Rückschlüsse über die Population ziehen zu können und für die Population unverzerrte Durchschnittswerte zu erhalten, wurden Schülergewichte berechnet.
Insgesamt konnten Hinweise auf hoch standardisierte Durchführungsbedingungen gefunden werden, die darauf hindeuten, dass die Testergebnisse sich nicht aufgrund von Messumständen unterscheiden. Des Weiteren werden Evidenzen für unverzerrte Schätzungen und Generalisierungen auf die Population gefunden. Angaben zur Reliabilität des Tests und zur Auswertung der Testleiterprotokolle bezüglich Störungen von Testsitzungen fehlen.

Neben den Maßnahmen, die im Bericht des LV 2012 beschrieben wurden, wurden auch externe Analysen zum Konstrukt des LV-Mathematiktests sowie zur Extrapolation der Ergebnisse auf die Zieldomäne durchgeführt. Diese Analysen wurden allerdings mit einer älteren Version des LV-Mathematiktests durchgeführt. Hartig und Frey (2012) untersuchten die Extrapolation des LV-Mathematiktests für den Mittleren Schulabschluss auf die Zieldomäne, indem sie die Zusammenhänge und die Abgrenzbarkeit der mit diesen Tests gemessenen Kompetenzen mit den in PISA 2006 gemessenen Kompetenzen analysierten. Die Autoren fanden hohe korrelative Zusammenhänge ($r = .94$) zwischen den Mathematiktests der Studien, die auf starke Gemeinsamkeiten schließen lassen. Niedrigere Korrelationen mit den PISA-Domänen Lesen und Naturwissenschaften ($r = .75$ und $r = .85$) wiesen auf die Abgrenzbarkeit der erfassten mathematischen Kompetenz von anderen Kompetenzen hin. Zusätzlich wurde der Konstruktbezug durch einen Vergleich der erklärten Varianzen im Ländervergleich und in PISA gegenüber den theoretischen Annahmen untersucht. Höhere durch Schulformen und Schulen erklärte Varianzanteile beim Ländervergleich ließen auf eine größere curriculare Nähe des Ländervergleich-Tests schließen. Dies stimme laut Hartig und Frey (2012) mit den theoretischen Erwartungen überein und könne als Hinweis für curriculare Validität gewertet werden. Zusammenfassend lässt sich festhalten, dass auch einige Evidenzen für die Interpretation der Testwerte bezüglich des Kompetenzkonstruktes und der Zieldomäne gefunden werden konnten.

1.2.4 Mathematisches Kompetenzstrukturmodell aus NEPS-K9

Das nationale Bildungspanel, auch National Educational Panel Study (NEPS) genannt, ist eine Panelstudie, die den Kompetenzerwerb und die Bildungsbiographie über die Lebensspanne untersucht. Die Schülerinnen und Schüler werden jährlich innerhalb der Schule befragt, bis diese die ursprünglich ausgewählte Schule oder das allgemeinbildende Schulsystem verlassen haben. Anschließend werden die Testpersonen außerhalb der Institution weiter befragt und getestet. Es werden Stichproben zu verschiedenen Zeiten für unterschiedliche Altersgruppen gezogen. Es handelt sich hierbei also um ein Multi-Kohorten-Sequenz-Design.

Neben dem Kompetenzbereich Mathematik werden auch die Bereiche Sprache (Orthografie, Lese- und Hörverstehen in deutscher und englischer Sprache, Kenntnisse in der

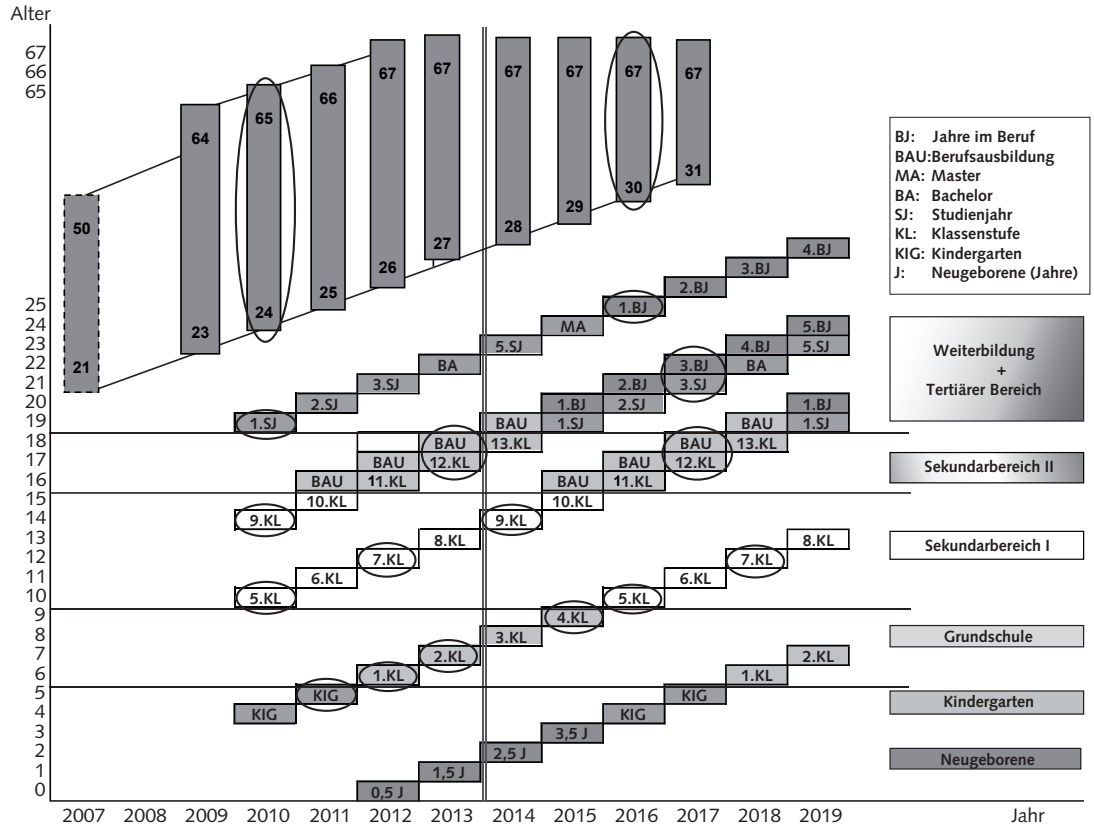


Abbildung 12: Multi-Kohorten-Sequenz-Design aus Ehmke et al. (2009, S.315)

Erstsprache bei Schülerinnen und Schülern mit Migrationshintergrund), Naturwissenschaften, Informationstechnologien (ICT)-Literacy und Problemlösen erfasst. Für den Kompetenzbereich Mathematik soll untersucht werden, wie sich mathematische Kompetenzen im Laufe der Entwicklung verändern und welche Einflüsse die Kompetenzen auf den Lebenslauf haben. Dabei wird mathematische Kompetenz als Ausmaß beschrieben, „[...] in dem Schülerinnen und Schüler, aber auch Erwachsene, die in der Schule gelernte Mathematik in problemhaltigen, vorwiegend außermathematischen Situationen flexibel anwenden können“ (Ehmke et al., 2009, S. 317). Der Begriff beschränkt sich allerdings nicht nur auf in der Schule gelernte Mathematik, sondern beinhaltet insbesondere auch die Fähigkeit, problematische Realsituationen in mathematische Sprache zu übersetzen, zu lösen und mit Blick auf die Realität zu interpretieren und zu validieren. Diese Vorstellung von mathematischer Kompetenz basiert auf dem Verständnis der Mathematical Literacy von PISA (Ehmke et al., 2009). Für eine gewisse Kompatibilität

mit dem curriculumsbasierten Ansatz des LV wurde das NEPS-Rahmenkonzept auch in Anlehnung an das LV-Rahmenkonzept entwickelt. Das NEPS-Kompetenzstrukturmodell beinhaltet eine inhaltliche und eine prozessbezogene Dimension. Die Zeit beziehungsweise Alterskohorte wird als eine implizite dritte Dimension betrachtet (Neumann et al., 2013). Die Prozesse und Inhaltsbereiche unterscheiden sich für die unterschiedlichen Alterskohorten. Im Folgenden werden die Teilbereiche für die Alterskohorte Sekundarstufe I beschrieben.

Das Kompetenzstrukturmodell des NEPS-K9-Mathematiktests beinhaltet vier mathematische Inhaltsbereiche, wie sie im OECD-Rahmenkonzept von 2003 beschrieben sind (Ehmke et al., 2009). Der Inhaltsbereich „Quantität“ bezieht sich auf alle Arten von Quantifizierungen, in denen Zahlen verwendet werden, um Situationen zu organisieren und zu beschreiben. Der Inhaltsbereich „Veränderung und Beziehungen“ bildet alle Arten von relationalen und funktionalen Beziehungen zwischen mathematischen Objekten ab. Alle Arten von ebenen und räumlichen Konfigurationen, Formen und Muster gehören zum Inhaltsbereich „Raum und Form“. Der letzte Inhaltsbereich „Daten und Zufall“ bezieht sich auf alle Arten von Phänomenen und Situationen, die statistische Daten beinhalten oder bei denen Zufall eine Rolle spielt (Ehmke et al., 2009). Jede Aufgabe kann einem dieser Inhaltsbereiche zugeordnet werden (Neumann et al., 2013).

Der Teilbereich Prozesse umfasst sechs Kompetenzen, die beim Lösen mathematischer Aufgaben erforderlich sein können. Dabei bezieht sich das NEPS-Rahmenkonzept auf die sechs kognitiven Komponenten wie sie von der KMK 2003 beschreiben wurden (Ehmke et al., 2009). Die Prozesse werden im Folgenden kurz erläutert. Der Prozess „mathematisch Argumentieren“ stellt den expliziten Umgang mit mathematischen Begründungen dar. Dazu gehören unter anderem das Nachvollziehen und Bewerten von gegebenen Argumentationen, das Begründen von Problemlösungen und das Durchführen von Beweisen. Der Prozess „mathematisch Kommunizieren“ beinhaltet das Verstehen von Aussagen jeglicher Art. Ein weiterer Prozess ist das „Modellieren“. Hierbei muss ein Situationsmodell aufgebaut und daraus ein mathematisches Modell entwickelt werden, welches anschließend in Realsituationen interpretiert und validiert werden soll. Strategien, die zur Wahl von konvergenten und divergenten Ansätzen, zum Betrachten von Spezialfällen, zum Verallgemeinern von Aussagen und Ähnlichem benötigt werden, gehören zum Prozess „mathematische Probleme lösen“. Der Prozess „Repräsentieren“ beinhaltet die Fähigkeit, mit unterschiedlichen Repräsentationsformaten von mathematischen Inhalten umgehen zu können. Der letzte Prozess „technische Fertigkeiten einsetzen“ wird

gebraucht, wenn Lösungsverfahren bekannt sind und abgerufen sowie umgesetzt werden müssen (Ehmke et al., 2009). Für die erfolgreiche Lösung einer Aufgabe können mehrere Prozesse benötigt werden (Neumann et al., 2013). Da der NEPS-Test aus praktischen Gründen lediglich geschlossene und halboffene Aufgaben verwendet, können vor allem aktive Fertigkeiten nicht abgedeckt werden. So ist es beispielsweise schwierig, die bei PISA definierte Kompetenz des Nutzens mathematischer Hilfsmittel, wie etwa das Nutzen eines elektronischen Taschenrechners, mit einem Papier-und-Bleistifttest als eigenständige Kompetenz zu messen (Neumann et al., 2013).

Trotz der Beschreibung der Teildimensionen im Kompetenzstrukturmodell werden diese nicht differenziert ausgewertet. Die achtundzwanzigminütige Testzeit lässt kein Facettendesign zu. Bei der Aufgabenentwicklung wurde jedoch eine gleichmäßige Verteilung über die Inhaltsbereiche sowie eine breite Abdeckung der Prozesse angestrebt (Ehmke et al., 2009). Das Kompetenzstrukturmodell wird mit 19 Multiple Choice Aufgaben, zwei Complex Multiple Choice Aufgaben und einer halboffenen Aufgabe abgebildet. Die Daten werden mit einem Multi-Kohorten-Sequenz-Design erfasst, wobei in unterschiedlichen Bildungsetappen unterschiedliche Instrumente eingesetzt werden (Blossfeld, von Maurice & Schneider, 2011; S. Weinert et al., 2011). Jede Schülerin und jeder Schüler bearbeitet alle 22 Mathematikaufgaben. Die so erhobenen Daten werden der Wissenschaft in Form von *scientific use files* für die Forschung zur Verfügung gestellt. Die erste Untersuchung von Schülerinnen und Schülern der neunten Klassenstufe wurde im Herbst 2010 durchgeführt.

1.2.5 Vergleich der mathematischen Kompetenzstrukturmodelle aus NEPS-K9, PISA 2012 und LV 2012

In Tabelle 3 werden die die Gemeinsamkeiten und Unterschiede zwischen den Studien zusammengefasst dargestellt. Zwischen den Studien lassen sich viele Gemeinsamkeiten erkennen. Sowohl das NEPS als auch PISA und der LV beanspruchen, die mathematische Kompetenz von Schülerinnen und Schülern am Ende der Sekundarstufe I zu messen. Das mathematische Kompetenzstrukturmodell des NEPS für die 9. Jahrgangsstufe orientiert sich sogar explizit an dem nationalen Kompetenzstrukturmodell des LV und an dem internationalen Kompetenzstrukturmodell von PISA. Die Kompetenzstruktur-

modelle von PISA und NEPS basieren auf einem Literacy-Ansatz, der explizit nicht curriculumsorientiert ist. Der LV bezieht sich in seinem Kompetenzstrukturmodell auf die KMK-Bildungsstandards für die Sekundarstufe I. Diese geben die Kompetenzen vor, die am Ende der neunten Klasse erreicht sein sollten und sind dementsprechend eine Richtlinie für die Kerncurricula der Länder. Der LV kann somit als curriculumsbasierter Test beschrieben werden. Durch die Bemühungen, den NEPS-Test auch an den LV anzulehnen, wird diesem daher auch in bedingtem Umfang eine Curriculumsbasierung zugeschrieben.

Für das Kompetenzstrukturmodell des NEPS werden, ebenso wie im LV und in PISA, Inhaltsbereiche und Prozesse definiert. NEPS bestimmt vier Inhaltsbereiche, deren Bezeichnung mit denen des PISA-Rahmenkonzeptes weitestgehend übereinstimmen, und ebenso sechs Prozesse, deren Benennungen mit den Prozessen des LV-Rahmenkonzeptes größtenteils übereinstimmen. Die Beschreibung der Teilkompetenzen in den Rahmenkonzeptionen der drei Studien ist nicht erschöpfend, was einen Vergleich der Inhalte erschwert. Die Teilbereiche in den jeweiligen Studien, besonders die Prozesse und Anforderungsbereiche, sind oft untereinander verzahnt und interagieren miteinander. Die Komplexität dieser Strukturen erschwert dadurch eine Abgrenzung der Teilbereiche zwischen den verschiedenen Studien.

Unterschiede, die trotz dieser Schwierigkeiten aufgedeckt werden können, werden im Folgenden beschrieben. Der größte Unterschied zwischen den Inhaltsbereichen der Studien ist, dass NEPS einen Inhaltsbereich weniger definiert als der LV. Der Inhaltsbereich „Messen“ wird im LV separat beschrieben, wogegen die Kompetenzen dieses Bereiches im NEPS, ebenso wie in PISA, in mehreren Inhaltsbereichen integriert sind. Das Messen von Längen, Flächen und Volumen sowie das Wählen von Einheiten fallen im NEPS beispielsweise unter den Bereich „Quantität“. Das Erkennen, Beschreiben und Analysieren von geometrischen Figuren und das gedankliche Operieren mit Strecken, Körpern und Formen gehören dagegen zum Bereich „Raum und Form“. Das systematische Messen von Daten im Zusammenhang mit Wahrscheinlichkeiten fällt wiederum unter den Bereich „Daten und Zufall“.
Bei den Prozessen werden mehrere Unterschiede zwischen den Studien deutlich. Zum einen können im NEPS durch die Verwendung von lediglich geschlossenen und halboffenen Aufgabenformaten nicht alle aktiven Kompetenzen, wie sie in den anderen beiden Tests beschrieben werden, abgedeckt werden. In diesem Zusammenhang definiert NEPS auch einen Prozess weniger als PISA. Mit den hauptsächlich geschlossenen Aufgaben

und dem Papier-und-Bleistift-Format im NEPS ist es schwierig, den Prozess „mathematische Hilfsmittel verwenden“ als eigenständige Kompetenz zu erfassen. Aus diesem Grund wird diese Kompetenz in verschiedene andere Prozesse integriert. Beim Prozess „mathematisch kommunizieren“ beschränkt sich das NEPS-Kompetenzstrukturmodell im Gegensatz zu den anderen beiden Studien auf das Verstehen und Evaluieren von Aussagen und Begründungen. Das aktive Präsentieren, Erklären und Begründen einer Lösung steht nicht im Fokus dieses Prozesses. Auch der Prozess „mathematische Darstellungen verwenden“ beinhaltet im NEPS vor allem passive Aktivitäten, in denen gegebene mathematische Repräsentationen genutzt werden. In PISA und im LV jedoch fallen unter die Kompetenz „Repräsentieren“ bzw. „mathematische Darstellungen verwenden“ auch das Erzeugen und Verändern von Darstellungen. Die Aufgaben im NEPS werden zwar anhand der Teildimensionen mit dem Ziel entwickelt, die Kompetenzen gleichmäßig abzudecken, jedoch nicht um diese auch trennscharf zu messen (Ehmke et al., 2009). Dagegen werden in PISA die Inhaltsbereiche und Anforderungsbereiche so operationalisiert, dass diese auch psychometrisch erfasst werden können. Der LV beschränkt sich auf die psychometrische Messung der Inhaltsbereiche, da die Prozesse und Anforderungsbereiche besonders stark und komplex interagieren und eine separate Messung daher nicht oder nur mit erheblichen Unschärfen möglich wäre (Roppelt et al., 2013).
Ein weiterer Unterschied zwischen den Studien ist, dass NEPS im Gegensatz zu PISA und dem LV keine Anforderungsbereiche definiert. Da diese eng mit den Prozessen verbunden sind und die drei Anforderungsbereich in PISA und im LV für jeden Prozess einzeln beschrieben werden, ist dies nicht weiter verwunderlich. Mit der beschränkten Testzeit und damit auch der beschränkten Aufgabenanzahl wäre es im NEPS nicht möglich, für jeden Anforderungsbereich in Kombination mit jedem Prozess genügend Aufgaben einzusetzen.
Auch Kontexte werden im NEPS, ebenso wie im LV, nicht definiert.

Zusammenfassend lässt sich festhalten, dass viele Gemeinsamkeiten aber auch einige Unterschiede zwischen den Studien aufgezeigt werden konnten. Der NEPS-K9-Mathematiktest kann mit seinen 22 Aufgaben und dem größtenteils geschlossenem Aufgabenformat die Mathematical Literacy nicht in einer vergleichbaren Genauigkeit wie PISA und das Curriculum nicht so präzise wie der LV erfassen. Dennoch scheint der NEPS-Test ähnliche mathematische Kompetenzen wie die aus PISA und dem LV zu messen. Welche Testwertinterpretationen mit dem NEPS-K9-Mathematiktest tatsächlich möglich sind, sollen die Analysen in dieser Arbeit zeigen. Im Folgenden wird daher ein Schema für

die Validierung der Testwertinterpretation des NEPS-K9-Mathematiktests angelehnt an Kane (2013) entwickelt. Dieses Schema soll die Grundlage für den Validierungsprozess darstellen.

Tabelle 3: Vergleich der Studien

		LV	PISA	NEPS
Testwert-interpretation	Intention	Überprüfung des Erreichens der länderübergreifenden Bildungsstandards	Erfassung und Vergleich, inwiefern Schülerinnen und Schüler in den verschiedenen Staaten auf das Erwachsenenleben, lebenslanges Lernen und auf die Anforderungen einer Teilhabe an der Gesellschaft vorbereitet werden	Untersuchung des Kompetenzerwerbs und der Bildungsbiographie über die Lebensspanne hinweg
	Referenzrahmen	sozial und kriterial	sozial und kriterial	sozial und ipsativ
	National/International	national	international	national
	Untersuchungsdesign	Querschnitt, Trend	Querschnitt, Trend	Längsschnitt
	Kompetenzbereiche	Deutsch, erste Fremdsprache, Mathematik, Naturwissenschaften	Lesen, Mathematik, Naturwissenschaften	Mathematik, Sprache, Naturwissenschaften, ICT-Literacy und Problemlösen
	Grundlegende Konzeption	Curriculumbasiert	Literacy	Literacy, in geringem Maße curriculumbasiert
Grundlage	Jahrgangsbasiert	Altersbasiert	Jahrgangsbasiert	

Fortsetzung auf der nächsten Seite

		LV	PISA	NEPS
	Stichprobenziehung	Mehrfach geschichtete Wahrscheinlichkeitsstichprobe	Mehrfach geschichtete Wahrscheinlichkeitsstichprobe	Zweifach geschichtete Wahrscheinlichkeitsstichprobe
	Größe der Stichprobe	ca. 45000	ca. 5000	ca. 15000
Durchführungsbedingungen	Testzeit Mathematikdomäne	60 bis 120 Minuten	30 bis 90 Minuten	28 Minuten
	Testdesign	Multi-Matrix-Design, 31 Testheftvarianten	Multi-Matrix-Design, 13 Testheftvarianten	2 Testhefte, Mathematik jeweils an vierter Stelle
	Erlaubte Hilfsmittel	Taschenrechner	Taschenrechner	Taschenrechner
Theoretische Konzeption	Inhaltsbereiche	Zahl; Messen; Raum und Form; funktionaler Zusammenhang; Daten und Zufall	Quantität; Raum und Form; Veränderung und Beziehungen; Unsicherheit und Daten	Quantität; Veränderung und Beziehungen; Raum und Form; Daten und Zufall
	Prozedurale Fähigkeiten	Mathematisch Argumentieren; Probleme mathematisch Lösen; mathematisch Modellieren; mathematische Darstellungen verwenden; mit symbolischen, formalen und technischen Elementen der Mathematik umgehen; mathematisch Kommunizieren	Situationen mathematisch formulieren; mathematische Konzepte, Fakten, Prozeduren und Schlussfolgerungen anwenden; mathematische Ergebnisse interpretieren, anwenden und bewerten	Mathematisch Argumentieren; mathematisch Kommunizieren; Modellieren; mathematische Probleme lösen; Repräsentieren; technische Fertigkeiten einsetzen

Fortsetzung auf der nächsten Seite

	LV	PISA	NEPS
kognitive Anforderungsbereiche	Reproduzieren; Zusammenhänge herstellen; Verallgemeinern und Reflektieren	Reproduzieren; Zusammenhänge herstellen; Verallgemeinern und Reflektieren	nicht definiert
Kontexte	nicht definiert	persönlich; ausbildungs- bzw. berufsbezogen; gesellschaftsbezogen; wissenschaftlich	nicht definiert
Kompetenzstufen	5 Kompetenzstufen	6 Kompetenzstufen	keine Kompetenzstufen

2 Ein Interpretation/ Use Argument für NEPS

In diesem Kapitel soll ein Interpretation/Use Argument (IUA) für den NEPS-K9-Mathematiktest angelehnt an den Argument Based Approach von Kane (1990, 2001, 2006, 2008, 2012, 2013) entwickelt werden. Eine Prüfung der Validität für die Tests, welche in NEPS angewendet werden, ist sehr relevant, da es sich bei der Studie um ein *Large Scale Assessment* handelt, mit welchem repräsentative Daten für relevante Bereiche des deutschen Bildungssystems gewonnen werden (Blossfeld & von Maurice, 2011). Eine wichtige Kompetenz, die über die Lebensspanne hinweg untersucht wird, ist die mathematische Kompetenz. Die zuverlässige Bestimmung der Mathematikkompetenzen am Ende der neunten Klasse ist von besonderem Belang, da die Pflichtschulzeit in Deutschland für die meisten Schülerinnen und Schüler nach der neunten Klasse endet und da Mathematikkompetenzen als eine wichtige Voraussetzung für die aktive Partizipation in der Gesellschaft gelten. Daher scheint die Validierung des Mathematiktests für die neunte Klasse besonders bedeutend. Der Argument Based Approach von Kane (1990, 2001, 2006, 2008, 2012, 2013) ist für die Validierung des NEPS-Mathematiktests insbesondere geeignet, da hierbei eine logisch zusammenhängende und transparente Argumentationskette aufgebaut wird, welche die Interpretation der Testwerte rechtfertigt. Dabei werden multiple Schlussfolgerungen überprüft und gestützt. Es besteht jedoch auch die Möglichkeit, Gegenargumente in die Argumentationskette einzubinden. Des Weiteren bietet das Modell Flexibilität für das Einfügen von Erweiterungen der Testwertinterpretationen, wie zum Beispiel die Schlussfolgerung des *Konstruktbezuges*, der *Skalierung* oder der *Domänenbeschreibung*. Dabei ist kein stark entwickeltes Konstrukt als Basis für die Validierung notwendig.

> „Using this approach, validation may be challenging, but it is doable“ (Brennan, 2013, S.75).

Eine weitere Möglichkeit, die sich durch die Validierung von Testwertinterpretationen des NEPS mit dem Argument Based Approach nach Kane (1990, 2001, 2006, 2008, 2012, 2013) ergibt, ist das Schaffen eines praxisbezogenen Beispiels und damit stärkerer Deutlichkeit über das Vorgehen nach dem Argument Based Approach von Kane (2013). Im Eingangskapitel wurde bereits beschrieben, dass Testinstrumente im *Low-Stake*-Bereich der Bildungswissenschaft oft nicht vollständig und nicht nach aktuellen Standards auf Validität untersucht werden (Cizek et al., 2010, 2008; Hogan & Agnello, 2004). Ein Grund für die Nichtumsetzung neuerer Ansätze kann unter anderem dem fehlenden praktischen Bezug und fehlenden Beispielen geschuldet sein. Beispielsweise sind die Texte zum Argument Based Approach nach Kane eher theoretischer Art und enthalten nur einige wenige konkrete Beispiele für die Umsetzung des Ansatzes (Kane, 1990, 2001, 2006, 2008, 2012, 2013). Auch haben bisher nur wenige Studien diesen Ansatz für die Validierung von Testwertinterpretationen implementiert und die konkrete Umsetzung des Ansatzes unterscheidet sich für die unterschiedlichen Tests voneinander (Chapelle et al., 2009; Shaw & Crisp, 2012). Im Folgenden wird daher ein adaptiertes Rahmenmodell zur Validierung des NEPS-Mathematiktests für die neunte Klasse nach dem Argument Based Approach von (Kane, 2013) entwickelt. Dabei wird Kanes theoretische Beschreibung des Modells als Basis für die Bildung des Interpretation / Use Argument für den NEPS-K9-Mathematiktest verwendet. Kane (2013) betonte jedoch, dass seine Beschreibung des Argument Based Approach keineswegs ausschöpfend sei und abhängig von der geplanten Testwertinterpretation angepasst werden solle. Dementsprechend wird die basale Argumentationsstruktur des IUA von Kane (2013) so erweitert und konkretisiert, dass die einzelnen Interpretationsschritte von den Rohwerten des NEPS-K9-Mathematiktests zur mathematischen Kompetenz, wie sie im NEPS-Rahmenkonzept beschrieben werden, transparent werden. Dafür wird der Ansatz von Kane (2013) unter anderem um die Interpretationsschritte *Domänenbeschreibung*, *Skalierung* und *Konstruktbezug* sowie um konkrete Hypothesen für den NEPS-K9-Mathematiktest erweitert. Die deskriptive Interpretation der Testergebnisse des NEPS-K9-Mathematiktests, die in dieser Arbeit validiert werden soll, ist:

> Die Unterschiede in der beobachteten Testleistung geben die Unterschiede in der mathematischen Fähigkeit wieder, wie diese im Rahmenkonzept definiert wird.

Andere Interpretationen der Testergebnisse sind sicherlich möglich, werden in dieser Arbeit jedoch nicht behandelt. Die Argumentationskette, welche in diesem Kapitel ent-

wickelt wird, besteht aus sechs Schlussfolgerungen, die das beobachtete Verhalten im NEPS-Mathematiktest mit der oben formulierten Testwertinterpretation verbinden (siehe Abbildung 13).

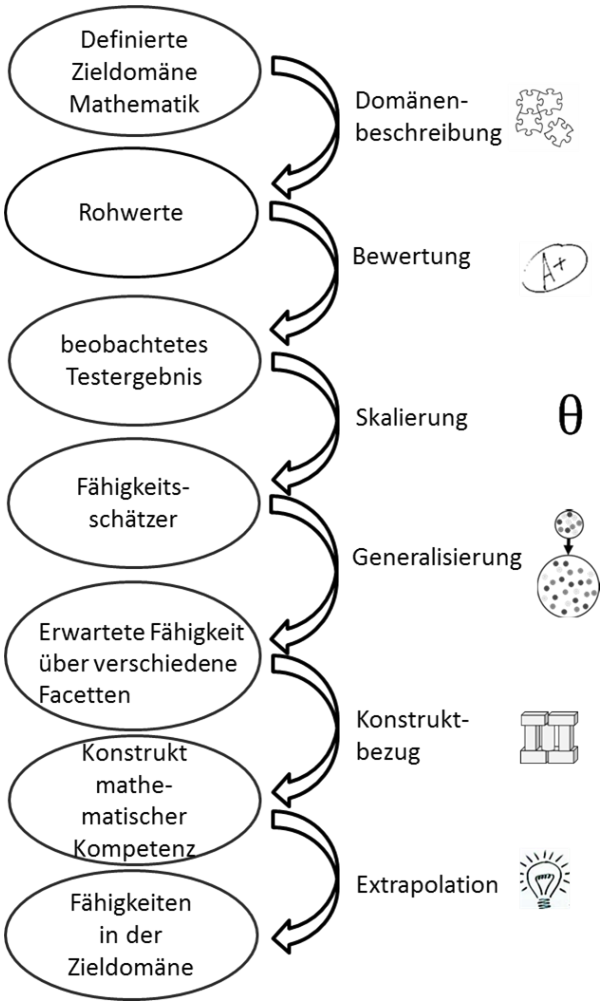


Abbildung 13: Argumentationskette für den NEPS-K9-Mathematiktest

Die Schlussfolgerungen basieren nach Kane (1990, 2001, 2006, 2008, 2012, 2013) jeweils auf einem Argument, welches sich wiederum auf Annahmen stützt (vgl. Abbildung 14, siehe auch Theoretischer Hintergrund, Kapitel 1.2). Es sei angemerkt, dass die Vorgehensweise sich bei der Bildung des IUA chronologisch von dem intuitiven Aufbau eines Argumentes unterscheidet. Der Ausgangspunkt des IUA ist die Schlussfolgerung, die für die Testwertinterpretation gerechtfertigt werden soll. Für die Rechtfertigung wird ein Argument aus der Schlussfolgerung abgeleitet, das die Schlussfolgerung stützt. Aus dem Argument lassen sich wiederum Annahmen ableiten, welche als Evidenz für das

Argument bewiesen werden müssen.

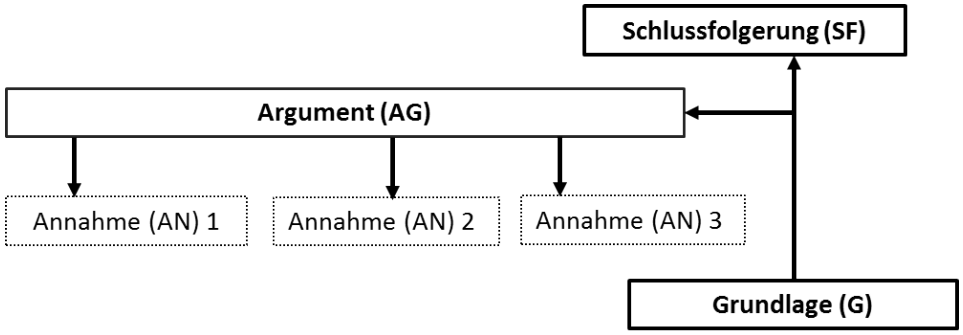


Abbildung 14: Grundmodell der Argumentation nach Kane (1990, 2001, 2006, 2008, 2012, 2013)

In den Abschnitten 2.1 bis 2.6 werden die Argumentationsstruktur mit den Schlussfolgerungen, Argumenten und Annahmen für das IUA des NEPS-K9-Mathematiktests sowie konkrete Hypothesen entwickelt. Dafür wird zunächst die zu ziehende Schlussfolgerung für den NEPS-Test beschrieben und das Argument, auf welchem die jeweilige Schlussfolgerung basiert, mit Hilfe von Literatur bezüglich des Argument Based Approach von Kane (1990, 2001, 2006, 2008, 2012, 2013) hergeleitet. Anschließend werden für jedes Argument die benötigten Annahmen für den NEPS-Test basierend auf Literatur zum Argument Based Approach und Literatur zur Validierung von Testwertinterpretationen für pädagogische und psychologische Tests entwickelt. Die in diesem Kapitel ausgearbeiteten Schlussfolgerungen, Argumente und Annahmen des IUA sind zwar auf den NEPS-K9-Mathematiktest zugeschnitten, sind jedoch relativ global formuliert. Für eine Validierung der Testwertinterpretation müssen die Annahmen nach Kane für den NEPS-K9-Mathematiktest bewiesen werden, sodass die jeweiligen Argumente gerechtfertigt und die Schlussfolgerungen angenommen werden können. Die Annahmen bilden daher die Grundlage für spezifische Hypothesen. Das auf (Kane, 2013) basierende Argumentationsmodell für das IUA des NEPS wird damit um konkrete Hypothesen erweitert (vgl. Abbildung 15).

Nach der Formulierung einer Annahme für das jeweilige Argument einer Schlussfolgerung werden daher zuerst mögliche Arten der Prüfung von Hypothesen bezüglich dieser Annahmen mittels einer Zusammenfassung bereits durchgeführter Studien dargestellt. Darauf folgend werden konkrete Hypothesen aus der jeweiligen Annahme für den NEPS-K9-Mathematiktest abgeleitet. Abschließend wird die Struktur der Schlussfolgerung mit

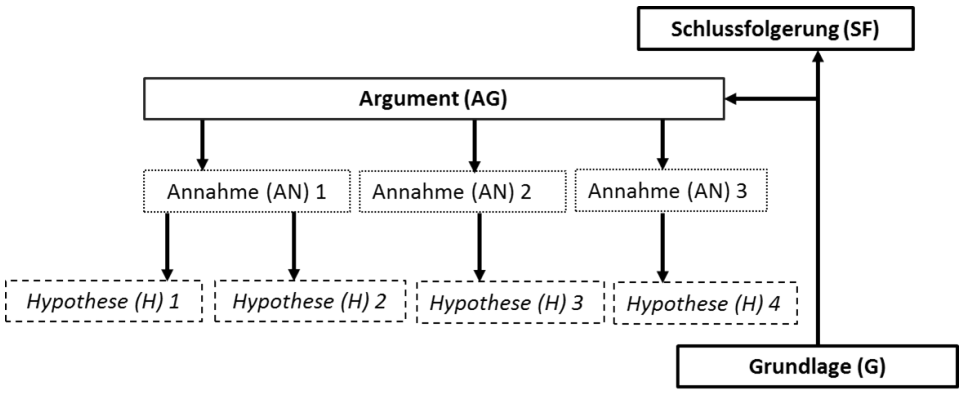
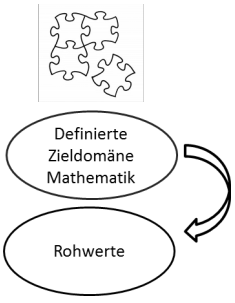


Abbildung 15: Erweiterung des Argumentationsmodells

ihrem Argument, den Annahmen und Hypothesen graphisch angelehnt an Kane (1990, 2001, 2006, 2008, 2012, 2013) und Toulmin (1958, 2003) dargestellt (vgl. Abbildung 15). Im Argumentationsschema nach Kane (2013), welches wiederum auf Toulmin (1958, 2003) basiert, werden die Elemente Schlussfolgerung, Argument und Annahme in Form von Behauptungen formuliert. Für diese Dissertation im Sinne einer Forschungsarbeit mit der Fragestellung nach der Validität der Testwertinterpretation können die Argumente für jede Schlussfolgerung als Forschungsfragen und die Annahmen als Subfragen betrachtet werden. Für eine einheitliche Verwendung des Argumentationsschemas werden die genannten Elemente jedoch nicht in Form von Fragen formuliert.

2.1 Domänenbeschreibung



Die Schlussfolgerung *Domänenbeschreibung* für den NEPS-Mathematiktest für die Klassenstufe 9 lautet: **„Von der definierten Zieldomäne Mathematische Kompetenz kann auf die Rohwerte in der Testdomäne geschlossen werden.“** (siehe auch Abbildung 13).

Die Erweiterung der Argumentationskette durch die Beschreibung der Domäne wird in dieser Studie angewendet, da sie wichtige Informationen für die Interpretation der Ergebnisse liefert. Obwohl Kane (2004) die Beschreibung des Bezugsbereiches nicht als

Schlussfolgerung aufnimmt, hebt er dennoch die Wichtigkeit einer solchen Beschreibung hervor.

> „[...] if the test is intended to be interpreted as a measure of competence in some domain, then efforts to describe the domain carefully and to develop items that reflect the domain (in terms of content, cognitive level, and freedom from potential sources of systematic errors) tend to support the intended interpretation." (Kane, 2004, S.141).

Diese Schlussfolgerung basiert damit auf folgendem Argument:

AG: Die Rohwerte im NEPS-Mathematiktest spiegeln relevante und repräsentative Elemente für die definierte Zieldomäne Mathematische Kompetenz wider.

2.1.1 Relevanz Teilkompetenzen

Aus dem oben genannten Zitat von Kane kann auch abgeleitet werden, dass für eine repräsentative Widerspiegelung der Zieldomäne die Teilkompetenzen durch die verwendeten Aufgaben angemessen abdeckt werden sollten. Auch die Standards (2014) heben die Wichtigkeit einer solchen Prüfung im Rahmen einer Validitätsuntersuchung hervor und empfehlen logische und empirische Analysen zur Repräsentativität des Testinhaltes für die Zieldomäne sowie Expertenreviews betreffend der Zuordnung von Testteilen beziehungsweise –aufgaben zum Testkonstrukt (Standards, 1999). Eine Annahme, die sich aus dem Argument der *Domänenbeschreibung* ableiten lässt, ist die folgende:

AN: Die Aufgaben des NEPS-Mathematiktests decken die für die Zieldomäne relevanten Teilkompetenzen angemessen ab.

Im Folgenden werden zunächst mögliche Arten der Prüfung von Hypothesen bezüglich dieser Annahme mittels einer Zusammenfassung bereits durchgeführter Studien dargestellt. Die Abdeckung relevanter Teilkompetenzen wurde bereits für mehrere Kompetenztests untersucht. Shaw und Crisp (2012) untersuchten beispielsweise die Relevanz der mit den Aufgaben gemessenen Teilkompetenzen des International A Level Physics Tests im Rahmen der Beschreibung der Domäne mit einem Expertenreview. Dabei wurde für alle Aufgaben von sechs Experten eingeschätzt, inwiefern diese die im Rahmenkonzept

beschriebenen Inhalte und kognitiven Prozesse messen. Die Autoren fanden heraus, dass alle im Rahmenkonzept definierten Inhalte und kognitiven Prozesse durch die Aufgaben abgedeckt wurden, wobei einige Inhaltsbereiche durch nur wenige Aufgaben repräsentiert wurden. Den Autoren zufolge stelle dies eine potentielle Gefährdung der Validität des Testes dar und es solle untersucht werden, ob die unterschiedliche Gewichtung der Inhalte theoretisch sinnvoll ist. In einer anderen Studie untersuchten Burstein, Aschbacher, Chen und Lin (1990) die Relevanz und Gewichtung der mit den Tests aus dem Mathematics Diagnostic Testing Project (MDTP) für die achte und zwölfte Klassenstufe gemessenen Teilkompetenzen, indem sie die verwendeten Aufgaben den Bereichen der MDTP-Rahmenkonzeption sowie den Inhaltsbereichen von drei anderen internationalen und nationalen Mathematiktests für die achte und zwölfte Klassenstufe zuwiesen und die so entstehenden Itemverteilungen mit denen der anderen Tests verglichen. Zusätzlich verglichen sie die Inhalte des MDTP-Rahmenkonzeptes mit den vier unterschiedlichen publizierten Stellungnahmen zu gewünschten Curriculumserwartungen. Die Testaufgaben verteilten sich logisch über die Inhalte und die vier Niveaustufen des Tests. Auch ließen sich beim Vergleich mit den drei Mathematiktests Ähnlichkeiten in der Verteilung auf die Inhaltsbereiche feststellen, jedoch auch Unterschiede, die mit dem Fokus des MDTP auf Algebra zusammenhängen. Nicht alle Beschreibungen der Curricula konnten im Rahmenkonzept des Tests identifiziert werden. Dies lässt sich ebenfalls meist mit dem Fokus des MDTP auf Algebra erklären und damit, dass es sich in den Beschreibungen der Curricula nicht um Inhalte, sondern Prozesse und Prozeduren handelt.

2.1.2 NEPS-Hypothesen zu Teilkompetenzen

In den oben genannten Studien werden für die Untersuchung der Relevanz der Teilkompetenzen vor allem Hypothesen über den Zusammenhang der Aufgaben in Bezug auf das jeweilige Rahmenkonzept beziehungsweise ähnliche Rahmenkonzepte untersucht. Für den NEPS-K9-Mathematiktest lassen sich konkrete Hypothesen aus der Annahme der *Domänenbeschreibung* ableiten. In der Rahmenkonzeption des NEPS-K9-Mathematiktests wird die Zieldomäne Mathematische Kompetenz beschrieben. Dabei werden eine kognitive und eine inhaltliche Komponente der mathematischen Kompetenz definiert. Bei der Definition der mathematischen Kompetenz und auch dieser Komponenten lehnt sich das Rahmenkonzept beabsichtigt stark an die Rahmenkonzeptionen von PISA und dem LV an (vgl. Kapitel 2.1.1). Aus diesem Grund und weil die PISA- und

LV-Rahmenkonzeption jeweils anerkannte internationale beziehungsweise nationale Definitionen mathematischer Kompetenz beinhalten, können diese Rahmenkonzepte als relevant für die Zieldomäne Mathematische Kompetenz in NEPS eingestuft werden. Sowohl die PISA- als auch die LV-Rahmenkonzeption wurde von Experten entwickelt und umfassend evaluiert und überarbeitet (OECD, 2010; Pant, Stanat, Pöhlmann & Böhme, 2013). Wird also im Zuge der Annahme der *Domänenbeschreibung* davon ausgegangen, dass die Aufgaben des NEPS-Mathematiktests die für die Zieldomäne relevanten Teilkompetenzen angemessen abdecken, so sollte sich die konzeptionelle Nähe zu den Rahmenkonzeptionen vom LV und von PISA auch in den Aufgaben des NEPS-Tests widerspiegeln. Folgende Hypothese kann für den NEPS-K9-Mathematiktest aufgestellt werden:

H1: In den Aufgaben des NEPS-K9-Mathematiktests lassen sich die relevanten Domänen mathematischer Kompetenz aus dem Rahmenkonzept von PISA und dem des LV identifizieren.

In der Rahmenkonzeption der NEPS-Mathematiktests werden ebenso wie in PISA und dem LV Inhaltsbereiche und kognitive Prozesse definiert. Die Operationalisierung dieser Bereiche durch die Aufgaben bestimmt den realisierten kognitiven beziehungsweise inhaltlichen Umfang der Bereiche. Die konzeptionelle Nähe der Rahmenkonzepte sollte sich daher auch in der Operationalisierung der Bereiche widerspiegeln. Folgende Hypothese lässt sich daher für den NEPS-K9-Mathematiktest ableiten:

H2: Die kognitiven und inhaltlichen Teilbereiche werden in NEPS und dem LV beziehungsweise NEPS und PISA auf ähnliche Weise operationalisiert.

Die PISA- und LV-Mathematiktests wurden jeweils anhand ihrer international beziehungsweise national anerkannten Rahmenkonzeption entwickelt, wobei die jeweiligen Komponenten so gewichtet wurden, dass die Zieldomäne durch die Aufgaben angemessen repräsentiert wird. Wird von der Annahme ausgegangen, dass die relevanten Teilkompetenzen für die Zieldomäne angemessen abgedeckt werden, so sollte auch die Gewichtung der Teilkompetenzen im NEPS mit der Gewichtung in den Studien LV und PISA übereinstimmen. Für die Annahme der *Domänenbeschreibung* ergibt sich daher folgende Hypothese:

H3: Die Gewichtung der Komponenten aus den PISA- und LV- Rahmenkonzepten im NEPS-K9-Mathematiktest unterscheidet sich nicht signifikant von der Gewichtung in den Mathematiktests aus den Haupterhebungen PISA 2012 und LV 2012.

Aus der Schlussfolgerung, dem Argument, der Annahme und den daraus abgeleiteten Hypothesen der *Domänenbeschreibung* lässt sich unten stehendes Argumentationsschema erstellen (Abbildung 16).

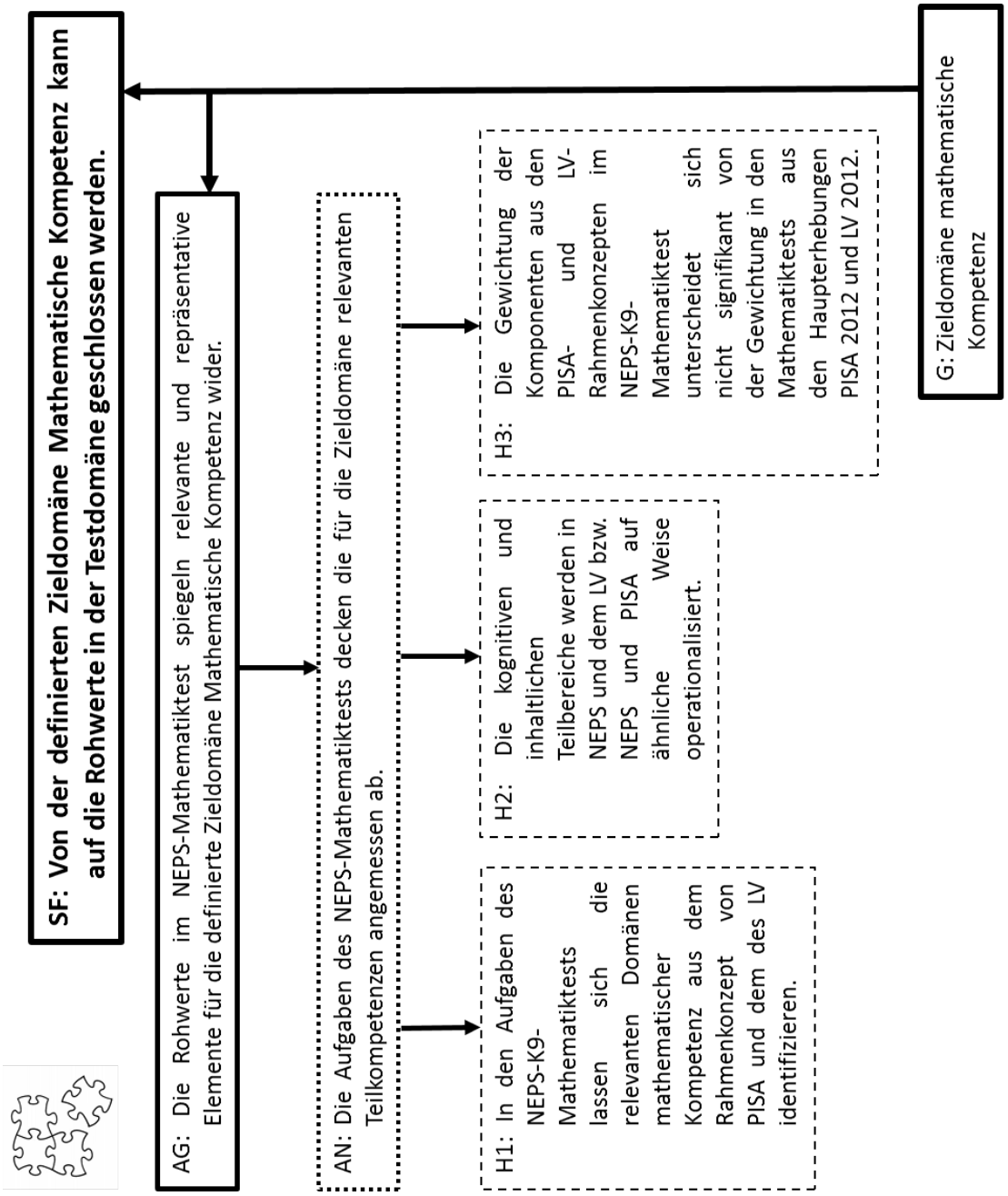
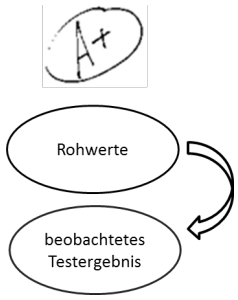


Abbildung 16: Argumentationsschema der *Domänenbeschreibung*

2.2 Bewertung



Die *Bewertung* steht an zweiter Stelle der Argumentationskette für die Interpretationen der Ergebnisse des NEPS Mathematiktests für die neunte Klasse. Diese Schlussfolgerung lautet für den NEPS-Mathematiktest für die Klassenstufe 9: **„Von den Rohwerten kann auf das Testergebnis geschlossen werden“** (siehe auch Abbildung 13).

In einem Test werden die Testpersonen normalerweise aufgefordert, Aufgaben zu bearbeiten. Die Lösungen der Schülerinnen und Schüler bilden die Rohwerte. Um die Rohwerte für Interpretationen nutzen zu können, müssen diese zunächst beispielsweise als richtig, falsch, fehlend etc. kodiert werden und aus diesen Kodierungen muss ein Testergebnis gebildet werden (Kane, 2013). Die Schlussfolgerung *Bewertung* basiert daher auf dem folgenden Argument:

AG: Die Rohwerte in den NEPS-Mathematikaufgaben führen zu Testergebnissen, die repräsentativ für die Zieldomäne Mathematische Kompetenz sind.

Das Argument bezieht sich an dieser Stelle jedoch nur auf die klassische Bewertung der Aufgaben, die Bildung von skalierten Testwerten ist eine eigenständige Schlussfolgerung der Argumentationskette.

2.2.1 Anwendung der Aufgabenkodierung

Kane (2013) hebt im Zusammenhang mit der Schlussfolgerung *Bewertung* die Wichtigkeit einer korrekten Anwendung der Bewertungskriterien hervor und fordert, dass diese frei von offenkundigen Fehlern ist. Cohen und Wollack (2006) stellen die Wichtigkeit der Prüfung einer solchen Annahme ebenfalls heraus und empfehlen, die Daten nach der Eingabe zu prüfen, zum Beispiel durch die Berechnung von Interraterreliabilität zwischen unterschiedlichen Eingaben, Kontrolle für große Sequenzen fehlender Werte oder Kontrolle für erwartete und unerwartete Muster in der Testergebnisverteilung. Aus dem Argument der *Bewertung* kann die folgende Annahme geschlossen werden:

AN1: Die Aufgabenkodierung wurde bestimmungsgemäß und den Kodieranweisungen entsprechend durchgeführt.

Nachstehend werden zwei Studien zusammengefasst, die Hypothesen bezüglich dieser Annahmen untersucht haben. Shaw und Crisp (2012) prüften in ihrer Studie die Anwendung der Bewertungskriterien für den International A Level Physics Test, indem sie ein Review von Dokumenten durchführten, in denen die Kodieranweisungen beschrieben waren. Insgesamt fanden sie deutliche und umfassende Anweisungen für alle Akteure und alle Stufen der Bewertung. Außerdem wurde die Reliabilität der Bewertung getestet, indem einige Testhefte doppelt kodiert wurden. Insgesamt fanden die Autoren hohe Pearson Interrater Korrelationen von $r = .80$ bis $r = .85$ zwischen den Gesamtergebnissen. Des Weiteren wurde die Raterübereinstimmung zusätzlich mit einem Multi-Facetten Rasch Modell und dem Rasch-Cohen's Kappa berechnet. Die Rasch-Analyse ergab, dass die Rater sich in ihren Einschätzungen ähnlich verhielten und dass es keine Rater gab, die nicht zum Modell passten. Das Rasch-Cohen's Kappa für exakte Übereinstimmungen lag mit $\kappa = .18$ in einem annehmbaren Bereich.
Die Anwendung der Bewertungskriterien wurde ebenfalls im Rahmen der PISA-Studie ausgewertet (OECD, 2009, 2012, 2014). In der PISA-Studie von 2006 wurden beispielsweise mehrere Blöcke aus 600 Testheften mehrfach kodiert. Anschließend wurde der Generalisierbarkeitskoeffizient sowie Komponenten der Varianz für jedes Land berechnet. Für die meisten Länder wurden akzeptable Werte nachgewiesen (OECD, 2012). Daraus kann geschlossen werden, dass die Bewertungskriterien von den Ratern auf gleiche Weise angewendet wurden.

2.2.2 NEPS-Hypothesen zur Anwendung der Aufgabenkodierung

In den oben beschriebenen Studien wird die Anwendung der Bewertungskriterien vor allem durch Hypothesen zur mehrfachen Kodierung und Eindeutigkeit des Prozesses überprüft. Für den NEPS-K9-Mathematiktest lässt sich eine Hypothese bezüglich der Eindeutigkeit des Prozesses ableiten. Für die Bewertung der Aufgaben wurden zunächst alle Testhefte in SPSS eingegeben und anschließend durch eine Syntax kodiert. Dieser Prozess muss für die Annahme, dass die Bewertung konsistent und den Kodieranweisungen gemäß angewendet wurde, fehlerfrei verlaufen sein. Folgende Hypothese lässt sich

für den NEPS-K9-Mathematiktest aufstellen:

H1: Fehler in der Eingabe und Kodierung der Testhefte aller Schülerinnen und Schüler können ausgeschlossen werden.

2.2.3 Angemessenheit der Aufgabenkodierung

Laut Kane (2013) müsse die Aufgabenkodierung angemessen sein, um bei der Schlussfolgerung *Bewertung* zu repräsentativen Testergebnissen zu gelangen. Unterschiedliche Aufgabenformate würden dabei unterschiedlich viel Stützung der Annahmen benötigen. So könne man davon ausgehen, dass in Multiple Choice Aufgaben die richtige Antwort vorhanden ist beziehungsweise mehrere richtige Antworten vorhanden sind und der Auswertungsschlüssel für die Kodierung der Antworten korrekt ist (Kane, 2004). Bei offenen Antwortformaten müsse jedoch ein detailliertes Aufgabenkodierschema entworfen werden, welches umfassend ausgearbeitet und geprüft werden müsse (Lane & Stone, 2006). Lane und Stone (2006) heben zudem ebenfalls die Wichtigkeit der Evaluation der Aufgabenkodierung hervor und unterstreichen dies mit einer Aussage von Messick (1989b), dass die Validität der Testwertinterpretation und -nutzung von der Genauigkeit der Wiedergabe des gemessenen Konstrukts durch die erreichten Testergebnisse abhänge. Die zweite Annahme, die dem Argument der *Bewertung* unterliegt, ist daher folgende:

AN2: Die Bewertungskriterien des NEPS-Mathematiktests für die 9. Klassenstufe ermöglichen die Bildung korrekter Testergebnisse.

Nachfolgend werden Studien vorgestellt, die Hypothesen bezüglich dieser Annahme geprüft haben. Im Rahmen der Validierung des Mayer-Salovey-Caruso Emotional Intelligence Test (MSCEIT) untersuchte Maul (2012) beispielsweise die verwendete Bewertungsmethode, indem er Literatur über die verwendete Methode des Consenus Scoring und eine bereits durchgeführt Studie über die Interraterreliabilität beim Consenus Scoring mit einer allgemeinen und einer Expertengruppe auswertete. Der Autor kommt auf Grund der Literatur zu dem Schluss, dass die Consenus Methode für das Bewerten von Emotionen ungeeignet sei. Emotionen würden nicht ausschließlich durch einvernehmliche Nutzung und Interpretation definiert, wodurch eine einvernehmliche Antwortmöglichkeit nicht zwingend korrekt sei. Außerdem deckt Maul auf, dass die Qualifikationen der Experten aus dem Experten-Consenus-Scoring nicht dokumentiert und somit nicht

hinreichend belegt seien, um für ein Experten-Scoring geeignet zu sein. Des Weiteren würden die im Vergleich niedrigeren bis gleich hohen Übereinstimmungen der Expertengruppe zu der allgemeinen Gruppe vermuten lassen, dass diese nicht mehr Expertise bezüglich der gefragten Inhalte besitzen als die allgemeine Gruppe (Maul, 2012). Wongwiwatthananukit, Popovich und Bennett (2000) untersuchten die Angemessenheit von Partial Credit Scoring für einen Test über das Fachwissen von Pharmaziestudentinnen und -studenten, indem sie diese mit einem dichotomen Scoring verglichen. Die Autoren fanden unter anderem für das Partial Credit Scoring höhere Itemschwierigkeiten und Reliabilitäten. Außerdem fanden sie einen Coefficient of effective length (CEL) zwischen 1.11 und 1.26, der angibt, dass die Partial Credit Methode für den Test über das Fachwissen von Pharmaziestudentinnen und -studenten effektiver im Erfassen von Fachwissen ist als die dichotome Methode.

2.2.4 NEPS-Hypothesen zur Angemessenheit der Aufgabenkodierung

In den oben genannten Studien zur Angemessenheit der Bewertungskriterien werden vor allem Hypothesen über die Methode der Zuweisung von Testscores und den Einfluss auf Test- und Itemeigenschaften geprüft. Für den NEPS-Test lassen sich Hypothesen mit einem ähnlichen Fokus ableiten. Der NEPS-K9-Mathematiktest besteht aus 22 Aufgaben, davon haben 19 ein einfaches Multiple Choice (MC)-Format und zwei ein Complex Multiple Choice (CMC)-Format. Eine Aufgabe hat ein Short Constructed Response (SCR)-Format. Für die Bewertung der Aufgaben wurde beschlossen, dass die verschiedenen Aufgabenformate unterschiedlich gewichtet werden und dass fehlende Werte ignoriert (das heißt als nicht vorliegend gewertet) werden. Wird von der Annahme ausgegangen, dass die Bewertungskriterien angemessen sind, so lässt sich folgende Hypothese für die *Bewertung* ableiten:

H2: Die unterschiedlichen Gewichtungen der Aufgabenformate und der Umgang mit fehlenden Werten führen zu unverfälschten Testergebnissen.

2.2.5 Psychometrische Itemeigenschaften

Nach Kane (2013) müsse davon ausgegangen werden können, dass die Interpretation der Ergebnisse nach der Durchführung der Bewertung nicht durch anderweitige Umstände beeinflusst werden kann. Kelava und Moosbrugger (2012) stellen heraus, dass die eingesetzten Testaufgaben dem Einsatzzweck des Messinstrumentes gerecht werden sollten. Nur wenn dies gegeben sei, könne die Bewertung der Aufgaben zu Testergebnissen führen, die repräsentativ für die Zieldomäne Mathematischen Kompetenz sind. So sollte in diesem Zusammenhang die Qualität der Distraktoren angemessen sein (Jonkisz, Moosbrugger & Brand, 2012), die Items sollten zwischen Testteilnehmern mit hohen und niedrigen Fähigkeiten differenzieren können (Kelava & Moosbrugger, 2012) und die Aufgaben des Tests sollten intern konsistent sein (Schermelleh-Engel & Werner, 2012). Die dritte Annahme, die aus dem Argument der *Bewertung* geschlossen werden kann, ist folgende:

AN3: Eine hohe psychometrische Qualität des Testinstruments ist gewährleistet.

Im Folgenden werden zwei Studien vorgestellt, die Hypothesen bezüglich dieser Annahme geprüft haben. Knigge (2010) untersuchte beispielsweise für den Test „Musik wahrnehmen und kontextualisieren“ unter anderem die Trennschärfe für jedes Item und für jede Antwortkategorie. Auf diese Weise sollte sichergestellt werden, dass Aufgaben den Gesamttest repräsentieren und zwischen Personen mit einer hohen und einer niedrigen Kompetenzausprägung unterscheiden können. Außerdem wurde analysiert, ob die Distraktoren ähnliche Schwierigkeiten und negative Trennschärfen aufweisen. In diesen Untersuchungen wurden einige Aufgaben identifiziert, welche die Bildung eines repräsentativen Testwertes gefährdeten. Diese Aufgaben wurden entweder eliminiert oder überarbeitet. Außerdem wurde die Reliabilität für jedes Testheft berechnet. Auch für den General Mathematics Placement Test (MPT-G) wurden die psychometrischen Itemeigenschaften in einer Pilotstudie analysiert. Dafür wurden acht Testversionen mit je 35 Aufgaben von 1566 Studenten bearbeitet. Die acht Versionen entstanden aus der Kombination von zwei unterschiedlichen Aufgabensets mit respektive drei und vier Distraktoren in zwei unterschiedlichen Reihenfolgen. Die Autoren prüften unter anderem die interne Konsistenz der Testversionen, die Trennschärfe für jede Aufgabe und die Häufigkeit der Auswahl sowie die Trennschärfe der Distraktoren. Die Mehrzahl der Aufgaben hatte gute beziehungsweise zufriedenstellende Trennschärfen. Die Analyse der Distraktoren

ergab, dass die Testversionen mit drei Distraktoren für den MPT-G zu bevorzugen sind (McGhee, Peterson, Gillmore & Lowell, 2008).

2.2.6 NEPS-Hypothesen zu den psychometrischen Itemeigenschaften

In den oben beschriebenen Studien zu Itemeigenschaften wurden vor allem Hypothesen über die Trennschärfe von Aufgaben und Distraktoren sowie deren Einfluss auf die Reliabilität des Tests untersucht. Auch für den NEPS-Mathematiktest lassen sich Hypothesen bezüglich dieser Itemeigenschaften ableiten. Alle Aufgaben des NEPS-K9-Mathematiktests sollten zur Messqualität des Tests beitragen. Durch die Bewertung der Aufgaben werden Itemeigenschaften sichtbar, welche die Repräsentativität des Ergebnisses erheblich beeinflussen können. Wird davon ausgegangen, dass die Itemeigenschaften die Bildung eines repräsentativen Testergebnisses ermöglichen, dann tragen die Itemeigenschaften zur Messqualität des Tests bei. Folgende Hypothesen lassen sich für den NEPS-K9-Mathematiktest ableiten:

H3: Die Aufgaben sind trennscharf.

H4: Die Qualität der Distraktoren ist angemessen.

H5: Die Aufgaben sind intern konsistent.

Aus dem Argument, den Annahmen und den Hypothesen für die Schlussfolgerung *Bewertung* lassen sich das unten stehende Argumentationsschema erstellen

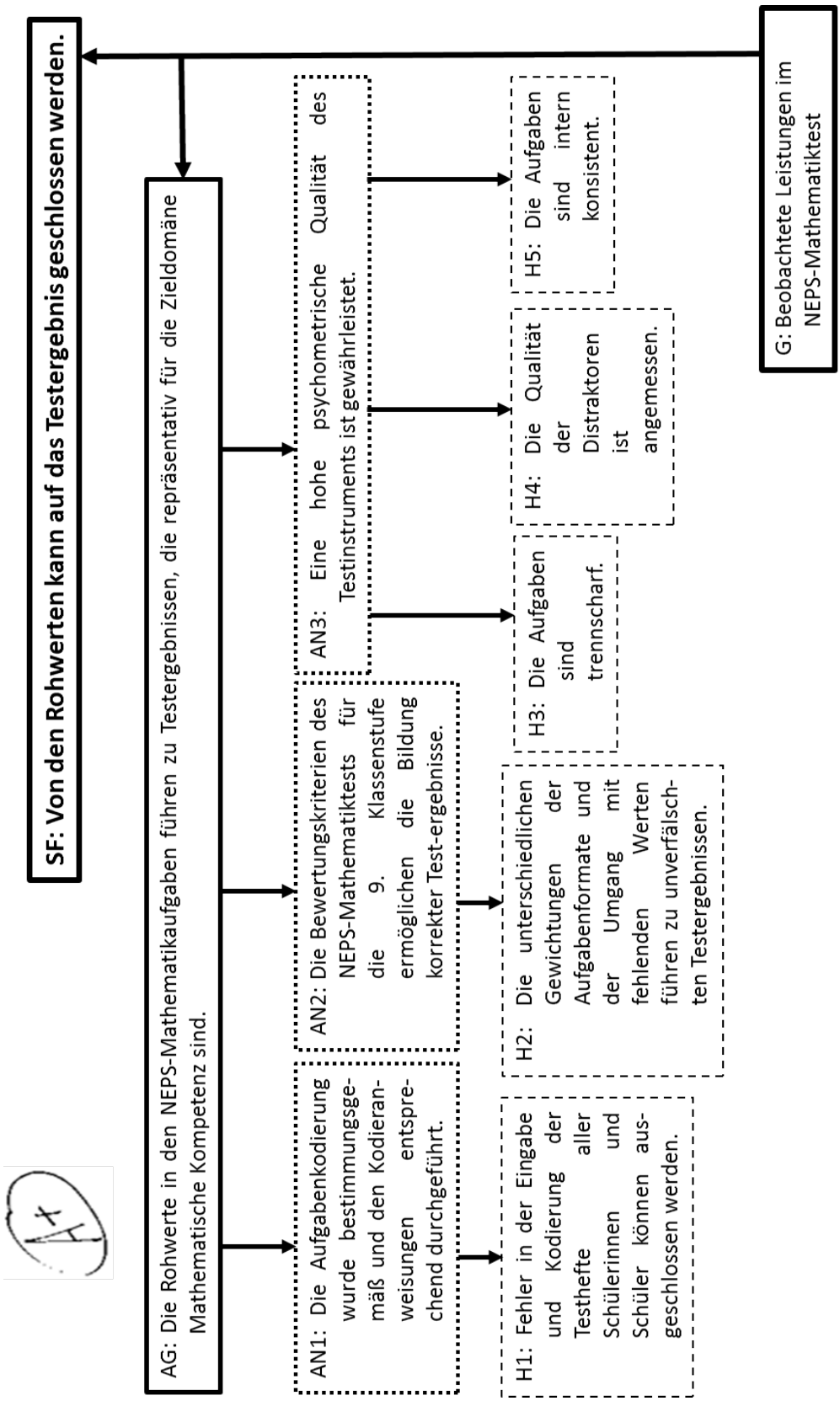
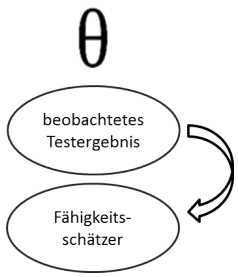


Abbildung 17: Argumentationsschema der *Bewertung*

2.3 Skalierung



Die Schlussfolgerung *Skalierung* lautet: **„Von den Testergebnissen kann auf Fähigkeiten auf einer latenten Skala geschlossen werden“** (siehe auch Abbildung 13).

Die Argumentationskette Kanes wird in der vorliegenden Arbeit um diese Schlussfolgerung erweitert, da die bewerteten Aufgaben im NEPS-Mathematiktest mittels IRT skaliert werden. Die Transformierung der beobachteten Ergebnisse auf skalierte Ergebnisse erweitert die Interpretationen der Testergebnisse. So kann für eine Schülerin beziehungsweise einen Schüler mit einem bestimmten Testscore eine Wahrscheinlichkeit vorausgesagt werden, mit welcher diese beziehungsweise dieser bestimmte Items korrekt lösen kann (Kane, 2013). Die skalenbasierte Schlussfolgerung basiert auf folgendem Argument:

AG: Das Testergebnis führt zu Fähigkeitsschätzern, welche die mathematische Kompetenz der Schülerinnen und Schüler widerspiegeln.

2.3.1 Modellpassung

Kane hebt bei dieser Schlussfolgerung hervor, dass die Genauigkeit der Fähigkeitsschätzer von der Passung der Daten zum Modell abhängig ist. Auch Yen und Fitzpatrick (2006) schreiben, dass die Qualität der Fähigkeitsschätzer von der Modellpassung des IRT-Modells zu den Daten abhängt. In diesem Zusammenhang müssen die Annahmen der Rasch-Homogenität erfüllt sein. Bei gegebener Rasch-Homogenität messen alle Aufgaben dieselbe Fähigkeit, das heißt die Itemparameter sind stichprobenunabhängig und lokal stochastisch unabhängig und alle Aufgaben haben die gleiche Trennschärfe im Sinne von Steigungsparametern. Außerdem weist der Itemfit bei Rasch-Homogenität eine gute Passung zum Modell auf (Yen & Fitzpatrick, 2006; Moosbrugger, 2012). Für das Argument der skalenbasierten Schlussfolgerung muss daher die folgende Annahme aufgestellt werden:

AN: Das verwendete IRT-Modell passt zu den Daten.

Nachstehend werden Studien zusammengefasst, welche die verwendeten IRT-Modelle auf ihre Angemessenheit untersucht haben. Senkbeil et al. (2013) prüften unter anderem die Skalierbarkeit und Passung des gewählten eindimensionalen IRT-Modells des Tests zur Erfassung technologischer und informationsbezogener Literacy (TILT) für Jugendliche am Ende der Sekundarstufe I. Sie kontrollierten beispielsweise die Passung der Items zu den Annahmen des Einparameter-Logistischen (1PL)-Modells nach Rasch, indem sie die gewichteten Abweichungsquadrate (WMNSQ; weighted mean squares) heranzogen. Außerdem analysierten die Autoren die Fairness des Tests, indem sie den Test auf Differentielles Item Funktionieren (DIF) für Geschlecht und Schultyp untersuchten. Eine ungenügende Modellpassung wurde in 6 der 140 Aufgaben nachgewiesen. In 7% der Aufgaben konnte DIF für Geschlecht und in 23% für Schultyp detektiert werden. Alle Aufgaben mit ungenügender Modellpassung und signifikantem DIF wurden aus dem Test ausgeschlossen. Um die Angemessenheit der Skalierung mit dem 1PL-Modell nach Rasch zu prüfen, verglichen die Autoren dieses mit einem Zweiparameter-Logistischen (2PL)-Modell. Für die Beurteilung wurden die Informationskriterien Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) und Consistent Aiken Information Criterion (CAIC) verwendet. Die Autoren fanden eine akzeptable Passung der Items mit dem Rasch-Modell. Das 1PL-Modell zeigte bessere BIC- und CAIC-Werte und wurde damit bestätigt. Monseur, Baye, Lafontaine und Quittre (2011) untersuchten, ob die lokale stochastische Unabhängigkeit (LSU) als Voraussetzung für die Rasch-Skalierung der Leseaufgaben aus PISA 2000 und der Mathematikaufgaben aus PISA 2003 gegeben war. Für die Aufdeckung lokaler stochastischer Abhängigkeit (LSA) zwischen den Aufgaben verwendeten die Autoren die Q3-Statistik von Yen (1984), welche auf den residualen Pearson-Produkt-Moment-Korrelationen basiert. Sie fanden in einem großen Teil der Mathematik- und Leseaufgabenverbunde moderate LSA bezüglich des globalen Kontextes der Unit. Außerdem fanden sie in einigen Mathematik- und Leseaufgaben substantielle LSA beruhend auf spezifischen paarweisen Abhängigkeiten. Die Autoren wiesen darauf hin, dass die Verletzung der LSA-Annahme in den PISA-Aufgaben eine Überschätzung der relativen Variabilität der schwach abschneidenden Länder zur Folge hat sowie eine Unterschätzung derer der stark abschneidenden Länder (Monseur et al., 2011).

2.3.2 NEPS-Hypothesen zur Modellpassung

In den beschriebenen Studien wurden die Annahmen des verwendeten Rasch-Modells und die Passung der Aufgaben zum verwendeten Modell getestet. Auch für den NEPS-K9-Mathematiktest ist die Überprüfung solcher Annahmen von Belang. Der Test wird mit Hilfe von IRT skaliert. Für die Generierung der Item- und Personenparameter wird ein 1PL-Modell verwendet. Wird davon ausgegangen, dass das verwendete Modell zu den Daten passt, so muss die Annahme des Rasch Modells erfüllt sein. Folgende Hypothesen können daher für den NEPS-K9-Mathematiktest aufgestellt werden:

H1: Für den Test kann spezifische Objektivität festgestellt werden.

H2: Die Aufgaben des Tests sind lokal stochastisch unabhängig.

H3: Die Aufgaben haben gleiche Trennschärfen.

H4: Die beobachtete Antwortwahrscheinlichkeit weicht nicht signifikant von der mit dem Modell vorhergesagten Antwortwahrscheinlichkeit ab.

In dem unten stehenden Argumentationsschema werden das Argument, die Annahmen und Hypothesen für die Schlussfolgerung der *Skalierung* visualisiert.

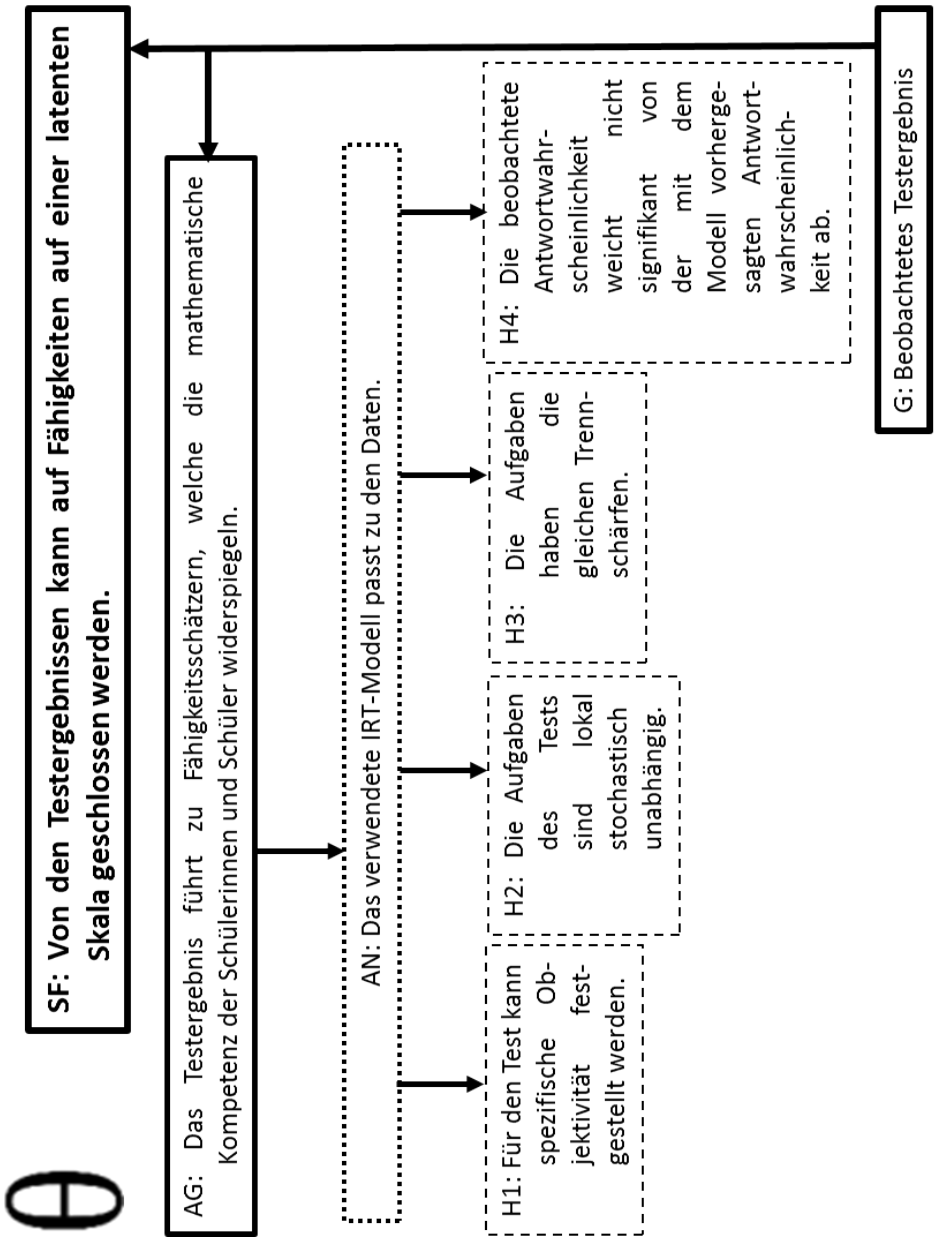
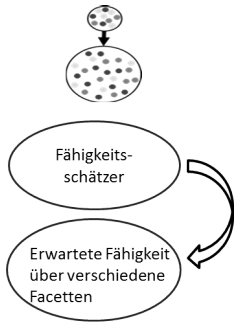


Abbildung 18: Argumentationsschema der *Skalierung*

2.4 Generalisierung



Diese Schlussfolgerung steht an vierter Stelle der Argumentationskette für den NEPS-Mathematiktest für die 9. Klassenstufe und lautet **„Die Fähigkeitsschätzer können über die Testung hinaus generalisiert werden“** (siehe auch Abbildung 13).

In den meisten Fällen soll mit einem Test nicht ausgewertet werden, wie gut die Leistung einer Testperson bezüglich spezifischer Aufgaben an einem bestimmten Tag oder in einem bestimmten Jahr war, sondern mit dem Test sollen Aussagen über die Leistung einer Testperson in einer Domäne und unter unterschiedlichen Bedingungen getroffen werden (Kane, 2013). Die Grundgesamtheit, über welche generalisiert werden soll, kann unter anderem Aufgaben, Testbedingungen, Gründe für die Testung und Rater beinhalten. Diese werden in der Generalisierbarkeitstheorie auch Facetten genannt (Kane, 2013). Die Generalisierbarkeit basiert daher auf folgendem Argument:

AG: Die Fähigkeiten auf der latenten Skala sind angemessene Schätzer für erwartete Ergebnisse in parallelen Messungen

2.4.1 Durchführungsbedingungen

Mit der Auswahl der Aufgaben eines Tests, der Durchführung zu bestimmten Zeitpunkten, dem Einsatz von bestimmten Testleitern etc. werden Stichproben der Facetten aus der Grundgesamtheit möglicher Aufgaben, Zeitpunkte, Testleiter etc. gezogen. Diese Stichproben sind selten zufällig, sondern werden oft aus praktischen Gründen gewählt. Um Fehler bei der Generalisierung zu vermeiden, sollten zum einen die Stichproben der Facetten so repräsentativ wie möglich gewählt werden. Zum anderen müssen Effekte, welche die Repräsentativität der Messbedingungen negativ beeinflussen, identifiziert und eliminiert werden. Eine Möglichkeit, solchen Effekten vorzubeugen beziehungsweise sie zu eliminieren ist, die Definition der Grundgesamtheit so anzupassen, dass bestimmte Facetten standardisiert sind (Kane, 2013). Auch Lane und Stone (2006) schreiben,

dass die Durchführungsbedingungen für alle Schülerinnen und Schüler standardisiert sein sollten, sodass alle Testergebnisse die gleichen Interpretationen zulassen. Auf diese Weise soll sichergestellt werden, dass die Variabilität in den Testleistungen lediglich durch die Fähigkeit der Schülerinnen und Schüler und nicht durch die Durchführungsbedingen beeinflusst wird (Lane & Stone, 2006; Moosbrugger, 2012). Aus dem Argument der *Generalisierung* kann somit die folgende Annahme hergeleitet werden:

AN1: Die Durchführungsbedingungen der Messung sind standardisiert.

Hierbei ist zu erwähnen, dass durch Standardisierung der Testbedingungen die Definition der Grundgesamtheit, über die generalisiert wird, nur diese standardisierten Bedingungen enthält. Die Standardisierung beugt somit unterschiedlichen Einflüssen von Durchführungsbedingungen auf die Testergebnisse vor und ermöglicht damit gleiche Interpretationen für diese Testergebnisse, jedoch sind diese Ergebnisse nur repräsentativ für die standardisierten Testbedingungen.

Im Folgenden werden Studien vorgestellt, die bereits Hypothesen zu den Durchführungsbedingungen untersucht haben. In großen Schulleistungsstudien wie dem LV und TIMSS werden die Umstände der Messung untersucht, um eine Generalisierung über diese Umstände zu rechtfertigen (Johansone, 2000; Richter et al., 2012). Im LV wurde im Rahmen eines Qualitätsmonitorings stichprobenartig die Einhaltung der Durchführungsbedingungen untersucht. Geschulte Beobachterinnen und Beobachter beobachteten und dokumentierten dabei unangekündigt den Testablauf und befragten Schulkoordinatorinnen und Schulkoordinatoren zur Vorbereitung und Durchführung der Studie. Die Ergebnisse weisen insgesamt auf eine gute Testvorbereitung durch Schulkoordinatorinnen und Schulkoordinatoren hin. Auch die standardisierten Durchführungsbedingungen wurden von den meisten Testleiterinnen und Testleitern eingehalten. Es wird davon ausgegangen, dass bei der Testdurchführung keine Probleme aufgetreten sind, die Auswirkungen auf die Interpretation der Ergebnisse haben (Richter et al., 2012). Im Rahmen der TIMS-Studie 2011 fand ein internationales Qualitätsmonitoring statt, bei dem 15 Schulen in jedem teilnehmenden Land bei der Testdurchführung observiert wurden. Insgesamt wurden die Tests den internationalen Richtlinien entsprechend durchgeführt. Die Testdurchführungsqualität wurde von den Beobachtern als gut eingeschätzt und in den meisten Testsitzungen konnten keine Probleme festgestellt werden (Johansone, 2000).

2.4.2 NEPS-Hypothesen zu den Durchführungsbedingungen

Die beschrieben Studien untersuchten die Umstände der jeweiligen Messung, indem sie die Standardisierung der Messungen prüften. Für den NEPS-K9-Mathematiktest können ebenfalls Hypothesen über die Standardisierung der Messung aufgestellt werden. Der Test wurde in verschiedenen Schulen, durch unterschiedliche Testleiter und zu unterschiedlichen Zeitpunkten durchgeführt. Um störende Einflüsse auf die Testleistung zu vermeiden, sollten Maßnahmen für die Standardisierung der Testung durchgeführt werden und diese müssen ausreichend für die Vermeidung störender Einflüsse sein (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 2014). Folgende Hypothese lässt sich für den NEPS-K9-Mathematiktest ableiten:

H1: Die Testung wurde sorgfältig anhand von angemessenen, standardisierten Prozeduren durchgeführt.

2.4.3 Messgenauigkeit

In der IRT wird die Generalisierung von Ergebnissen über Testwiederholungen mit Hilfe von Item Characteristic Curves (ICCs) und dem darauf basierenden Standardfehler zur Schätzung analysiert (Kane, 2013; Yen & Fitzpatrick, 2006). Der Standardfehler repräsentiert dabei die Variabilität über hypothetische Wiederholungen mit bestimmten Testaufgaben oder mit unterschiedlichen Aufgaben unter der Voraussetzung, dass die Ergebnisse zum Modell passen. Aus den Standardfehlern abgeleitete Konfidenzintervalle erlauben Schlüsse zu Einschränkungen der Generalisierung. Geschätzte Standardmessfehler zeigen Limitationen der Generalisierung auf. Mit Ergebnissen einer Testsitzung ist es jedoch nicht möglich, die Variabilität über Testbedingungen, Gründe für Tests oder andere Facetten zu erfassen (Kane, 2013; Yen & Fitzpatrick, 2006). Die zweite Annahme, die aus dem Argument der *Generalisierung* abgeleitet werden kann, ist daher:

AN2: Die Messgenauigkeit des Tests ist angemessen.

Die nachstehenden Studien analysierten Hypothesen zur Messgenauigkeit von Tests. Köller, Eßel-Ullmann und Paasch (2012) prüften die Zuverlässigkeit der Messungen des Tests zur standardbasierten Diagnostik mathematischer Kompetenzen, indem sie den

Kruder-Richardson-Koeffizienten als Reliabilitätsmaß für die beiden Testversionen berechneten. Außerdem untersuchten sie die Messgenauigkeit des Tests in verschiedenen Leistungsbereichen durch eine Analyse der Standardfehler für die Personenparameter. Insgesamt fanden die Autoren zufriedenstellende Kruder-Richardson-Koeffizienten zwischen KR-20 = .80 und KR-20 = .86. Die Analyse der Standardfehler ergab, dass beide Testversionen im oberen bis mittleren Leistungsbereich präziser definieren als im unteren Leistungsbereich. Auch für den Florida Comprehensive Assessment Test (FCAT) in Mathematik und Lesen wurde die Messgenauigkeit des Tests untersucht. In dieser Studie wurden die Tests in Lesen und Mathematik für die Klassenstufen 3 bis 10 analysiert. Die Daten wurden für MC-Aufgaben mit einem Dreiparameter-Logistischen (3PL)-Modell und für die übrigen Formate mit einem 2PL-Partial Credit Modell skaliert. Anschließend wurden die Standardmessfehler berechnet und konditionelle Standardfehlerkurven betrachtet. Zusätzlich wurde aus dem mittleren Standardmessfehler aller Schülerinnen und Schüler ein Maß für die Homogenität des Tests generiert, welches ähnlich wie Cronbach Alpha interpretiert werden kann. Die Autoren fanden interne Konsistenz um .90, die ähnlich hoch war wie in anderen landesweiten Tests. Die Standardfehlerkurven zeigten, dass individuelle Testergebnisse in der Mitte der Verteilung weniger variierten als an den Rändern der Verteilung.

2.4.4 NEPS-Hypothesen zur Messgenauigkeit

Die Unsicherheiten, die bei der Generalisierung auftreten, können auch statistisch quantifiziert werden. So ist es wichtig, die Reliabilität der individuellen Fähigkeitsmessungen zu bestimmen (Rost, 1996). Daher können folgende Hypothesen formuliert werden:

H2: Die individuellen Fähigkeitsmessungen sind reliabel.

In dem unten stehenden Argumentationsschema werden das Argument, die Annahmen und die Hypothesen für die Schlussfolgerung der *Generalisierung* visualisiert.

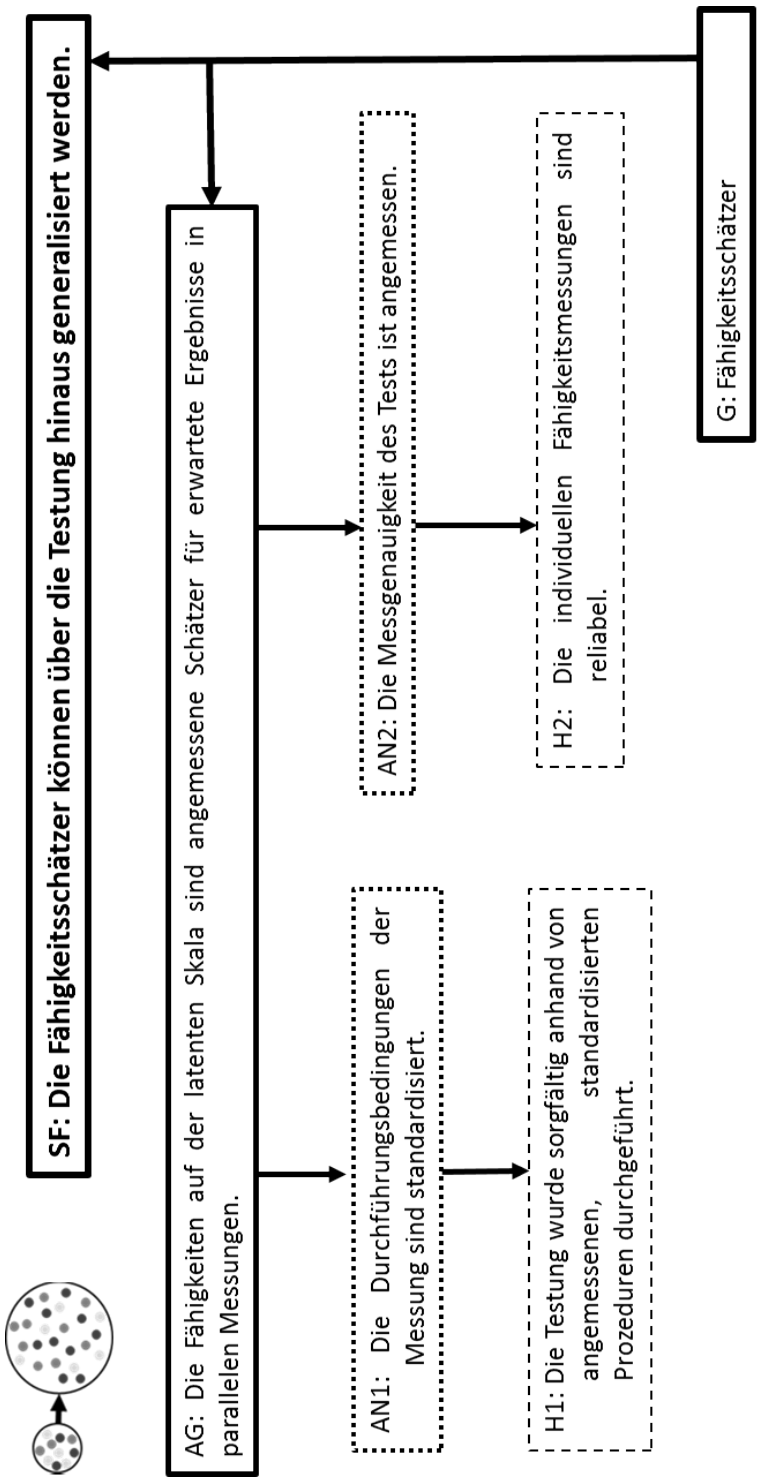
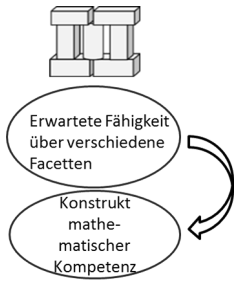


Abbildung 19: Argumentationsschema der *Generalisierung*

2.5 Konstruktbezug



Der *Konstruktbezug* steht zusammen mit der *Extrapolation* an vierter Stelle der Argumentationskette. Diese Schlussfolgerung lautet: **„Vom (generalisierten) Testergebnis des NEPS Mathematiktests für die 9. Klassenstufe lässt sich auf das Konstrukt der mathematischen Kompetenz, wie es im NEPS-Rahmenkonzept. definiert wird, schließen.“** (siehe auch Abbildung 13).

Wenn der Test auf einem theoretischen Konstrukt basiert, sollte dieses Konstrukt kritisch auf Plausibilität untersucht werden. Die Konstrukttheorie beinhaltet die Annahmen für die konstruktbasierte Schlussfolgerung (Kane,2012, 2004). Die Schlussfolgerung *Konstruktbezug* basiert daher auf folgendem Argument:

AG: Die (generalisierten) Ergebnisse im NEPS-Mathematiktest für die neunte Klasse sind auf das Konstrukt der mathematischen Kompetenz zurückzuführen.

2.5.1 Dimensionale Struktur

Kane bezieht sich in seinen Beschreibungen dieser Schlussfolgerung nicht speziell auf die dimensionale Struktur des Konstruktes. Jedoch schreibt er, dass eine Schlussfolgerung bezüglich des *Konstruktbezuges* Nachweise für das Konstrukt und für die Beziehungen zwischen der beobachteten Testleistung und dem Konstrukt benötigt (Kane, 2002). Das Konstrukt eines Tests kann eine einzige Dimension beinhalten, aber auch mehrere homogene und trotzdem separierbare Dimensionen. Die Validität der Testwertinterpretation wird durch das Ausmaß, in welchem die Annahmen des Rahmenkonzeptes durch Wechselbeziehungen der Aufgaben und Testkomponenten repräsentiert werden, beeinflusst (Standards, 2014). Eine Annahme, die sich aus dem Argument des *Konstruktbezuges* ableitet, ist folgende:

AN: Die angenommene dimensionale Struktur des NEPS-Mathematikkonstrukts lässt sich analytisch bestätigen.

Nachstehend werden Studien zusammengefasst, welche die dimensionalen Strukturen von Tests prüften. Winkelmann, Robitzsch, Stanat und Köller (2012) analysierten bei-

spielsweise die Dimensionalität des Tests zum Erreichen der Bildungsstandards in der Primarstufe, indem sie unter anderem verschiedene konkurrierende dimensionale Modelle prüften. Die Autoren berechneten zwei einparametrische Modelle, wovon eines der Annahme der Within-Item-Dimensionality unterlag und eines der Annahme der Between-Item-Dimensionality. Zusätzlich berechneten sie ein zweiparametrisches Modell mit der Annahme der Within-Item-Dimensionailty. Die Autoren verglichen die Korrelationen zwischen den Inhaltsbereichen in diesen Modellen mit Korrelationen bei randomisierter Itemklassifikation. Aus den Korrelationsmustern schlossen die Autoren, dass die Annahme eines einparametrischen Modells mit Within-Item-Dimensionality in vielen Fällen nicht plausibel sei. Um die Homogenität der inhaltlichen Kompetenzbereiche zu überprüfen, führten die Autoren nicht parametrische Dimensionalitätsanalysen mithilfe der DETECT-Statistik aus. Die Ergebnisse lieferten Hinweise für die Eindimensionalität der Inhaltsbereiche, deckten aber auch kleinere Verletzungen dieser Eindimensionalität auf. Auch für die prozessbezogenen Kompetenzen wurden die Korrelationen in einem Modell mit Within-Item-Dimensionality berechnet. Die Ergebnisse wiesen auf eine analytische Separierbarkeit der Kompetenzen hin. Für einen Vergleich der verschiedenen Modelle wurden die Modellgütekriterien AIC, BIC und CAIC für ein eindimensionales, globales Modell, für ein Modell der inhaltsbezogenen Kompetenzen mit Within-Item-Dimensionality und für eines mit Between-Item-Dimensionality sowie für ein Modell mit den prozessbezogenen Kompetenzen unter Annahme der Within-Item-Dimensionality berechnet. Die Modellgütekriterien wiesen die besten Werte für ein Modell mit den fünf inhaltlichen Kompetenzen unter Annahme der Between-Item-Dimensionality auf (Winkelmann et al., 2012). Shaw und Crisp (2012) untersuchten die Dimensionalität des International A Level Physics Tests. Um zu kontrollieren, ob die Testhefte jeweils ein einziges Merkmal oder mehrere unterschiedliche Merkmale messen, führten sie separate explorative Faktorenanalysen für 5 Testhefte durch. Maximal 4600 und minimal 209 achtzehnjährige Schülerinnen und Schüler bearbeiteten jeweils eines der Testhefte. Die Autoren fanden für zwei Testhefte eine Eindimensionalität, für zwei Testhefte zwei gemessene Faktoren und für ein Testheft drei Faktoren. Insgesamt ließen sich die unterschiedlichen Faktoren durch die Autoren sinnvoll erklären.

2.5.2 NEPS-Hypothesen zur dimensionalen Struktur

Die beschriebenen Studien prüften die Dimensionalität der Tests empirisch, um die Ergebnisse dann mit den theoretischen Annahmen aus dem Konstrukt abzugleichen. Auch für den NEPS-K9-Mathematiktest können Hypothesen zur Dimensionalität des Tests aus dem Konstrukt abgeleitet werden. Für die mathematische Kompetenz des NEPS werden im Rahmenkonzept zwar Teildimensionen in Form von Inhaltsbereichen und Prozessen beschrieben, dennoch wird das Konstrukt als eindimensional beschrieben und in der Skalierung auch eindimensional modelliert. Wird also von der Annahme ausgegangen, dass sich die dimensionale Struktur des Tests analytisch bestätigen lässt, so ist auch davon auszugehen, dass die Teildimensionen sehr stark zusammenhängen (r= >.95, Carstensen, 2013) und eine eindimensionale Skalierung am besten zu den Daten passt, da mit den Dimensionen eine einzige mathematische Kompetenz gemessen wird. Folgende Hypothesen lassen sich für den NEPS-K9-Mathematiktest formulieren:

H1: Innerhalb der NEPS-Teildimensionen können sehr hohe Zusammenhänge gefunden werden.

H2: Für den NEPS-Mathematiktest ist eine eindimensionale Skalierung einer mehrdimensionalen Skalierung vorzuziehen.

Das NEPS-Rahmenkonzept definiert für die Domäne Mathematik ein eigenes Konstrukt mathematischer Kompetenz, welches sich von den anderen in NEPS gemessenen Kompetenzen unterscheidet. Für den NEPS-K9-Mathematiktest kann daher angenommen werden, dass dieser sich dimensional von den anderen Kompetenztests des NEPS im Lesen, in den Naturwissenschaften und in ICT abgrenzen lässt.

H3: Für eine mehrdimensionale Skalierung der mathematischen, naturwissenschaftlichen, Lese- und ICT-Kompetenzen aus dem NEPS auf separaten Dimensionen wird eine bessere Passung gefunden als bei einer eindimensionalen Skalierung der Kompetenzen auf einer gemeinsamen Dimension.

Auf Grund der großen Nähe der naturwissenschaftlichen und mathematischen Rahmenkonzeptionen zwischen NEPS und PISA [vgl. Kapitel 1.1] (van den Ham, Nissen, Ehmke, Sälzer & Roppelt, 2014; Wagner, Schöps, Hahn, Pietsch & Köller, 2014) werden die Zusammenhänge zwischen den Literacy-Domänen Mathematik und Naturwissenschaften im NEPS in einer Höhe erwartet, die auch zwischen den Literacy-Domänen Mathematik

und Naturwissenschaften in PISA nachgewiesen wurde [$r =$. 88] (OECD, 2012). In der PISA-Studie wird außerdem ein kleinerer Zusammenhang zwischen den Kompetenzen Lesen und Mathematik ($r = .85$) gefunden als zwischen den mathematischen und naturwissenschaftlichen Kompetenzen ($r = .89$) (OECD, 2014). Die folgende Hypothese kann aufgestellt werden:

H4: Die mathematische Kompetenz im NEPS hängt höher mit der naturwissenschaftlichen Kompetenz zusammen als mit der Kompetenz Lesen.

Für die Schlussfolgerung des *Konstruktbezugs* kann aus dem Argument, den Annahmen und den Hypothesen untenstehendes Argumentationsschema erstellt werden.

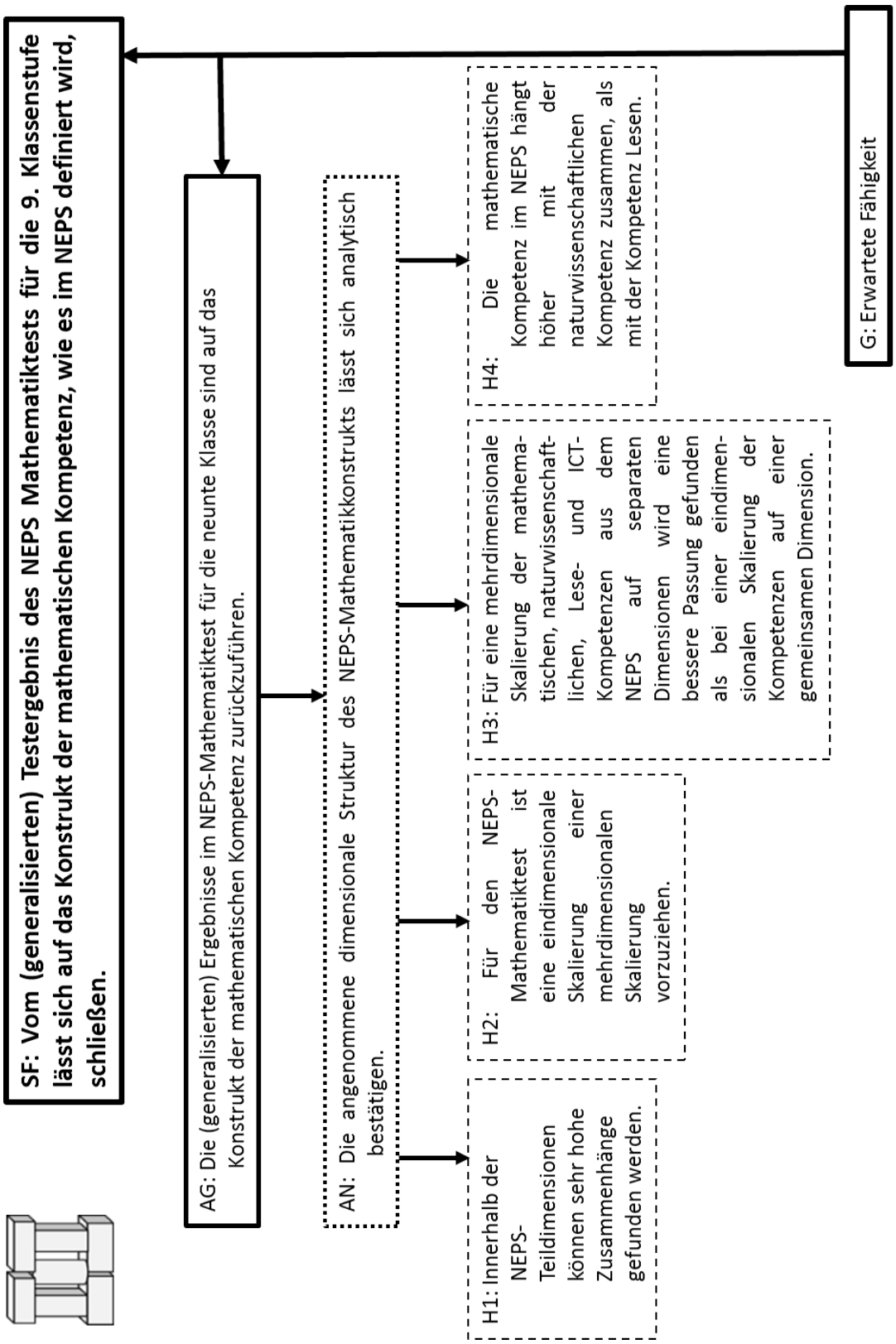
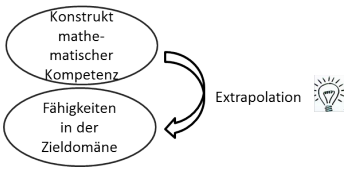


Abbildung 20: Argumentationsschema des *Konstruktbezugs*

2.6 Extrapolation



Die Schlussfolgerung der *Extrapolation* lautet: **„Von dem NEPS-Mathematiktestergebnis als Repräsentation des Konstrukts kann auf mathematische Kompetenz in der Zieldomäne geschlossen werden.“**(siehe auch Abbildung 13).

Die Testergebnisse sollen in der Regel als Leistung in der Zieldomäne interpretiert werden und beinhalten damit auch eine Erwartung für zukünftige Leistungen in der Domäne (Kane, 2006). Dabei ist zu beachte, dass die Zieldomäne Mathematische Kompetenz, wie sie im NEPS-Rahmenkonzept definiert wird, über die Testsituation hinausgeht und sich auf die mathematische Kompetenz im Alltag bezieht. Die Schlussfolgerung basiert dementsprechend auf dem folgenden Argument:

AG: Die Kompetenz der Zieldomäne, wie sie mit dem NEPS-Test erfasst wird, ist ein Indikator für die Leistung in der Zieldomäne Mathematische Kompetenz

2.6.1 Zusammenhang mit Kriterien mathematischer Kompetenz

Aus dem Argument der *Extrapolation* lässt sich ableiten, dass eine leistungsstarke Testperson auch vergleichbare Aufgaben in der Realsituation bewältigen kann (Kane, 2013). Mit dem Test soll also die Leistung in der Zieldomäne korrekt vorausgesagt werden können (Standards, 2014). Eine Annahme, die sich aus dem Argument der *Extrapolation* ableiten lässt, ist daher folgende:

AN: Die Testleistung im NEPS-Mathematiktest für die neunte Klasse hängt mit der mathematischen Kompetenz in der Zieldomäne zusammen.

Der Zusammenhang eines Tests mit Leistungen in der Zieldomäne wurde bereits mehrfach in anderen Studien untersucht. Stalder, Meyer und Hupka-Brunner (2008) analysierten beispielsweise mittels deskriptiver und multivariater Analysen den Zusammenhang zwischen den Lesekompetenzen bei PISA und der Bildungslaufbahn in der Sekundarstufe II in der Schweiz. Die Datengrundlage bildete eine an PISA 2000 gekoppelte

Längsschnittstudie TREE, in der 6343 Schülerinnen und Schüler nach ihrer Teilnahme an PISA 2000 bis 2007 jährlich befragt wurden. Durch jährlichen Ausfall nahmen noch 3982 Personen in der letzten Erhebungswelle teil. Die Autoren fanden heraus, dass Jugendliche mit mittleren und höheren Lesekompetenzen häufiger eine anspruchsvollere Ausbildung beginnen und einen höheren Bildungsabschluss erreichen als Jugendliche mit niedrigeren Kompetenzen. Jedoch beginnt mehr als die Hälfte der Jugendlichen mit Lesekompetenzen unterhalb der ersten Kompetenzstufe in PISA 2000 ohne Verzögerung eine Berufsausbildung. Die durch PISA formulierte potentielle Risikogruppe von Schülerinnen und Schülern mit Kompetenzen unterhalb der ersten Kompetenzstufe kann nicht belegt werden. Köller et al. (2012) untersuchten im Rahmen der Validierung des Tests zur standardbasierten Diagnostik mathematischer Kompetenzen in der Primarstufe unter anderem die Zusammenhänge des Tests mit Zeugnisnoten, Schulform und anderen Leistungsmaßen. Insgesamt wurden in der Studie 687 Fünftklässlerinnen und Fünftklässler in unterschiedlichen Schulformen am Anfang des Schuljahres getestet. Die Autoren fanden eine höhere Korrelation des Testes mit der Mathematiknote als mit der Deutschnote. Im Mittel erzielten die Gymnasiastinnen und Gymnasiasten höhere Testwerte als Schülerinnen und Schüler anderer Schulformen. Die Mathematikleistung ist außerdem in einem Probit-Regressionsmodell mit einem Pseudo-R^2 zwischen .29 und .28 ein statistisch signifikanter Prädiktor für die besuchte Schulform. Für die testbasierten Messungen mit dem Intelligenztest, dem Lesetest und dem Orthographietest wurden substantielle Korrelationen zwischen $r = .50$ und $r = .65$ nachgewiesen. Aus diesen Ergebnissen schlossen die Autoren, dass der Test Potenziale für die Schullaufbahndiagnostik und Lernausgangslagenuntersuchung aufweist, jedoch eine mögliche Neuauflage des Tests auch einige Austauschaufgaben enthalten sollte.

2.6.2 NEPS-Hypothesen zum Zusammenhang mit Kriterien mathematischer Kompetenz

Die beschriebenen Studien leiten aus dem jeweiligen Testkonstrukt Annahmen über die Zusammenhänge mit Kriterien für die Kompetenz in der Zieldomäne ab und prüfen Hypothesen über die Zusammenhänge der Testergebnisse mit diesen Kriterien. Auch für den NEPS-K9-Mathematiktest können Hypothesen über die Zusammenhänge mit Kriterien aufgestellt werden.

Mathematische Kompetenz im NEPS beschreibt unter anderem das Ausmaß, in dem Schülerinnen und Schüler, aber auch Erwachsene, die in der Schule gelernte Mathematik in problemhaltigen, vorwiegend außermathematischen Situationen flexibel anwenden können (Ehmke et al., 2009). Da die Mathematiknote eine Bewertung der in der Schule gelernten mathematischen Kompetenz ist, kann von einem substantiellen Zusammenhang der Mathematiknote mit der Testleistung ausgegangen werden. Das Konstrukt mathematischer Kompetenz im NEPS umfasst durch den Literacy-Aspekt jedoch mehr als nur in der Schule gelernte Fähigkeiten. Diese Tatsache sollte sich in der Höhe der Korrelation abzeichnen. Des Weiteren ist zu erwarten, dass die Zusammenhänge der mathematischen Testleistung mit Schulnoten aus anderen Fächern geringer sind, da in diesen Fächern andere Kompetenzen gefördert und benotet werden. Auf Basis des mathematischen Konstruktes des NEPS kann die nachfolgende Hypothese abgeleitet werden.

H1: Die Leistung der Schülerinnen und Schüler im NEPS-Mathematiktest hängt stärker mit der Mathematiknote zusammen als mit Zeugnisnoten anderer Fächer.

Die Zieldomäne Mathematischer Kompetenz im NEPS wird angelehnt and das PISA-Rahmenkonzept, das auf dem mathematischen Literacy-Konzept basiert. Jedoch sind auch die wichtigsten curricularen Inhalte aus dem LV-Rahmenkonzept in der Definition der Zieldomäne enthalten. Hohe Kompetenzen im Sinne des NEPS werden durch hohe Kompetenzstufen in PISA sowie im LV widergespiegelt. Die Messungen mathematischer Kompetenz aus PISA und dem LV kann daher als Indikator für die mathematische Kompetenz in der Zieldomäne interpretiert werden.

H2: Die mathematische Kompetenz im NEPS hängt stark mit den mathematischen Kompetenzwerten, gemessen durch PISA und LV, zusammen.

H3: Die mathematische Kompetenz im NEPS hängt deutlich niedriger mit den naturwissenschaftlichen Kompetenzwerten, gemessen durch PISA und LV, zusammen.

Studien haben gezeigt, dass mathematische Schülerleistung einen spezifischen Faktor mathematischer Kompetenz beinhaltet, dass sich die Leistungsvarianz jedoch auch durch einen Faktor kognitiver Grundfähigkeit erklären lässt (Brunner, 2006; Köller et al., 2012; Winkelmann et al., 2012). Auch für den PISA-Mathematiktest, welcher dem NEPS-Test konzeptionell sehr nahe ist und ebenfalls das Ziel hat, Mathematical Literacy zu messen, wird ein Zusammenhang mit kognitiver Fähigkeit zwischen $r = .74$ gefunden (Prenzel et al., 2004). Kognitive Fähigkeit kann aus diesen Gründen als Kriterium für

mathematische Kompetenz in der Zieldomäne betrachtet werden und es wird ein ähnlich hoher Zusammenhang mit den Kompetenzen aus dem NEPS-K9-Mathematiktest wie mit dem PISA-Mathematiktest erwartet. So lässt sich für den NEPS-Mathematiktest folgende Hypothese aufstellen:

H4: Die Testergebnisse im NEPS-K9-Mathematiktest hängen mit dem Kriterium kognitive Fähigkeiten zusammen.

Metakognitives Wissen ist nicht nur ein Prädiktor für kognitive Leistungen, sondern auch für Mathematikleistung (Lingel, Neuenhaus, Artelt & Schneider, 2014). In der PISA-Studie aus 2003 wurde für eine nationale Substichprobe eine Korrelation von $r = .43$ zwischen dem Strategiewissen in Mathematik und der Mathematical Literacy nachgewiesen (Schneider & Artelt, 2010). Lingel, Neuenhaus und Schneider fanden Korrelationen von $r = .38$ und $r = .39$ zwischen der mit dem DEMAT 5+ gemessenen Mathematikleistung und metakognitivem Wissen bei Fünftklässlerinnen und Fünftklässlern (Lingel et al., 2014). Metakognitives Wissen kann daher als Kriterium für mathematische Kompetenz angenommen werden. Folgende Hypothese wird abgeleitet:

H5: Die Testergebnisse im NEPS-K9-Mathematiktest hängen mit dem Kriterium deklarative Metakognition zusammen.

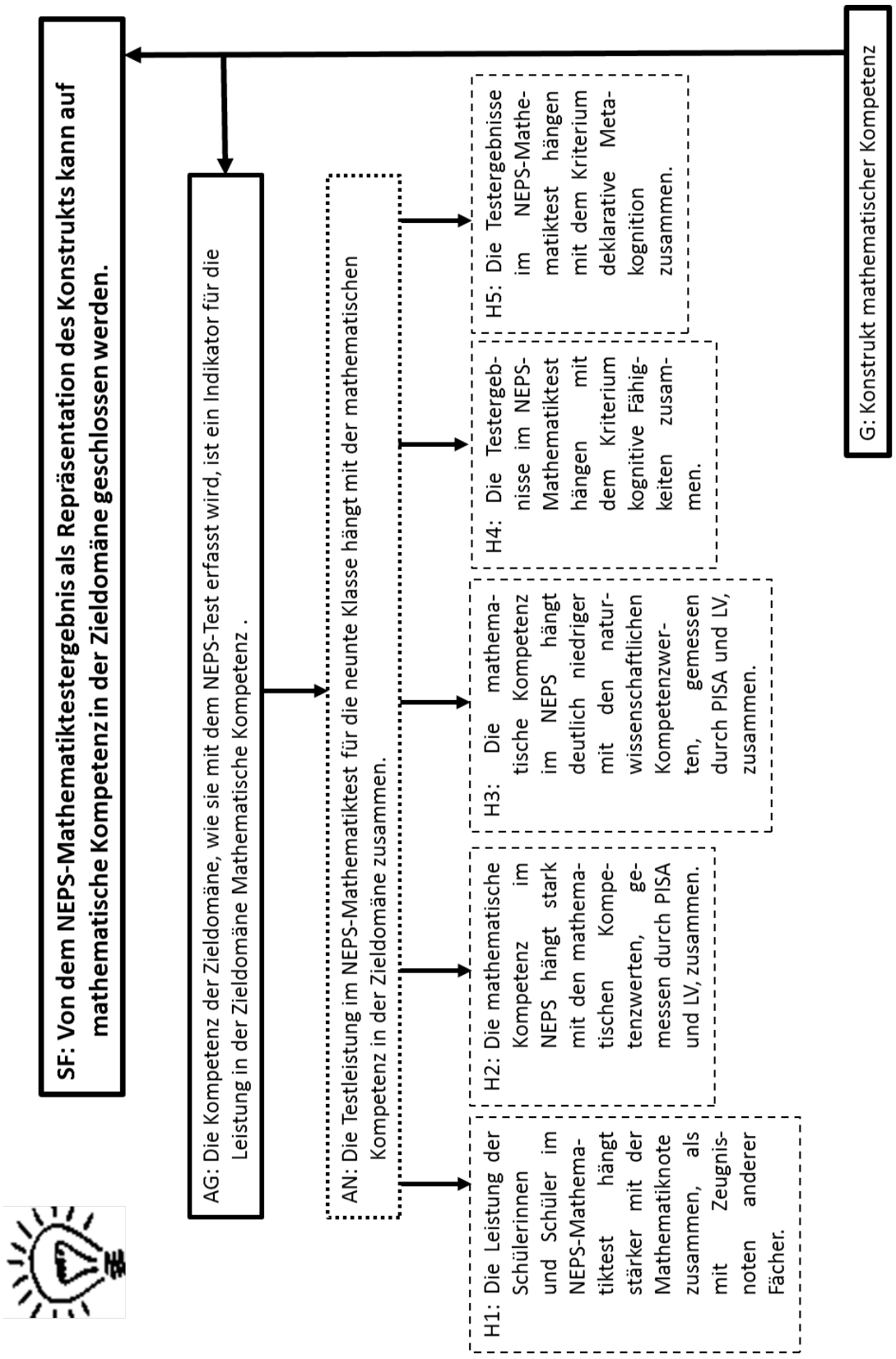


Abbildung 21: Argumentationsschema der *Extrapolation*

2.7 Entscheidung

Die hier vorgestellte Argumentationskette dient zur Validierung der folgenden Interpretation der Testergebnisse:

> Die Unterschiede in der beobachteten Testleistung geben die Unterschiede in der mathematischen Fähigkeit wieder, wie diese im Rahmenkonzept definiert wird.

Diese Testverwendung und Interpretation bleibt auf deskriptiver Ebene und beinhaltet keine Entscheidungsregeln. Für diese Interpretation wird die Schlussfolgerung *Entscheidung* daher nicht in die Argumentationskette aufgenommen. Die mit dem NEPS-Test erhobenen Daten werden Wissenschaftlern und Wissenschaftlerinnen als *Scientific-Use-Files* zu Verfügung gestellt. Sollten Analysen in diesem Kontext mehr als die oben genannte deskriptive Interpretation beinhalten, so müssen diese speziell evaluiert werden.

2.8 Fazit

Die hier vorgestellte Argumentationskette bildet die Grundlage für die Validierung der Testwertinterpretationen des NEPS-Mathematiktests für die neunte Klasse. Jede Schlussfolgerung erweitert die Interpretation der Testergebnisse von der beobachteten Leistung der Schülerinnen und Schüler bis hin zur Leistung in der Zieldomäne Mathematische Kompetenz. Die Schlussfolgerungen basieren jeweils auf einem Argument, welches wiederum durch Annahmen gestützt wird. Aus den Annahmen können konkrete Hypothesen für den NEPS-K9-Mathematiktest abgeleitet werden. Im nächsten Schritt werden diese Hypothesen geprüft und somit ein Validitätsargument für die Testwertinterpretation des NEPS-K9-Mathematiktests gebildet. Bevor jedoch die Prüfung der Hypothesen vorgestellt wird, sollen die Datengrundlage der Studie, das Design und grundlegende Methoden beschrieben werden.

3 Studienbeschreibung

Für die Validierung der Testwertinterpretationen des NEPS-K9-Mathematiktests werden Daten aus zwei Studien herangezogen. Zum einen werden die Daten der originalen NEPS-Erhebung aus 2010 für die neunte Jahrgangsstufe verwendet. Diese wurden nach Abschluss eines Datennutzervertrages in Form eines *scientific use files* von der Website des NEPS heruntergeladen. Dieser *Scientific Use File* ist ein Datensatz mit hohem Anonymisierungsgrad. Sensible Angaben wie zum Beispiel geographische Angaben auf Bundesländerebene oder offene Textfelder sind in diesem Datensatz nicht enthalten. Andere Informationen wie beispielsweise Angaben zur Staatsangehörigkeit oder Landessprache sind in dem Datensatz aggregiert. Mit diesen Daten können jedoch nicht alle Forschungsfragen beantwortet werden. Für die Untersuchung der Zusammenhänge des NEPS-Mathematiktests mit anderen Messungen mathematischer und nicht mathematischer Kompetenz wurde eine Validierungsstudie durchgeführt. In dieser Studie wurden neben den NEPS-Instrumenten zur Erfassung der mathematischen und naturwissenschaftlichen Kompetenz unter anderem Aufgaben aus PISA und dem LV zur Messung der mathematischen und naturwissenschaftlichen Kompetenz sowie ein kognitiver Fähigkeitstest verwendet. Im Folgenden wird zuerst das Design der Haupterhebung und anschließend das Design der Validierungsstudie beschrieben. Abschließend werden einige grundlegende Methoden für die Analysen des Interpretation/Use Argument (IUA) dargelegt. Die konkreten Methoden zur Auswertung der einzelnen Schlussfolgerungen und den dazugehörigen Hypothesen werden im Kapitel 4, vorausgehend zu den jeweiligen Ergebnissen, dargelegt.

3.1 Studie 1: NEPS 2010

Bevor das Design und die Stichprobe der Erhebung aus 2010 beschrieben werden, die den Analysen in dieser Arbeit zu Grunde liegt, wird ein kurzer Überblick über die Stichprobenziehung der NEPS-Erhebung gegeben.

3.1.1 Stichprobenziehung in NEPS 2010

Im Herbst 2010 wurden im NEPS erstmalig Daten in der neunten Jahrgangsstufe (Startkohorte 4) erhoben. Die Stichprobenziehung wurde in Zusammenarbeit des Leibniz-Instituts für Bildungsverläufe e.V. (LIfBi) und dem infas Institut für angewandte Sozialwissenschaft Bonn (zuständig für die Eltern-Befragung) durchgeführt. Die Ausgangsstichprobe für die erste Erhebung bestand aus Schülerinnen und Schülern der neunten Klasse an sowohl Regel- als auch Förderschulen. Für beide Schularten wurden getrennte Stichproben gezogen. In dieser Untersuchung wird lediglich die Stichprobe der Regelschulen betrachtet. Für die Zielpopulation der Schülerinnen und Schüler an Regelschulen wurde eine geschichtete Klumpenstichprobe gezogen. Innerhalb der so ausgewählten neunten Klassen wurden alle Schülerinnen und Schüler zur Teilnahme eingeladen. Die Realisierung der Stichprobe ist im Methodenbericht A46_A47_A83 dargestellt. Eine Stichprobenbeschreibung wird in Tabelle 4 gegeben.

Tabelle 4: Schüleranzahl, Geschlechterverteilung und Teilnahmestatus in Abhängigkeit der Schulform für NEPS 2010

Schulform	Schülerzahl	Anzahl an Schulen	Anteil der Jungen (%)	Teilnahme (%)
Hauptschule	3805	181	56	93.8
Schule mit mehreren Bildungsgängen	1190	56	51.5	94.7
Realschule	3249	104	51.5	95.7
Integrierte Gesamtschule	1703	55	45.5	95
Gymnasium	5292	149	45.4	96.7
Gesamt	15239	545		

3.1.2 Testdesign

In der NEPS-Erhebung von Neuntklässlerinnen und Neuntklässlern im Jahr 2010 wurden, wie auch in anderen Erhebungen des NEPS, unterschiedliche Kompetenzdomänen gleichzeitig getestet. Um möglichen Positionseffekten durch die Reihenfolge der Kompetenzdomänen in einem Testheft vorzubeugen, wurde die Reihenfolge der Domänen rotiert. Nach einer ersten zufälligen Zuordnung der Schülerinnen und Schüler zu den Testheften wurde die Reihenfolge der Kompetenzdomänen für nachfolgende Erhebungswellen fixiert. Schülerinnen und Schüler, die zum Beispiel in der ersten Erhebung die Mathematikaufgaben vor den Leseaufgaben bearbeitet hatten, werden in der folgenden längsschnittlichen Erhebung die Domänen in ebenfalls dieser Reihenfolge bearbeiten. In der Erhebung von Neuntklässlerinnen und Neuntklässlern im Jahr 2010 gab es zwei Testhefte. Im ersten Testheft wurde ICT-Literacy an erster Stelle, naturwissenschaftliche Kompetenz an zweiter Stelle, Lesegeschwindigkeit an dritter Stelle, mathematische Kompetenz an vierter Stelle, Hörverstehen auf Wortebene an fünfter und domänenspezifische Metakognitionen an letzter Stelle getestet. Im zweiten Testheft wurde die naturwissenschaftliche Kompetenz an erster, ICT-Literacy an zweiter, Lesegeschwindigkeit an dritter, mathematische Kompetenz an vierter, Hörverstehen auf Wortebene an vorletzter und domänenspezifische Metakognitionen an letzter Stelle getestet. Die Testung erfolgte normalerweise in einem Klassenraum mit Einzelsitzplätzen. Die reine Bearbeitungszeit der Testhefte betrug 112 Minuten. Es gab eine Pause von 15 Minuten vor der Lesegeschwindigkeitstestung (Pohl & Carstensen, 2012). Für die Analysen dieser Studie wurden lediglich die Daten des NEPS-Mathematiktests verwendet. Die Aufgaben zur mathematischen Kompetenz waren in beiden Testheften an gleicher Stelle platziert (A16). Der NEPS-Mathematiktest beinhaltete 22 Aufgaben, davon waren 19 im einfachen MC, 2 im CMC und eine im SCR. Alle Schülerinnen und Schüler bearbeiteten die 22 Mathematikaufgaben in der gleichen Reihenfolge und hatten dafür eine Bearbeitungszeit von 28 Minuten.

Die Schülerinnen und Schüler erhielten zusammen mit dem Kompetenztest einen Fragebogen, für den sie ca. 40 Minuten Bearbeitungszeit hatten (Methodenbericht A46). Auch die Lehrkräfte und Schulleitungen erhielten Fragebögen. Diese wurden im Rahmen dieser Dissertation jedoch nicht verwendet. Die Eltern der Schülerinnen und Schüler wurden in einer Teilstudie telefonisch befragt (nähere Informationen siehe Methodenbericht computergestütztes Telefoninterview (CATI)). Für die Analysen dieser Studie wurden

Tabelle 5: Informationen zum Kompetenztest der Haupterhebung 2010/11, Klasse 9 in Regelschulen aus Pohl und Carstensen (2012)

Informationen zu den einzelnen Tests				
Konstrukt	**Anzahl der Items**	**vorgegebene Bearbeitungszeit**	**Erhebungsmodus**	**Nächste Messung (bis 2013)**
ICT-Literacy	40	29 min	paper-pencil	nach 3 Jahren
Naturwissenschaftliche Kompetenz	28	29 min	paper-pencil	
Lesegeschwindigkeit	51	2 min	paper-pencil	nach 3 Jahren
Mathematische Kompetenz	22	28 min	paper-pencil	nach 2 Jahren
Hörverstehen auf Wortebene: rezeptiver Wortschatz	89	20 min	paper-pencil	
Domänenspezifische prozedurale Metakognition				Entsprechend den jeweiligen Domänen
Zur Domäne ICT-Literacy	1	1 min	paper-pencil	s.o.
Zur Domäne Naturwissenschaftliche Kompetenz	1	1 min	paper-pencil	s.o.
Zur Domäne Mathematische Kompetenz	1	1 min	paper-pencil	s.o.
Zur Domäne Rezeptiver Wortschatz	1	1 min	paper-pencil	s.o.

Informationen aus der Schüler- und der Elternbefragung bezüglich der Anzahl der Bücher zu Hause, des Geschlechts, des Migrationshintergrundes und der Schulform der Zielperson verwendet. Auf Grund des längsschnittlichen Designs des Nationalen Bildungspanels ist es vorgesehen die Schülerinnen und Schüler jährlich innerhalb der Schule zu befragen, bis diese die ursprünglich ausgewählte Schule oder das allgemeinbildende Schulsystem verlassen haben. Anschließend sollen die Testpersonen außerhalb der Institution weiter befragt und getestet werden. Die längsschnittlichen Daten werden in dieser Studie nicht betrachtet.

3.2 Studie 2: Validierungsstudie 2012

3.2.1 Stichprobenziehung in der Validierungsstudie 2012

Die Validierungsstudie wurde im Frühjahr 2012 durchgeführt. Diese Studie hat neben der Validierung der Testwertinterpretationen des NEPS Mathematik- und NaWi -Tests auch die Evaluation langfristiger Effekte des Bund-Länder-Kommission für Bildungsplanung und Forschungsförderung (BLK)-Modellversuchsprogramms Steigerung der Effizienz des mathematisch-naturwissenschaftlichen Unterrichts (SINUS) beziehungsweise des Nachfolgeprojektes SINUS-Transfer auf die mathematische und naturwissenschaftliche Kompetenz von Schülerinnen und Schülern zum Ziel. Aus diesem Grund wurde die Stichprobe für die Validierungsstudie aus Schulen gezogen, die am BLK-Programm SINUS beziehungsweise SINUS-Transfer teilgenommen und das Programm dauerhaft etabliert haben. Schulen, die bereits an den Hauptstudien des Ländervergleichs und PISA 2012 teilgenommen haben, wurden für die Stichprobenziehung nicht berücksichtigt. Für das Bundesland Thüringen ergab sich auf Wunsch des Bundesministeriums für Bildung eine Besonderheit im Erhebungsdesign. In diesem Bundesland wurden nur Schulen ausgewählt, die an der originalen Testung der Länderübergreifenden Bildungsstandards teilgenommen hatten. Insgesamt wurden 80 Schulen mit 1965 Schülerinnen und Schülern für die Teilnahme an der Studie ausgewählt. Für 1718 Schülerinnen und Schüler lag eine Einverständniserklärung für die Teilnahme an der Studie vor. Für 17 Schülerinnen und Schüler ist das Geschlecht unbekannt.

Tabelle 6: Schüleranzahl, Geschlechterverteilung und Vorliegen der Einverständniserklärung in Abhängigkeit der Schulform für die Validierungsstudie 2012

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

3.2.2 Testdesign

Die Testung der Schülerinnen und Schüler wurde an zwei Tagen durchgeführt (siehe Abbildung 22). Am ersten Testtag bearbeiteten die Schülerinnen und Schüler nach einer Einweisung von ca. 10 Minuten Mathematik- und NaWi -Testhefte aus PISA 2012. Die reine Bearbeitungszeit betrug 120 Minuten, wobei nach einer Zeitstunde eine Pause von 60 Minuten eingelegt wurde. Ein Testheft enthielt aus methodischen Gründen ebenfalls Aufgaben zum Kompetenzbereich Lesen. Diese Aufgaben wurden jedoch nicht in den Berechnungen berücksichtigt. Anschließend wurden die Schülerinnen und Schüler nach einer weiteren Pause von 15 Minuten mittels eines Fragebogens zu Hintergrundmerkmalen befragt. Hierfür standen 45 Minuten zur Verfügung. Der zweite Testtag begann ebenfalls mit einer Einführung von 15 Minuten. Anschließend beantworteten die Testpersonen entweder erst Aufgaben aus dem LV 2012 Mathematik- und NaWi -Test und danach den Mathematik- und NaWi -Test des NEPS oder sie bearbeiteten erst die NEPS-Tests und nachfolgend die Aufgaben aus den LV -Tests. Für die Bearbeitung der LV -Aufgaben und der NEPS-Tests standen den Schülerinnen und Schülern jeweils 60 Minuten zur Verfügung, zwischen den Tests gab es eine Pause von 15 Minuten. Im Anschluss an die Testbearbeitung und nach einer weiteren Pause von 15 Minuten beantworteten die Schülerinnen und Schüler 20 Minuten lang Aufgaben aus dem Subtest des Berliner Tests zur Erfassung fluider und kristalliner Intelligenz für die 8. Bis 10. Jahrgangsstufe (BEFKI)

zur Erfassung figuraler Aspekte. Darauffolgend füllten die Testpersonen einen Fragebogen mit zusätzlichen Informationen zu Hintergrundmerkmalen aus. Für die Schülerinnen und Schüler aus Thüringen ergab sich aufgrund der Besonderheit im Stichprobendesign ein abweichendes Testdesign. Diese Testpersonen hatten bereits an der originalen Erhebung der Länderübergreifenden Bildungsstandards 2012 teilgenommen und den dazugehörigen Fragebogen bearbeitet. Im Rahmen der Validierungsstudie bearbeiteten die Jugendlichen daher an einem Testtag Aufgaben aus dem PISA 2012 Mathematik- und NaWi -Test und die NEPS Mathematik- und NaWi -Tests. Nach einer Einführung von ca. 10 Minuten wurden 60 Minuten lang Aufgaben bearbeitet, woraufhin eine Pause von 15 Minuten folgte. Danach wurde die zweite Hälfte der Aufgaben innerhalb von weiteren 60 Minuten beantwortet.

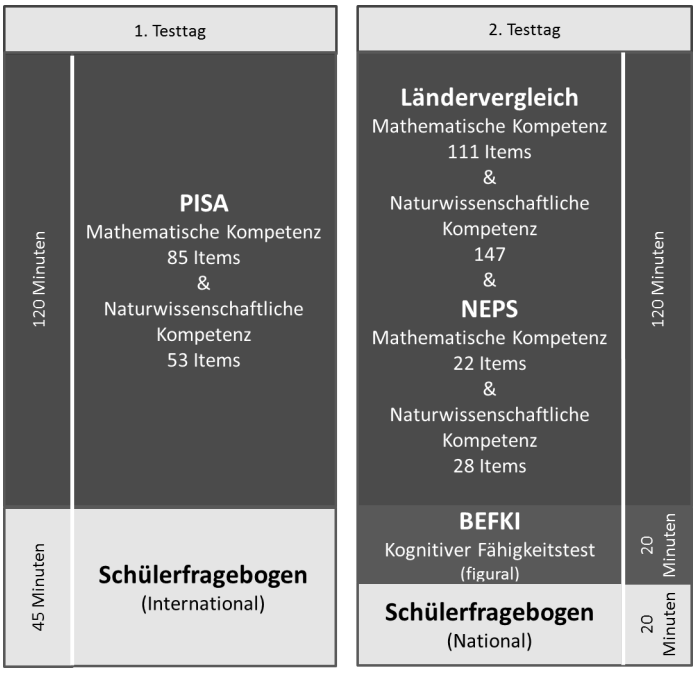


Abbildung 22: Testdesign der Validierungsstudie

Testheftdesign Testtag 1

Für die Testhefte des ersten Testtages wurden sieben Mathematik-Aufgabenblöcke, drei NaWi-Aufgabenblöcke und ein Lese-Aufgabenblock in Form eines Multi-Matrix Designs über fünf Testhefte rotiert (siehe Abbildung 1). Ein Teil der Blöcke kam in mehreren Testheften vor, anderer Blöcke befanden sich in nur einem Testheft. Insgesamt wur-

den 85 Mathematikaufgaben, 53 NaWi-Aufgaben und 14 Lese-Aufgaben eingesetzt. Die Aufgaben zum Kompetenzbereich Lesen waren eine methodische Konsequenz des Testheftdesigns und wurden nicht in den Berechnungen berücksichtigt. Die Schülerinnen und Schülern bearbeiteten PISA-Mathematikaufgaben in einer Zeit von 30 bis 90 Minuten und PISA NaWi-Aufgaben in einer Zeit von 30 bis 60 Minuten.

Testheftdesign Testtag 2

Für den zweiten Testtag wurden 12 Mathematik- und 12 NaWi -Aufgabenblöcke aus dem Ländervergleich 2012 sowie die NEPS Mathematik- und LV -Tests in Form eines Multi-Matrix Designs über 16 Testhefte rotiert (siehe Abbildung 2). Für eine vollständige Verankerung über die Testheftversionen wurden im Rahmen des Multi-Matrix Designs außerdem vier Testhefte erstellt, die ausschließlich Aufgabenblöcke aus dem Ländervergleich beinhalteten. Auf diese Weise entstanden 20 Testhefte, die den Schülerinnen und Schülern zufällig zugeteilt wurden. Auch hier kamen einige Aufgabenblöcke in mehreren Testheften vor und andere Aufgaben nur in einem Testheft. Insgesamt wurden 111 Mathematikaufgaben aus dem Ländervergleich verwendet. Je nach Testheft bearbeiteten die Schülerinnen und Schüler Mathematikaufgaben in einer Zeit von 60 bis 120 Minuten. In der Domäne Naturwissenschaften wurden 147 Aufgaben aus dem Ländervergleich verwendet. Je nach Testheft bearbeiteten die Schülerinnen und Schüler NaWi -Aufgaben in einer Zeit von 60 bis 120 Minuten. Neben den Aufgaben aus dem Ländervergleich wurden alle Aufgaben aus dem NEPS Mathematik- und Naturwissenschaftstest der HE 2010 verwendet. Der NaWi -Test bestand aus 28 Aufgaben. Zur Bearbeitung standen den Schülerinnen und Schülern jeweils 30 Minuten für die Mathematik- und NaWi -Aufgaben zur Verfügung.

Testheftdesign Bundesland Thüringen

Für die Testung im Bundesland Thüringen wurden zwei Mathematik-, zwei NaWi-Aufgabenblöcke aus PISA 2012 sowie die Mathematik und NaWi-Tests aus NEPS über fünf Testhefte rotiert (siehe Abbildung 3). Insgesamt wurden 25 Mathematik- und 35 NaWi-Aufgaben aus PISA verwendet. Je nach Testheft bearbeiteten die Schülerinnen und Schüler PISA -Mathematikaufgaben in einer Bearbeitunszeit von 30 bis 60 Minuten

und NaWi-Aufgaben in einer Bearbeitungszeit von 30 bis 60 Minuten. Der Mathematik- und NaWi-Test aus NEPS wurden von den Schülerinnen und Schülern vollständig bearbeitet. Die Daten der originalen Testung der Länderübergreifenden Bildungsstandards, an der diese Schülerinnen und Schüler zuvor teilgenommen hatten, wurden der Validierungsstudie zur Verfügung gestellt. Dabei wurden nur die Aufgaben ausgewertet, die auch in der Validierungsstudie eingesetzt wurden.

3.3 Analysen zur Auswertung des IUA

Bei dem Einsatz des NEPS-Mathematiktests und auch der Tests aus PISA und dem LV wurde von einer latenten, nicht direkt sichtbaren Personeneigenschaft ausgegangen, die mit einem beobachteten Verhalten zusammenhängt. Dieser Zusammenhang von Ausprägungen auf einer latenten Variable und manifestem, beobachteten Verhalten lässt sich mit Hilfe von Item Response Modellen modellieren. Aus diesem Grund und weil die eingesetzten Tests auch in den Haupterhebungen mit IRT modelliert wurden, wurden für die Beantwortung einiger Forschungsfragen IRT-Modelle eingesetzt. Daher wurden vor einer Beschreibung der Analysemethoden für jede Schlussfolgerung in der Argumentationskette des IUA einige wichtige Grundlagen der hier verwendeten IRT-Modellierungen beschrieben.

3.3.1 Skalierung der Kompetenzdaten

Bei der Skalierung der Kompetenzdaten wurde das sogenannte Partial-Credit–Modell von Masters (1982) verwendet. Dieses ist eine Erweiterung des Rasch-Modells für mehrstufige Antwortkategorien. Dichotome und polytome Aufgaben können gleichzeitig berücksichtigt werden. Das Partial-Credit-Model (PCM) zerlegt ein ordinales Item in mehrere Schwellen-Parameter. Die Abstände zwischen den Schwellenparametern beschreiben die Schwierigkeit des Gelangens von einer zur nächsten Antwortkategorie.

Bei der Verwendung eines eindimensionalen Modelles wurde davon ausgegangen, dass die Items eine einzige latente Personeneigenschaft repräsentieren und sich einzig in den Schwierigkeiten unterscheiden. Für die Analyse mehrerer latenter Personeneigenschaften

wurden mehrdimensionale Skalierungen durch eine Erweiterung des PCM durchgeführt. Bei dem sogenannten Mixed Coefficient Multinomial Logit Model (MCMLM) werden die kategorialen Antworten in den meisten Anwendungen als unabhängige Variablen modelliert. Die Antwortmuster werden mit Hilfe einer logistischen Regression vorhergesagt, wobei die Personenfähigkeiten und Itemschwierigkeiten als unabhängige Variablen fungieren (Wu, Adams, Wilson & Haldane, 2007, S. 57-75).

Die Item Parameter wurden sowohl bei einer eindimensionalen als auch bei einer mehrdimensionalen Skalierung mit Hilfe von Weighted Maximum Likelihood Estimate (MML) gewonnen. Bei der MML wird eine bestimmte Verteilung der Personenparameter zu Grunde gelegt und es werden zuerst nur die Itemparameter geschätzt. In einem zweiten Schritt werden dann die so gewonnen Itemparameterschätzungen eingesetzt, um wiederum die Personenparameter zu schätzen. Die Personenfähigkeitsschätzer wurden in Form von Weighted Maximum Likelihood Estimates (WLEs) (Warm, 1989) und Plausible Values (PVs) (Mislevy, 1991) ausgegeben.
Die Modellparameter für Berechnungen mit der Validierungsstichprobe wurden in dieser Arbeit frei geschätzt und nicht auf die Parameter der NEPS-, PISA- beziehungsweise LV-Haupterhebungen fixiert. Der Grund hierfür ist, dass die freie Schätzung eine bessere Modellanpassung an die Daten ermöglichte. Als Konsequenz befanden sich die geschätzten Personenparameter nicht auf den originalen PISA-, LV- bzw. NEPS-Metriken. Jedoch wurde eine Verzerrung der gefundenen Korrelationen durch eine Fixierung ausgeschlossen. Auch wurden für die Berechnungen in Übereinstimmung mit dem Vorgehen der NEPS-Haupterhebung die WLE-Werte berechnet und nicht wie bei den Haupterhebungen aus PISA und dem LV die PV-Werte. Es musste bei der Interpretation der Ergebnisse zwar beachtet werden, dass die Testergebnisse nicht mit den Testwerten der Haupterhebungen vergleichbar sind. Da jedoch lediglich die Zusammenhänge zwischen den Testkonstrukten analysiert werden sollten, ist dieser Nachteil nicht als gravierend einzuschätzen.

Für die Durchführung der Skalierung wurde die Software ConQuest (Wu et al., 2007) verwendet. Mit diesem Programm ist es möglich, zu berücksichtigen, dass Aufgaben aufgrund des verwendeten Multi-Matrix-Designs zufällig fehlen.

3.3.2 Scoring und fehlende Werte

Sowohl in der NEPS Haupterhebung im Jahr 2010 als auch in der Validierungsstudie wurde jede korrekt gelöste polytome Aufgabe im NEPS-NaWi- und Mathematiktest für die neunte Klassenstufe mit 0.5 Punkten gewertet und jede korrekte MC- und SCR-Aufgabe mit einem Punkt (Duchhardt & Gerdes, 2013). In beiden Studien wurden vier Arten von fehlenden Werten gewertet: Items, die aufgrund des Testdesigns nicht vorlagen (a), ungültige Antworten (b), ausgelassene Antworten (c) und aufgrund von Zeitbeschränkungen nicht erreichte Antworten (d). Für die Skalierung der Kompetenzdaten wurden die unterschiedlichen Arten der fehlenden Werte ignoriert (Pohl & Carstensen, 2012). Alle in der Validierungsstudie eingesetzten LV-Mathematik- und LV-NaWi-Aufgaben wurden dichotom gewertet (0 für falsch und 1 für richtig). Für beide Bereiche wurden vier Arten von fehlenden Werten administriert: Items, die aufgrund des Testdesigns nicht vorliegen (a), ungültige Antworten (b), ausgelassene Antworten (c) und aufgrund von Zeitbeschränkungen nicht erreichte Antworten (d). Bei der Skalierung der WLE-Modelle wurden fehlende Werte, die aufgrund des Testdesigns entstanden, ignoriert. Die übrigen fehlenden Werte (b bis d) wurden als falsch (mit 0) kodiert. Von in der Validierungsstudie eingesetzten Mathematikaufgaben aus PISA wurden 77 Aufgaben dichotom gewertet (0 für eine falsche und 1 für eine richtige Antwort). Die übrigen acht Aufgaben hatten eine polytome Wertung mit drei Abstufungen (0,1 und 2 Punkte). Von den NaWi-Aufgaben wurden 50 dichotom (0 für eine falsche und 1 für eine richtige Antwort) und drei polytom gewertet mit drei Abstufungen (0,1 und 2 Punkte). Es wurden drei Arten von fehlenden Werten erfasst: Items die aufgrund des Testdesigns nicht vorlagen (a), ungültige Antworten (b), ausgelassene Antworten (c). Für die Auswertung der Kompetenzdaten mit WLE-Modellen wurden die fehlenden Werte, die aufgrund des Testdesigns entstehen, ignoriert und die übrigen fehlenden Werte als falsch kodiert.

4 Ein Validitätsargument für NEPS-K9-Mathematiktest

In diesem Kapitel wird die Evaluation des Interpretation / Use Argument (IUA), welches das beobachtete Verhalten mit den Interpretationen der Testwerte verbindet, für den NEPS-K9-Mathematiktest vorgestellt. Die Hypothesen der ersten fünf Schlussfolgerungen *Domänenbeschreibung*, *Bewertung*, *Skalierung*, *Generalisierung* und *Konstruktbezug* wurden basierend auf der Stichprobe der Haupterhebung ausgewertet. Für die Evaluation der Schlussfolgerung *Extrapolation* wurden zwei Hypothesen basierend auf beiden Stichproben, eine basierend auf der Stichprobe der Validierungsstudie und eine basierend auf der Stichprobe der Haupterhebung im Jahr 2010, ausgewertet (vgl. Kapitel 4.6).

4.1 Domänenbeschreibung

Die Argumentationsstruktur für die Schlussfolgerung *Domänenbeschreibung* wurde im Kapitel 2.1 entwickelt und dargelegt. Eine visuelle Darstellung der Schlussfolgerung, des Argumentes, der Annahme und der drei Hypothesen wurde in Kapitel 2.1 in der Abbildung 16 geboten. Nachfolgend wird zunächst die Methode für die Evaluation der drei Hypothesen dargestellt. Anschließend wird im Ergebnisteil beschrieben, inwiefern relevante Teilkompetenzen aus den PISA- und LV-Rahmenkonzepten in den NEPS-Aufgaben identifiziert werden können. Außerdem wird ein Vergleich der Operationalisierung der inhaltlichen und kognitiven Teilbereiche aus NEPS mit denen aus PISA und dem LV vorgestellt. In einem weiteren Schritt wird ein Vergleich der Gewichtung der Teilkompetenzen aus den PISA- und LV-Rahmenkonzeptionen im NEPS-Mathematiktest mit der Gewichtung der Teilkompetenzen in den Studien PISA 2012 und LV 2012 dargelegt. Das Unterkapitel schließt mit einer Zusammenfassung und Diskussion der Ergebnisse ab.

4.1.1 Methode

Um die Repräsentativität des NEPS-K9-Mathematiktests für die Zieldomäne zu untersuchen, ist es von Belang, die Abdeckung der Teilkompetenzen der Zieldomäne durch die verwendeten Aufgaben zu analysieren (vgl. Kapitel 2.1).

NEPS-Hypothesen zu Teilkompetenzen

Um die Hypothesen zu der Annahme der *Domänenbeschreibung*: „Die Aufgaben des NEPS-Mathematiktest decken die für die Zieldomäne relevanten Teilkompetenzen angemessen ab“ zu prüfen, wurde ein Expertenreview der NEPS-Mathematikaufgaben durchgeführt.

Die erste Hypothese „In den Aufgaben des NEPS-K9-Mathematiktests lassen sich die relevanten Domänen mathematischer Kompetenz aus dem Rahmenkonzept von PISA und dem des LV identifizieren.“ wurde untersucht, indem drei in der Konstruktion von Aufgaben erfahrene Experten die 22 Mathematikaufgaben anhand eines vorher erstellten Fragebogens einschätzten. Um die relevanten mathematischen Teildomänen zu identifizieren, wurden die 22 Mathematikaufgaben in die Teildomänen der LV-Rahmenkonzeption und der PISA-Rahmenkonzeption eingeordnet. Die Experten wählten für jede Aufgabe des NEPS-Mathematiktests jeweils aus, zu welchem der fünf Inhaltsbereiche aus dem LV und zu welchem der vier Inhaltsbereiche aus PISA diese Aufgabe passt. Anschließend beurteilten die Experten für jede Aufgabe, welche der sechs kognitiven Prozesse aus dem LV und welche der sieben kognitiven Prozesse aus PISA für die Lösung der Aufgabe benötigt werden. Für jeden dieser Prozesse schätzten die Experten ein, auf welchem Anforderungsbereich nach PISA beziehungsweise LV dieser Prozess gefordert wird (für Beschreibungen der Inhaltsbereiche, kognitiven Prozesse und Anforderungsbereiche aus PISA und LV siehe Kapitel 1.2). Für die Einordnungen der NEPS-Aufgaben in die Inhaltsbereiche, kognitiven Prozesse und Anforderungsbereiche wurde eine Expertenübereinstimmung berechnet. Dafür wurde der Median der Prozentualen Übereinstimmung (PÜ) zwischen jeweils dem ersten und zweiten Experten, dem ersten und dritten Experten und dem zweiten und dritten Experten berechnet. Außerdem wurde das zufallskorrigierte Übereinstimmungsmaß Cohens Kappa κ verwendet. Auch hier wurde der Median der Cohens Kappa- Werte zwischen jeweils zwei Experten in den drei

möglichen Kombinationen berechnet. Anschließend wurden die NEPS-Aufgaben basierend auf den Experteneinschätzungen in die jeweiligen Kategorien der Inhaltsbereiche, Prozesse und Anforderungsbereiche des LV und in die Inhaltsbereiche, Prozesse, Anforderungsbereiche und Kontexte aus PISA eingeordnet. Dabei wurde jeweils die Kategorie beziehungsweise Abstufung für eine Aufgabe gewählt, die von mindestens zwei Experten für diese Aufgabe eingeschätzt wurde. Auf diese Weise war es möglich, alle Aufgaben in die Teildimensionen aus LV und PISA einzuordnen.

Um die zweite Hypothese „Die kognitiven und inhaltlichen Teilbereiche werden in NEPS und dem LV beziehungsweise NEPS und PISA auf ähnliche Weise operationalisiert.“ zu untersuchen, wurde die aus dem Expertenreview resultierende Verteilung in den inhaltlichen Teilkompetenzen von LV und PISA der Verteilung in den originalen NEPS-Teilkompetenzen gegenübergestellt. Unterschiede zwischen diesen Verteilungen gaben Aufschluss über Unterschiede in der Operationalisierung der Teilkompetenzen zwischen den Studien. Es wurde sichtbar, inwiefern sich die jeweiligen inhaltlichen Teilkompetenzen aus dem NEPS von denen aus PISA- bzw. LV-Teilkompetenzen unterscheiden. Die Operationalisierung der kognitiven Teilkompetenzen konnte nicht untersucht werden, da die Einordnung der NEPS-Aufgaben in die kognitiven Prozesse des NEPS nicht vorlag.

Die Hypothese „Die Gewichtung der Komponenten aus den PISA- und LV-Rahmenkonzepten für den NEPS-K9-Mathematiktest unterscheidet sich nicht signifikant von der Gewichtung in den Mathematiktests aus den Haupterhebungen PISA 2012 und LV 2012.“ wurde getestet, indem die aus dem Expertenreview gewonnene Verteilung der NEPS-Mathematikaufgaben in den Teilkompetenzen von PISA und dem Ländervergleich mit der Verteilung der PISA- und LV-Aufgaben aus den originalen Studien im Jahr 2012 verglichen wurde. Die Unterschiede zwischen den Verteilungen wurden mit einem Chi-Quadrat-Test auf Signifikanz geprüft.

Aufgrund fehlender Informationen über die Zugehörigkeit von NEPS-, PISA- und LV-Aufgaben zu Prozessen, Kontexten und Anforderungsbereichen konnten die beschriebenen Berechnungen für die Testung der zweiten und dritten Hypothese nicht für alle Teildimensionen durchgeführt werden. Für die NEPS-Aufgaben waren beispielsweise lediglich die Einordnungen in die Inhaltsbereiche bekannt. Aus diesem Grund konnte lediglich die Verteilung der NEPS-Aufgaben in den NEPS-Inhaltsbereichen mit der Verteilung der NEPS-Aufgaben in den PISA- und LV-Inhaltsbereichen verglichen werden. Des Weiteren war die Verteilung der PISA-Mathematikaufgaben aus 2012 in den Prozes-

sen der PISA-Rahmenkonzeption nicht bekannt und konnte diese Verteilung nicht mit der Verteilung der NEPS-Mathematikaufgaben in den PISA-Prozessen verglichen werden. Für die LV-Mathematikaufgaben aus 2013 war die Verteilung in den Prozessen und Anforderungsbereichen nicht vorhanden, weshalb hier nur ein Vergleich der Verteilungen mit den Inhaltsbereichen möglich war.

4.1.2 Ergebnisse

Die Ergebnisse zur Relevanz der mit den NEPS-Mathematikaufgaben abgedeckten Teilkompetenzen werden im Folgenden für die Hypothesen H1, H2 und H3 zur *Domänenbeschreibung* dargestellt.

Ergebnisse zu H1: Domänen mathematischer Kompetenz in NEPS

Im Folgenden werden die Ergebnisse der Experteneinordnung der NEPS-Mathematikaufgaben in die Teilbereiche der PISA-Rahmenkonzeption und anschließend für die Teilbereiche der LV-Rahmenkonzeption für die Überprüfung der Hypothese 1 „In den Aufgaben des NEPS-K9-Mathematiktests lassen sich die relevanten Domänen mathematischer Kompetenz aus dem Rahmenkonzept von PISA und dem des LV identifizieren.“ dargelegt [1].

Ergebnisse zur Einordnung der NEPS-Mathematikaufgaben in die PISA-Teilkompetenzen. Die Experten konnten jeder NEPS-K9-Mathematikaufgabe einen Inhaltsbereich aus PISA zuordnen. Dabei ordneten mindestens zwei Experten jede Aufgabe in den gleichen Inhaltsbereich ein. Für die Übereinstimmung wurden eine $P\ddot{U}$ von 90 % und ein $\kappa = 0.74$ gefunden. Die Verteilung der NEPS-Mathematikaufgaben in den PISA-Inhaltsbereichen wird in Abbildung 23 dargestellt.

In allen NEPS-K9-Mathematikaufgaben konnten die Experten Prozesse der PISA-Rahmenkonzeption identifizieren. Bei der Zuordnung jeder Aufgabe zu einem Prozess oder

[1] Teile dieser Auswertungen wurden bereits in van den Ham et al. (2014) veröffentlicht

mehreren Prozessen stimmten mindestens zwei Experten in ihrer Wahl überein. Insgesamt wurde eine $P\dot{U}$ von 75% und ein $\kappa = 0.49$ festgestellt. Es gab somit keine NEPS-Aufgaben, die andere als die in PISA definierten Prozesse beinhalteten. Die Verteilung der NEPS-Mathematikaufgaben in den Prozessen der PISA-Rahmenkonzeption wird in Tabelle 8 dargestellt. Für die erfolgreiche Lösung einer Aufgabe können in der PISA-Rahmenkonzeption mehrere Prozesse pro Aufgabe benötigt werden. In den NEPS-Mathematikaufgaben ließen sich fünf der sechs PISA-Prozesse identifizieren. Lediglich der Prozess „Mathematische Hilfsmittel verwenden“ wurde von den Experten keiner der 22 NEPS-Mathematikaufgaben zugeordnet. Dies ist nicht weiter verwunderlich, da schon in der Rahmenkonzeption angegeben wird, dass mit dem NEPS-Test keine separate Kompetenz zum Umgang mit mathematischen Hilfsmitteln gemessen werden soll (vgl. Kapitel 1.2).

Auch die Anforderungsbereiche der PISA-Rahmenkonzeption konnten in allen NEPS-K9-Mathematikaufgaben von den Experten identifiziert werden. Bei der Zuordnung jeder Aufgabe zu einem Anforderungsbereich stimmten mindestens zwei Experten in ihrer Wahl überein. Es konnten somit keine NEPS-Aufgaben gefunden werden, die einem anderen Anforderungsbereich als den in PISA definierten zugeordnet werden müssen. Insgesamt wurde eine $P\dot{U}$ von 60 % und ein $\kappa = 0.34$ gefunden. Die Einordnung der NEPS-Aufgaben in die Anforderungsbereiche wird in Tabelle 9 aufgezeigt.

Die Kontexte aus der PISA-Rahmenkonzeption wurden in allen NEPS-K9-Mathematikaufgaben durch die Experten identifiziert. Mindestens zwei Experten stimmten bei allen Aufgaben in ihrer Wahl des Kontextes überein. In den NEPS-K9-Aufgaben ließen sich also international anerkannte Kontexte wiederfinden. Dabei wurde eine $P\dot{U}$ von 79% und ein $\kappa = 0.42$ realisiert. Die Einordnung in die Kontexte wird in Tabelle 10 wiedergegeben.

Einordnung der NEPS-Mathematikaufgaben in die LV-Teilkompetenzen. Die Experten konnten jeder NEPS-K9-Mathematikaufgabe einen Inhaltsbereich aus dem LV zuordnen. Dabei ordneten mindestens zwei Experten die jeweiligen Aufgaben in den gleichen Inhaltsbereich ein. Für die Übereinstimmung wurden eine $P\dot{U}$ von 90 % und ein κ = 0.69 gefunden. Abbildung 24 zeigt, inwiefern sich die NEPS-Mathematikaufgaben in die inhaltlichen Teilkompetenzen der LV-Rahmenkonzeption einordnen ließen.

In allen NEPS-K9-Mathematikaufgaben konnten die Experten Prozesse der LV-Rahmenkonzeption identifizieren. Für die erfolgreiche Lösung einer Aufgabe können mehrere Prozesse benötigt werden. Bei der Zuordnung jeder Aufgabe zu einem Prozess oder mehreren Prozessen stimmten mindestens zwei Experten in ihrer Wahl überein. Insgesamt wurde eine PÜ von 76 % und ein $\kappa = 0.53$ gefunden. Die Verteilung der NEPS-Mathematikaufgaben in den Prozessen der LV-Rahmenkonzeption wird in Tabelle 12 dargestellt. Die NEPS-Aufgaben decken keine Prozesse ab, die nicht auch im LV definiert sind.

Die Verteilung der NEPS-Aufgaben auf die LV-Anforderungsbereiche wird in Tabelle 13 dargestellt. Bei der Zuordnung jeder Aufgabe zu einem Prozess oder mehreren Prozessen stimmten mindestens zwei Experten in ihrer Wahl überein. Insgesamt wurde eine PÜ von 64% und ein $\kappa = 0.33$ gefunden. Die Experten konnten zwei der drei Anforderungsbereiche aus dem LV in den NEPS-Aufgaben identifizieren. Der Anforderungsbereich „Verallgemeinern und Reflektieren“ wurde von den Experten nicht gewählt, was unter anderem durch die verwendeten Aufgabenformate im NEPS bedingt sein kann (s. Seite 120).

Ergebnisse zu H2: Operationalisierung der Teilbereiche

Zur Prüfung der Hypothese 2 „Die kognitiven und inhaltlichen Teilbereiche werden in NEPS und dem LV beziehungsweise NEPS und PISA auf ähnliche Weise operationalisiert.“ wurden die Experteneinordnungen der NEPS-Mathematikaufgaben in die inhaltlichen Teilbereiche der LV- und PISA-Rahmenkonzeptionen den Einordnungen der NEPS-Mathematikaufgaben in die inhaltlichen Teilbereiche der NEPS-Rahmenkonzeption gegenübergestellt [2].

Vergleich der Operationalisierung der inhaltlichen Teilkompetenzen in NEPS und PISA. Die Einordnung der NEPS-Mathematikaufgaben in die inhaltlichen Teilkompetenzen der PISA-Rahmenkonzeption wird in Abbildung 23 dargestellt. Diese Verteilung wird außerdem der Verteilung der Aufgaben auf die Inhaltsbereich der NEPS-Rahmenkonzeption gegenübergestellt. Die fett gedruckten Pfeile in der Abbildung zeigen, in welchen Inhaltsbereich der PISA-Rahmenkonzeption die meisten Aufgaben aus

[2] Teile dieser Auswertungen wurden bereits in van den Ham et al. (2014) veröffentlicht

dem jeweiligen NEPS-Inhaltsbereich von den Experten eingeordnet wurden. Die dünn gedruckten Pfeile zeigen, in welche PISA-Inhaltsbereiche einzelne NEPS-Aufgaben eingeordnet wurden. Es wurde kein signifikanter Unterschied zwischen der Verteilung der NEPS-Aufgaben in den PISA-Inhaltsbereichen und der Verteilung der NEPS-Aufgaben in den NEPS-Inhaltsbereichen gefunden ($\chi^2 = 0.62$ $df = 3$, ns). In den mathematischen Rahmenkonzeptionen von NEPS und PISA werden die Inhaltsbereiche auf ähnliche Weise operationalisiert. Kleinere Abweichungen wurden zwischen den Inhaltsbereichen „Quantität“ in NEPS und „Quantität“ in PISA sowie zwischen „Raum und Form“ in NEPS und „Raum und Form“ in PISA gezeigt. Im NEPS scheint der Inhaltsbereich „Raum und Form“ mehr Inhalte zu umfassen, da diesem auch Aufgaben aus den PISA-Inhaltsbereichen „Unsicherheit und Daten“, „Veränderung und Beziehungen“ und „Raum und Form“ zugeordnet wurden. Umgekehrt umfasst der Inhaltsbereich „Raum und Form“ im PISA auch Aufgaben, die im NEPS den Inhaltsbereichen „Veränderung und Beziehungen“ und „Quantität“ zugeordnet wurden. Diese jedoch nicht signifikanten Abweichungen weisen auf eine weitgehend gleiche Operationalisierung der Teildimensionen hin.

Für den Teilbereich Prozesse lagen keine Einordnungen in die NEPS-Rahmenkonzeption vor. Aus diesem Grund konnte die Operationalisierung der Teildimensionen nicht verglichen werden.

Vergleich der Operationalisierung der inhaltlichen Teilkompetenzen in NEPS und im LV. Die Verteilung der NEPS-Aufgaben in den LV-Inhaltsbereichen wurde der Verteilung der Aufgaben auf die Inhaltsbereiche der NEPS-Rahmenkonzeption gegenübergestellt (siehe Abbildung 24). Die fett gedruckten Pfeile beschreiben, in welchen Inhaltsbereich der LV-Rahmenkonzeption die meisten NEPS-Aufgaben des jeweiligen NEPS-Inhaltsbereiches eingeordnet wurden. Die dünn gedruckten Pfeile illustrieren diese Einordnung ebenfalls, jedoch für einzelne Aufgaben. Die meisten Aufgaben wurden von den Experten dem verwandten Inhaltsbereich der anderen Rahmenkonzeption zugeordnet (siehe Abbildung 24). So wurden die meisten Aufgaben des NEPS-Inhaltsbereiches „Quantität“ dem verwandten Inhaltsbereich „Zahl“ der LV-Rahmenkonzeption zugeordnet. Auffällig ist jedoch, dass die meisten Aufgaben des NEPS-Inhaltsbereiches „Raum und Form“ dem LV-Inhaltsbereich „Messen“ zugeordnet wurden. Es handelte sich hierbei vor allem um Aufgaben, bei denen Flächen oder Winkel berechnet werden müssen. Das Berechnen von Inhalten oder Umfängen geometrischer Strukturen beinhaltet oft auch

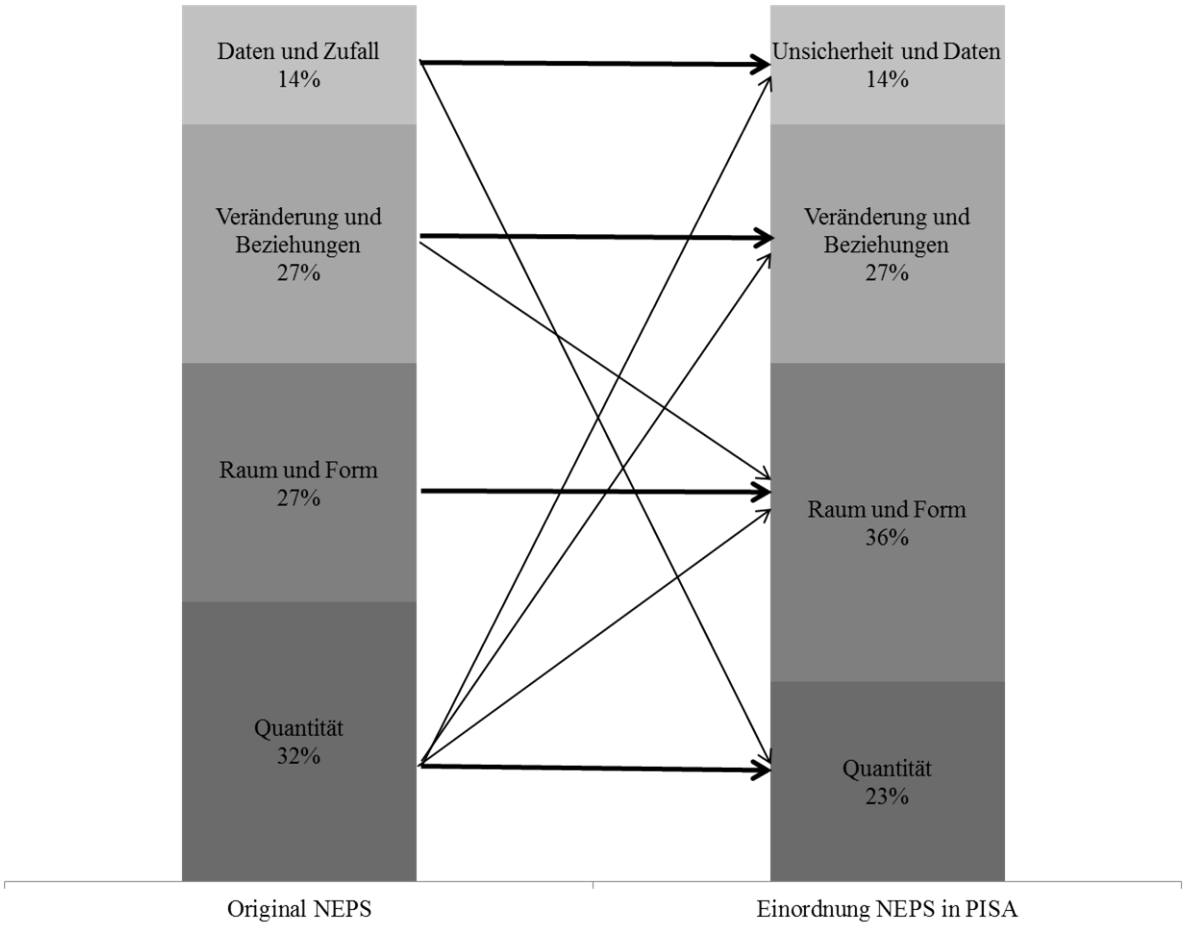


Abbildung 23: Prozentuale Verteilung der Mathematikaufgaben aus NEPS in die Inhaltsbereiche aus NEPS und PISA

das Einbeziehen von speziellen Eigenschaften, wodurch zum Lösen der Aufgabe Zugänge aus sowohl „Raum und Form“ als auch „Messen“ benötigt werden. Das Bestimmen von Maßen wird in der LV-Rahmenkonzeption unter dem Bereich „Messen“ definiert, in der NEPS-Rahmenkonzeption jedoch nicht explizit benannt. In der NEPS-Rahmenkonzeption wird für den Inhaltsbereich „Raum und Form“ jedoch das Analysieren von und das gedankliche Operieren mit geometrischen Strukturen beschrieben. In der Definition des Ländervergleichs sind die meisten „Raum und Form“-Aufgaben aus dem NEPS also Aufgaben, die auch „Messen“ nach Definition des LV beinhalten. Insgesamt konnten für den Inhaltsbereich „Quantität“ im NEPS mehr Inhalte als für den Inhaltsbereich „Zahl“ im LV gefunden werden. Die Bereiche „Daten und Zufall“ aus dem NEPS und „Daten und Zufall“ aus dem LV sowie „Veränderung und Beziehungen“ aus NEPS und

„Funktionaler Zusammenhang“ aus dem LV scheinen sehr ähnlich definiert zu sein. Die Testung auf signifikante Unterschiede zwischen den Studien mit dem Chi-Quadrattest war nicht möglich, da die Rahmenkonzeptionen eine unterschiedliche Anzahl an Kategorien definieren. Es konnten allerdings relevante Teildimensionen in den Aufgaben des NEPS-Tests identifiziert werden. Außerdem ließ sich feststellen, dass die Inhaltsbereiche in den jeweiligen Rahmenkonzeptionen ähnlich definiert sind, wobei der Ländervergleich mit der Definition des Inhaltsbereiches „Messen“ einen stärkeren Fokus auf diese Inhalte legt.

Für die Teilbereiche Prozesse lagen keine Einordnungen in die NEPS-Rahmenkonzeption vor. Aus diesem Grund konnte die Operationalisierung der Teildimensionen an dieser Stelle nicht verglichen werden.

Ergebnisse zu H3: Gewichtung der Komponenten

Für die Prüfung der Hypothese 3 „Die Gewichtung der Komponenten aus den PISA- und LV-Rahmenkonzepten im NEPS-K9-Mathematiktest unterscheidet sich nicht signifikant von der Gewichtung in den Mathematiktests aus den Haupterhebungen PISA 2012 und LV 2012“ wurden die prozentualen Verteilungen der NEPS-Mathematikaufgaben in den Teilbereichen der LV- und PISA-Rahmenkonzeption den Verteilungen der LV- und PISA-Aufgaben in der LV- und PISA-Rahmenkonzeption gegenübergestellt [3].

Gewichtung der PISA-Teilkompetenzen in den Mathematikaufgaben aus NEPS und PISA 2012. Die prozentuale Verteilung der Mathematikaufgaben aus PISA 2012 und aus NEPS-K9 auf die PISA-Inhaltsbereiche ist in Tabelle 7 dargestellt. Es wurden keine signifikanten Unterschiede zwischen den Gewichtungen der inhaltlichen Teildimensionen gefunden ($\chi^2 = 1.78$ $df = 3$, ns). Die Mathematikaufgaben verteilen sich relativ gleichmäßig über die Inhaltsbereiche. In NEPS wurde eine tendenziell stärkere Gewichtung für den Inhaltsbereich „Raum und Form“ und eine geringere Gewichtung für den Inhaltsbereich „Daten und Zufall“ gefunden.

Die prozentuale Verteilung der NEPS-Mathematikaufgaben in den Prozessen der PISA-Rahmenkonzeption ist in Tabelle 8 dargestellt. Für den Prozess „Argumentieren“ wurde

[3] Teile dieser Auswertungen wurden bereits in van den Ham et al. (2014) veröffentlicht

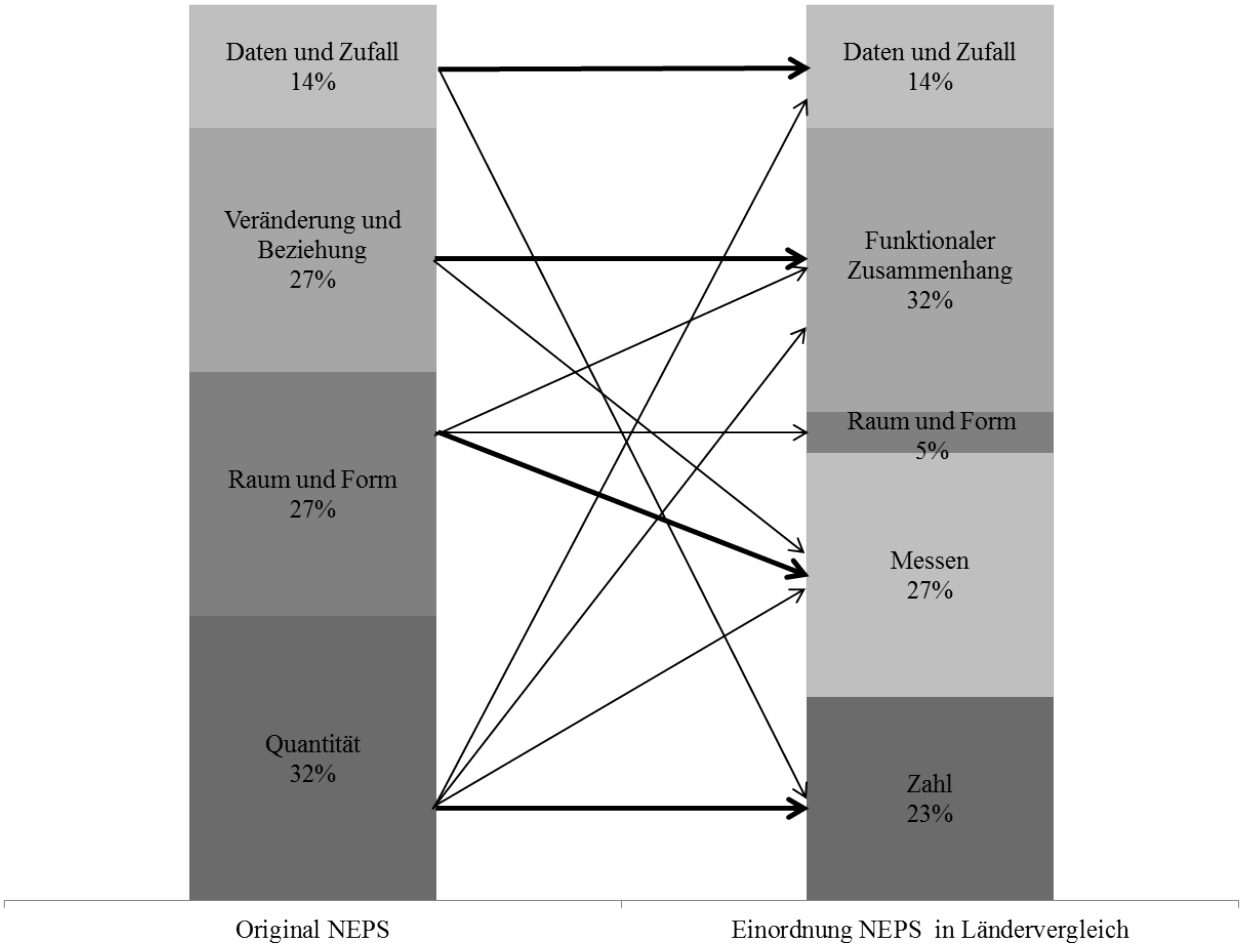


Abbildung 24: Prozentuale Verteilung der Mathematikaufgaben aus NEPS in die Inhaltsbereiche aus NEPS und LV

mit nur vier Aufgaben eine geringe Gewichtung gefunden. Der Grund hierfür lässt sich in der Rahmenkonzeption des NEPS-Testes finden. Der NEPS-Mathematiktest besteht aus 21 Aufgaben mit geschlossenem Format und einer halboffenen Aufgabe (Duchhardt & Gerdes, 2013). Dementsprechend beschränken sich die Aufgaben im Bereich „Argumentieren“ auf das Nachvollziehen und Bewerten von gegebenen Aussagen und Argumenten (Ehmke et al., 2009). Die stärksten Gewichtungen im NEPS konnten für die Prozesse „Mathematisieren“ und „Mit Mathematik symbolisch, formal und technisch umgehen“ nachgewiesen werden. Nur für die Lösung von drei beziehungsweise vier der 22 Mathematikaufgaben werden diese Prozesse nicht benötigt. Den meisten Prozessen konnten die Experten mindestens 20% der NEPS-Aufgaben zuordnen, sodass davon ausgegangen

Tabelle 7: Vergleich der prozentualen Verteilung der Mathematikaufgaben aus PISA 2012 und aus NEPS-K9 auf die PISA-Inhaltsbereiche

	PISA 2012-Mathematik		NEPS-K9-Mathematik	
PISA-Inhaltsbereiche	Anzahl	Prozent	Anzahl	Prozent
Quantität	29	26	5	23
Raum und Form	27	25	8	36
Veränderung und Beziehungen	29	26	6	27
Unsicherheit und Daten	25	23	3	14
Gesamt	110	100	22	100

werden kann, dass der NEPS-Test diese wichtigen Dimensionen abdeckt. Die Verteilung der PISA 2012-Aufgaben über die Prozesse war nicht bekannt, wodurch kein direkter Vergleich der Gewichtungen zwischen den Studien möglich war.

In Tabelle 9 wird die Einordnung der NEPS-Mathematikaufgaben aus dem Expertenreview in die Anforderungsbereiche der PISA-Rahmenkonzeption dargestellt und der Verteilung der Mathematikaufgaben aus PISA 2012 gegenübergestellt. Die Verteilung der NEPS-Mathematikaufgaben in den PISA-Anforderungsbereichen unterscheidet sich nicht signifikant von der Verteilung der PISA-Mathematikaufgaben aus dem Jahr 2012 auf die Anforderungsbereiche ($\chi^2 = 1.44$, $df = 2$, ns). Eine tendenziell stärkere Gewichtung konnte für den Prozess „Mathematische Konzepte, Prozeduren und Überlegungen anwenden“ im NEPS gefunden werden. Eine schwächere Gewichtung zeigte sich hingegen für den Prozess „Situationen mathematisch formulieren“.

Die Verteilung der NEPS-Mathematikaufgaben in den PISA-Kontexten unterscheidet sich signifikant von der Verteilung der PISA-Mathematikaufgaben aus dem Jahr 2012 ($\chi^2 = 19.10$, $df = 3$, $p < .01$). Während sich die PISA-Aufgaben aus 2012 relativ gleichmäßig über die verschiedenen Kontexte verteilen, wobei der gesellschaftliche Kontext leicht stärker gewichtet wird, wurden die NEPS-Mathematikaufgaben von den Experten vor allem dem persönlichen Kontext zugeordnet (Tabelle 10). Der berufliche, der gesellschaftliche und der wissenschaftliche Kontext wird jeweils nur durch wenige Aufgaben abgedeckt.

Tabelle 8: Prozentuale Verteilung der NEPS-Mathematikaufgaben in den Prozessen der PISA-Rahmenkonzeption

NEPS-K9-Mathematik		
PISA-Prozesse	Anzahl	Prozent
Kommunizieren	15	68
Mathematisieren	18	82
Repräsentieren	7	32
Argumentieren	4	18
Problemlösestrategien entwickeln	6	27
Mit Mathematik symbolisch, formal und technisch umgehen	17	77
Mathematische Hilfsmittel verwenden	0	0

Tabelle 9: Vergleich der prozentualen Verteilung der NEPS-Mathematikaufgaben und der Mathematikaufgaben aus PISA 2012 in den Anforderungsbereichen der PISA-Rahmenkonzeption

	PISA 2012-Mathematik		NEPS-K9-Mathematik	
PISA - math. Anforderungsbereiche	Anzahl	Prozent	Anzahl	Prozent
Situationen mathematisch formulieren	32	29	4	18
Mathematische Konzepte, Prozeduren und Überlegungen anwenden	51	46	13	59
Mathematische Ergebnisse interpretieren, anwenden und bewerten	27	25	5	23
Gesamt	110	100	22	100

Fett hervorgehobene Kennwerte zeigen statistisch signifikante Unterschiede zwischen PISA und NEPS an ($p < .05$)

Tabelle 10: Vergleich der prozentualen Verteilung der NEPS-Mathematikaufgaben und der Mathematikaufgaben aus PISA 2012 in den Aufgabenkontexten der PISA-Rahmenkonzeption

	PISA 2012-Mathematik		NEPS-K9-Mathematik	
PISA-Kontexte	Anzahl	Prozent	Anzahl	Prozent
Persönlicher Kontext	21	19	14	**64**
Beruflicher Kontext	24	22	3	**14**
Gesellschaftlicher Kontext	36	33	2	**9**
Wissenschaftlicher Kontext	29	26	3	14
Gesamt	110	100	22	100

Fett hervorgehobene Kennwerte zeigen statistisch signifikante Unterschiede zwischen PISA und NEPS an ($p < .05$)

Gewichtung der NEPS-Mathematikaufgaben in den LV-Teilkompetenzen. Tabelle 11 stellt die prozentuale Verteilung der NEPS-Mathematikaufgaben auf die LV-Inhaltsbereiche der Verteilung der Mathematikaufgaben aus dem LV 2012 auf die LV-Inhaltsbereiche gegenüber. Die Verteilungen unterscheiden sich nicht statistisch signifikant voneinander ($\chi^2 = 5.01$ $df = 4$, ns). Der Vergleich verdeutlicht jedoch noch einmal den Unterschied zwischen den „Raum und Form“-Inhaltsbereichen beider Studien. Im NEPS konnte nur eine Aufgabe dem Bereich „Raum und Form“, wie er im LV definiert wird, zugeordnet werden. Dagegen decken im LV 2012 17 % der Aufgaben diesen Bereich ab. Entsprechend konnten im NEPS 12 % mehr Aufgaben mit den Inhalten, die im LV dem Inhaltsbereich „Messen“ angehören, gefunden werden. Für die übrigen Inhaltsbereiche wurden mit einem maximalen Unterschied von 7 % sehr ähnliche Gewichtungen gefunden.

Tabelle 11: Vergleich der prozentualen Verteilung der Mathematikaufgaben aus LV 2012 und aus NEPS-K9 auf die LV-Inhaltsbereiche

	LV 2012-Mathematik		NEPS-K9-Mathematik	
LV - Inhaltsbereiche	Anzahl	Prozent	Anzahl	Prozent
Zahl	67	22	5	23
Messen	44	15	6	27
Raum und Form	52	17	1	5
Funktionaler Zusammenhang	75	25	7	32
Daten und Zufall	62	21	3	14
Gesamt	300	100	22	100

In den NEPS-Mathematikaufgaben konnten alle wichtigen Prozesse aus dem LV identifiziert werden. Für die Lösung der meisten Mathematikaufgaben (86 %) wurde der Prozess „mathematisch modellieren“ als notwendig identifiziert. Auch die Prozesse „Mit Mathematik symbolisch, formal, technisch umgehen“ und „Kommunizieren“ wurden von den Experten in über 70% der Aufgaben identifiziert. Die übrigen Prozesse wurden in rund 30% der NEPS-Aufgaben als Voraussetzung für die Bearbeitung angesehen. Die Prozesse aus dem LV werden durch die NEPS-Mathematikaufgaben abgedeckt.

Die Gewichtung der NEPS-Mathematikaufgaben in den LV-Prozessen konnte der Ver-

Tabelle 12: Prozentuale Verteilung der NEPS-Mathematikaufgaben in den Prozessen der LV-Rahmenkonzeption

	NEPS-K9-Mathematik	
LV - Prozesse	Anzahl	Prozent
Mathematisch argumentieren	6	27
Probleme mathematisch lösen	6	27
Mathematisch modellieren	19	86
Mathematische Darstellungen verwenden	7	32
Mit Mathematik symbolisch, formal, technisch umgehen	17	77
Mathematisch Kommunizieren	16	73

teilung der Aufgaben aus LV 2012 nicht gegenübergestellt werden, da die Einordnung der LV-Aufgaben in die Prozesse nicht vorlag.

Die meisten Aufgaben des NEPS (64%) konnten dem Bereich „Zusammenhänge herstellen“ zugeordnet werden (Tabelle 13). Dass die Aufgaben aus dem NEPS dem Bereich „Verallgemeinern und Reflektieren“ nicht zugeordnet werden konnten, liegt insbesondere am geschlossenen Aufgabenformat der NEPS-Aufgaben. In der LV-Rahmenkonzeption werden diesem Anforderungsbereich vor allem kognitive Aktivitäten zugeordnet, die in geschlossenen und halboffenen Antwortformaten schwierig zu realisieren sind, wie zum Beispiel das Nutzen, Erläutern und Entwickeln von komplexen Argumentationen oder das Reflektieren über Lösungswege. Dass für die NEPS-Aufgaben dieser Prozess nicht benötigt wird, bedeutet nicht, dass die Aufgaben deswegen einfacher sind. Auch die anderen zwei Prozesse können eine empirisch hohe Schwierigkeit aufweisen. Es muss jedoch festgehalten werden, dass der national als wichtig erachtete Prozess „Verallgemeinern und Reflektieren“ nicht in den NEPS-Aufgaben identifiziert werden konnte. Da in der NEPS-Rahmenkonzeption keine Anforderungsbereiche definiert werden, konnte auch die Operationalisierung solcher Anforderungsbereiche nicht analysiert werden. Ein Vergleich der Gewichtungen mit dem LV 2012 war nicht möglich, da die prozentuale Verteilung der Anforderungsbereiche im LV nicht vorlag.

Tabelle 13: Prozentuale Verteilung der NEPS-Mathematikaufgaben in den Anforderungsbereichen der LV-Rahmenkonzeption

LV - Anforderungsbereiche	Anzahl	Prozent
Reproduzieren	8	36
Zusammenhänge herstellen	14	64
Verallgemeinern und Reflektieren	0	0
Gesamt	22	100

4.1.3 Diskussion und Fazit *Domänenbeschreibung*

Der Validierungsschritt „Domänenbeschreibung“ basiert auf dem Argument, dass die beobachteten Leistungen im NEPS-Mathematiktest relevante und repräsentative Inhalte sowie kognitive Prozesse für die Zieldomäne Mathematische Kompetenz widerspiegeln. Um dieses Argument zu evaluieren, wurden die Hypothesen bezüglich der Identifizierung relevanter Teilkompetenzen, Operationalisierung der Teilkompetenzen und Gewichtung der Teilkompetenzen getestet.

In den Aufgaben des NEPS-K9-Mathematiktests konnten relevante Teilkompetenzen mathematischer Kompetenz aus den PISA- und LV-Rahmenkonzeptionen identifiziert werden. Keine der NEPS-Aufgaben beinhaltet von der PISA- bzw. LV-Rahmenkonzeption abweichende Inhaltsbereiche, Prozesse, Anforderungsbereiche oder Kontexte. Die NEPS-Aufgaben beinhalten jedoch nicht den PISA-Prozess „mathematische Hilfsmittel verwenden“ und den LV-Anforderungsbereich „Verallgemeinern und Reflektieren“. Im Gegensatz zu PISA werden im NEPS die Fähigkeiten in den Bereichen „mathematische Hilfsmittel verwenden“ nicht als eigenständiger Prozess definiert, sondern durch mehrere Prozesse abgedeckt.
Die inhaltlichen Teilkompetenzen im NEPS werden ähnlich operationalisiert wie die inhaltlichen Teilkompetenzen im LV und in PISA. Es wurden keine signifikanten Unterschiede zwischen der Verteilung der NEPS-Aufgaben auf die NEPS-Inhaltsbereiche und der Verteilung der NEPS-Aufgaben auf die PISA-Inhaltsbereiche gefunden. In der Rahmenkonzeption des LV wird jedoch der Inhaltsbereich „Messen“ als eigenständige Kompetenz formuliert, wohingegen diese Kompetenz im NEPS über mehrere Inhaltsbereiche verteilt ist. Die übrigen Inhaltsbereiche sind sehr ähnlich operationalisiert. Es wur-

de kein signifikanter Unterschied zwischen der Verteilung der NEPS-Mathematikaufgaben in den Inhaltsbereichen und Anforderungsbereichen der PISA-Rahmenkonzeption und der Gewichtung dieser Teilkompetenzen in der PISA-2012-Studie gefunden. Ebenso wurden keine signifikanten Unterschiede zwischen der Verteilung der NEPS-Mathematikaufgaben in den Inhaltsbereichen der LV-Rahmenkonzeption und der Gewichtung dieser Teilkompetenzen in der LV 2012-Studie aufgezeigt. Zwischen den Mathematiktests aus NEPS und PISA 2012 wurde hingegen ein signifikanter Unterschied bezüglich der Gewichtung der PISA-Kontexte nachgewiesen, da die NEPS-Aufgaben hauptsächlich dem persönlichen Kontext aus PISA zugeordnet wurden. Die Gewichtungen der LV-Anforderungsbereiche und der LV- und PISA-Prozesse konnten nicht verglichen werden, da die Einordnung der LV- und PISA-Aufgaben in die Prozesse nicht vorlag.

Die hier durchgeführten Analysen basieren auf der Annahme, dass die PISA- und LV-Domänen sowie ihre Gewichtung jeweils die Zieldomäne mathematischer Kompetenz für den NEPS-Mathematiktest repräsentieren. Diese Annahme liegt darin begründet, dass der NEPS-Mathematiktest sowohl das Konzept Mathematical Literacy basierend auf PISA als auch grundlegende curriculare Kompetenzen basierend auf dem LV erfassen soll (vgl. Kapitel 1.2). Das sorgfältige Vorgehen bei der Entwicklung der Rahmenkonzeptionen und der Testaufgaben in PISA und LV bekräftigt die Annahme, dass die Inhalte dieser beiden Tests wichtige Teilkompetenzen der Zieldomäne des NEPS-Tests repräsentieren (vgl. Kapitel 1.2). Die hier durchgeführte Untersuchung liefert daher aussagekräftige Validitätsevidenz.
Wünschenswert wären jedoch noch weitere Studien, welche die durch die NEPS-Aufgaben erfassten Inhalte, Prozesse, Anforderungen und Kontexte noch detaillierter untersuchen. Denkbar wären beispielsweise Think-Aloud-Studien zur Untersuchung der Inhalte, Prozesse etc., welche durch die Neuntklässlerinnen und Neuntklässler tatsächlich zum Lösen der Aufgaben angewendet werden.

Insgesamt konnten Hinweise für die Validität der Annahme: „Die Aufgaben des NEPS-Mathematiktests decken die für die Zieldomäne relevanten Teilkompetenzen angemessen ab.“ gefunden werden. So ließen sich die relevanten Teilkompetenzen mathematischer Kompetenz aus den PISA- und LV-Rahmenkonzepten in den Aufgaben des NEPS-K9-Mathematiktests identifizieren. Auch wurden ähnliche Operationalisierungen der inhaltlichen Teilkompetenzen im NEPS und der inhaltlichen Teilkompetenzen aus PISA und LV gefunden. Des Weiteren wurden keine signifikanten Unterschiede zwischen der Gewichtung der Inhaltsbereiche aus dem LV 2012 sowie der Gewichtung der Inhalts- und

Anforderungsbereiche aus PISA 2012 und der Gewichtung dieser Teilkompetenzen im NEPS-Test nachgewiesen. Insgesamt zeigt der NEPS-Mathematiktest also große Überschneidungen mit den PISA- und LV-Mathematikmessungen. Der NEPS-Test scheint jedoch ein ein etwas weniger breites Konzept mathematischer Kompetenz mit leicht unterschiedlichen Aufgabenanforderungen zu erfassen, welches die aktiven Komponenten der Prozesse aus PISA und LV, die Kontexte aus PISA und die Anforderungsbereiche aus LV nur eingeschränkt abdeckt. Bei der Interpretation der Testergebnisse ist es daher von großer Wichtigkeit, zu beachten, dass nur die Aspekte der Zieldomäne Mathematische Kompetenz mit dem NEPS-Mathematiktest erfasst werden, welche durch die NEPS-Aufgaben repräsentiert werden. Daher können auf die genannten Teilkomponenten der LV- und PISA-Rahmenkonzeption keine Rückschlüsse gezogen werden.

4.2 Bewertung

Die Argumentationsstruktur für die Schlussfolgerung *Bewertung* wurde in Kapitel 2.2 entwickelt und beschrieben. Eine Darstellung der Argumente, Annahmen und Hypothesen wird in Kapitel 2.2.6 in der Abbildung 17 aufgezeigt. Nachfolgend wird die Methode für die Evaluation der Hypothesen für die Schlussfolgerung *Bewertung* beschrieben. Anschließend werden die Ergebnisse zu den Gewichtungen der Aufgabenformate und zum Umgang mit fehlenden Werten dargelegt. Außerdem wird analysiert, ob Fehler in der Aufgabenkodierung ausgeschlossen werden können. Daraufhin werden die Ergebnisse zu der psychometrischen Qualität des Mathematiktests zusammengetragen. Schließlich werden die Ergebnisse kurz zusammengefasst und diskutiert.

4.2.1 Methode

Das Argument der Schlussfolgerung *Bewertung* „Die Rohwerte in den NEPS-Mathematikaufgaben führen zu Testergebnissen, die repräsentativ für die Zieldomäne mathematische Kompetenz sind.“ basiert auf drei Annahmen und fünf Hypothesen (vgl. Kapitel 2.2). Im Folgenden wird das methodische Vorgehen zur Prüfung dieser fünf Hypothesen beschrieben.

Anwendung der Aufgabenkodierung

Für die Analyse der ersten Hypothese „Fehler in der Eingabe und Kodierung der Testhefte aller Schülerinnen und Schüler können ausgeschlossen werden“ wurde ein Review des methodischen Berichts für die Startkohorte 4 (IEA Data Processing and Research Center, 2013) und des technischen Berichts für den NEPS-K9-Mathematiktest (Duchhardt & Gerdes, 2013) durchgeführt. Das Sicherstellen einer korrekten Eingabe und Kodierung der Testhefte ist vor der Testung vorbereitet und bei der Dateneingabe und Kodierung gewährleistet worden. Der Methodenbericht für die Startkohorte 4 und der technische Bericht für den NEPS-K9-Mathematiktest enthalten Informationen über die Eingabe und Kodierung sowie Bewertung der NEPS-K9-Daten. Diese Informationen wurden in den Ergebnissen zusammengetragen und anhand von Qualitätskriterien aus der Literatur für die Eingabe-, Kodierungs- und Bewertungsprozeduren untersucht.

Angemessenheit der Aufgabenkodierung

Die zweite Hypothese „Die unterschiedlichen Gewichtungen der Aufgabenformate und der Umgang mit fehlenden Werten führen zu unverfälschten Testergebnissen.“ wurde mit Hilfe eines Reviews von Dokumenten bezüglich der Gewichtungen von Aufgabenformaten und den Antwortformaten im NEPS untersucht. In diesen Dokumenten sind bereits die Auswertungen und Ergebnisse, die zur Evaluation der Hypothese benötigt wurden, beschrieben worden:

- Pohl & Carstensen, Scaling of competence tests in the National Educational Panel Study – Many questions, some answers, and further challenges, 2013
- Pohl, Gräfe & Rose, Dealing with omitted and not reached items in competence tests – Evaluating approaches accounting for missing responses in IRT models, 2014
- Haberkorn, Pohl, Carstensen & Wiegand, Incorporating different response formats of NEPS competence tests in an IRT model, 2013;

Für die Evaluation der zweiten Hypothese wurden die Ergebnisse der zu diesem Thema veröffentlichten Dokumente zusammengetragen.

Psychometrische Itemeigenschaften

Für die Prüfung der dritten Annahme der *Bewertung* „Eine hohe psychometrische Qualität des Testinstruments ist gewährleistet.“ wurden die drei Hypothesen H3: „Die Aufgaben sind trennscharf.“, H4: „Die Qualität der Distraktoren ist angemessen.“ und H5: „Die Aufgaben sind intern konsistent.“ formuliert. Berechnungen zur Trennschärfe der NEPS-K9-Mathematikaufgaben sowie Auswertungen zur Qualität der Distraktoren sind bereits im technischen Bericht für den NEPS-K9-Mathematiktest veröffentlicht worden. Aus diesem Grund wurden die Ergebnisse aus dem technischen Bericht für die Evaluation dieser Hypothesen zusammengetragen.

Die interne Konsistenz der Aufgaben wurde anhand des Cronbach-α-Koeffizienten berechnet (Schermelleh-Engel & Werner, 2012; Cronbach, 1951). Diese Berechnung ist eine Verallgemeinerung der Testhalbierungsmethode. Dabei wird der Test mit m Items in m Teile zerlegt, wobei jedes Item als separater Testteil betrachtet wird. Eine Voraussetzung für die korrekte Schätzung von Cronbachs α ist die essentielle τ-Äquivalenz. Von einer essentiellen τ-Äquivalenz kann gesprochen werden, wenn die wahren Werte (τ_i) aller Items (x_i, $i = 1, ..., m$) aus einem über alle Items gleichen wahren Wert und einer itemspezifischen, additiven Konstante bestehen (Schermelleh-Engel & Werner, 2012)):

$$\tau_i = \tau + c_i (i = 1, ..., m).$$

Cronbachs α berechnet sich dann wie folgt:

$$\alpha = \frac{m}{m-1} \left(1 - \frac{\sum_{i=1}^{m} Var(x_i)}{Var(x)} \right)$$

Die Höhe des Cronbach-α-Koeffizienten hängt von der Stärke der Korrelation der Testteile untereinander ab. Je höher diese Korrelationen ausfallen, desto größer wird der Koeffizient (Moosbrugger & Kelava, 2012). Cronbachs α wird bei der Skalierung der Daten der NEPS-Haupterhebung mit der Software Conquest berechnet.

4.2.2 Ergebnisse

Die Ergebnisse zur Repräsentativität der Rohwerte des NEPS-K9-Mathematiktests werden im Folgenden für die fünf Hypothesen dargelegt.

Ergebnisse zu H1: Ausschließen von Fehlern in der Eingabe- und Kodierung

Für die Analyse der ersten Hypothese „Fehler in der Eingabe und Kodierung der Testhefte aller Schülerinnen und Schüler können ausgeschlossen werden.“ wurden die Informationen bezüglich dieser Prozeduren aus dem methodischen Bericht für die Startkohorte 4 (IEA Data Processing and Research Center, 2013) und aus dem technischen Bericht für den NEPS-K9-Mathematiktest (Duchhardt & Gerdes, 2013) anhand von Qualitätskriterien aus der Literatur bewertet.

In der Literatur werden (1) standardisierte Eingabe- und Kodierungsprozeduren empfohlen, um eine konsistente Aufgabenkodierung sicher zu stellen. Dafür sollten präzise Anweisungen entwickelt werden (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 2014). Außerdem sollten Daten auf ihre (2) Integrität geprüft werden, um so mögliche Eingabefehler aufzudecken. Dafür können die Daten auf Vollständigkeit, große Mengen an Missing Response, Abweichungen in den Verteilungen, Antwortschlüssel und Iteminformationen, Anzahl der Schülerinnen und Schüler, Anzahl Jungen und Mädchen und Anzahl Schulen geprüft werden (Cohen & Wollack, 2006). Zusätzlich sollte (3) die Richtigkeit der Kodierung zu Beginn sichergestellt und während des Prozesses beobachtet werden. Periodische Kontrollen statistischer Eigenschaften, wie zum Beispiel Mittelwerte, Standardabweichung etc., können die Kodierer als Feedback für die Richtigkeit der Daten geben (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 2014).

Die Eingabe und Kodierung der Testhefte erfolgte standardisiert und anhand von präzisen Vorgaben. So wurden die Testhefte im Erhebungsinstitut scannergestützt eingegeben. Da der NEPS-K9-Mathematiktest vor allem geschlossene und nur eine halboffene Aufgabe beinhaltet, war bei der Eingabe jedes Testheftes nur einmal die Eingabe einer freien Antwort in Form einer Zahl vorzunehmen. Anschließend wurden die Daten auf-

bereitet und an das Datenzentrum übergeben. Variablen, Variablennamen, Werte und Wertebereiche wurden dabei durch Codebooks definiert. Die Codebooks wurden von den Erhebungskoordinatorinnen und Erhebungskoordinatoren zur Verfügung gestellt. Es ist davon auszugehen, dass bei der Aufbereitung der Daten diese auch in gewisser Weise auf ihre Integrität geprüft wurden. Genaue Informationen über die Prüfung der Integrität der Daten wurden bisher jedoch nicht veröffentlicht. Die Bewertung der NEPS-K9-Aufgaben wurde anhand einer automatisierten Syntax durchgeführt. Dabei wurde die Bewertung durch das Bilden von Kreuztabellen zwischen den Rohwerten und den bewerteten Aufgaben und durch das bilden von Kontrollvariablen auf Fehler überprüft (IEA Data Processing and Research Center, 2013). Durch die Standardisierung im Eingabeprozess und Kodierungsprozess sowie durch konkrete Vorgaben anhand von Codebooks für die Aufbereitung der Daten sind in diesen Schritten keine systematischen Fehler zu erwarten.

Ergebnisse zu H2: Gewichtung von Aufgabenformaten und Umgang mit fehlenden Werten

Analysen zur Evaluation der Hypothese H2: „Die unterschiedlichen Gewichtungen der Aufgabenformate und der Umgang mit fehlenden Werten führen zu unverfälschten Testergebnissen.“ sind bereits in drei Artikeln veröffentlicht worden. Nachfolgend werden daher die Ergebnisse zur Gewichtung der Aufgabenformate aus den Artikeln „Scaling of competence tests in the National Educational Panel Study – Many questions, some answers, and further challenges“ von Pohl und Carstensen (2013) und „Incorporating different response formats of NEPS competence tests in an IRT model“ von Haberkorn, Pohl, Carstensen und Wiegand (2013) berichtet. Anschließend werden die Ergebnisse zum Umgang mit fehlenden Werten aus dem Artikel „Dealing with omitted and not reached items in competence tests – Evaluating approaches accounting for missing responses in IRT models“ von Pohl et al. (2014) zusammengefasst.

Haberkorn et al. (2013) untersuchten die Passung unterschiedlicher Bewertungsregeln für Aufgaben mit CMC-Format und Matching-Aufgaben für die Kompetenzdaten im NEPS und die Gewichtung dieser Aufgabenformate im Vergleich zu den dichotom gewerteten MC-Aufgaben. Die Autoren fanden eine höhere Diskrimination für die Aufgaben, wenn die Anzahl der richtigen Antworten einer Aufgabe bei der Bewertung berücksichtigt

wurde. Eine Bewertungsregel, bei der eine Aufgabe nur dann als richtig zählt, wenn alle Subantworten korrekt sind, führte zu einem Informationsverlust. Des Weiteren wurde untersucht, welche Bewertungsregel am besten zu den Kompetenzdaten passt. Dafür wurden zwei Bewertungsregeln mittels des Weighted Mean Squares (WMNSQ) verglichen. Die erste Regel wertet jede korrekt gelöste Subaufgabe mit einem Punkt und die zweite Regel wertet jede korrekt gelöste Subaufgabe mit einem halben Punkt. Für die zweite Bewertungsregel wurden bessere WMNSQ-Werte gefunden. Die Ergebnisse wurden durch eine 2PL-Analyse der Daten bestätigt. Die mit dem 2PL-Modell berechneten Trennschärfen waren den Trennschärfen aus dem 1PL-Modell mit der zweiten Bewertungsregel sehr ähnlich. Die Ergebnisse ließen sich für verschiedene Alterskohorten und für verschiedene Kompetenzdomänen reproduzieren (Pohl & Carstensen, 2013).
Die im NEPS-K9-Mathematiktest gehandhabte Bewertungsregel für die zwei CMC-Aufgaben mit einem halben Punkt für jede korrekte Subaufgabe ist daher angemessen.

Pohl et al. (2014) untersuchten die Verwendung unterschiedlicher Modelle für die Modellierung von fehlenden Werten durch Auslassen und nicht Erreichen von Aufgaben. Die Datengrundlage dieser Untersuchung waren 4 976 Schülerinnen und Schüler, die das NEPS-Lesetestheft für die Klassenstufe 5 (K5) bearbeiteten und 5 194 Schülerinnen und Schüler, die das NEPS-Mathematiktestheft für die K5 bearbeiteten. Die Autoren untersuchten die Eindimensionalität der latenten Modellierung der fehlenden Werte, da implizit davon ausgegangen wurde, dass die fehlenden Werte ein eindimensionales Konstrukt der Tendenz zum Auslassen messen. Um diese Annahme zu testen, wurden die fehlenden Werte in einem Rasch-Modell skaliert und Item-Fit Kriterien wie WMNSQ, ICC und die Trennschärfe ausgewertet. Diese Berechnungen wurden für den Mathematik- und den Lesetest separat durchgeführt. Um die Möglichkeit des Behandelns fehlender Werte als *systembedingt fehlend* zu testen, wurde der Zusammenhang zwischen den *Missing Propensities* und der Fähigkeit in verschiedenen Modellen analysiert. Dafür wurde die Korrelation der Fähigkeit mit der latenten *Propensity* in einem latenten Modell berechnet sowie die manifeste Korrelation aus dem Regressionsschätzer in einem manifesten Modell. Des Weiteren untersuchten die Autoren die Leistung fünf unterschiedlicher Modelle: Bewerten der fehlenden Werte als *falsch* (1), Behandeln fehlender Werte als *systembedingt fehlend* für die Schätzung der Itemparameter und Bewerten der fehlenden Werte als *falsch* für die Schätzung der Personenparameter (2), Behandeln fehlender Werte als *systembedingt fehlend* für die Schätzung (3), modellbasierter, manifester Ansatz (4) und modellbasierter latenter Ansatz (5). Die Analysen wurden getrennt für die ausgelasse-

nen Items, für die nicht erreichten Items, für die ausgelassenen und nicht erreichten Items zusammen sowie für die ausgelassenen und nicht erreichten Items in einem Modell getrennt durchgeführt. Zusätzlich wurde eine Simulationsstudie durchgeführt. Dafür wurden 1 443 Testpersonen ohne fehlende Werte ausgewählt und die Item- und Personenparameter in einem Rasch-Modell geschätzt. In einem weiteren Schritt wurden systematisch fehlende Werte eingeführt und die unterschiedlichen Modelle zum Umgang mit diesen fehlenden Werten wurden evaluiert. Die Ergebnisse bezüglich der Modellannahmen zeigten, dass die fehlenden Werte durch Auslassen zu einem eindimensionalen Raschmodell passen. Die Personenfähigkeit korrelierte sowohl mit den ausgelassenen als auch mit den nicht erreichten Aufgaben. Die Autoren schlossen daraus, dass fehlende Werte nicht ignoriert werden können. Pohl et al. (2014) fanden unterschiedliche Item- und Personenparameterschätzungen für das Modell, in dem die fehlenden Werte als *falsch* bewertet werden und das Modell, in dem die fehlenden Werte als *systembedingt fehlend* behandelt wurden. Zwischen dem Behandeln fehlender Werte als *systembedingt fehlend* und den anderen modellbasierten Ansätzen wurde kein Unterschied gefunden. Die IRT-Modellierung durch das Behandeln fehlender Werte als *systembedingt fehlend* schien daher robust bei Verstößen gegen die Annahme der Ignorierbarkeit zu sein. Die zusätzliche Simulationsstudie bestätigte diese Ergebnisse. Die Personenparameterschätzungen aus dem IRT-Modell, welches fehlende Werte als *systembedingt fehlend* behandelt, korrelierten ebenso hoch mit dem kompletten Datensatz wie die Schätzungen aus den modellbasierten Ansätzen. Das Bewerten der fehlenden Werte als *falsch* führte zu Fähigkeitsschätzern, die sich deutlich von den Schätzern des vollständigen Datensatzes unterschieden. Die Autoren berichteten außerdem, dass für den NEPS-K9-Mathematiktest und Lesetest ähnliche Ergebnisse gefunden wurden.
Diese Ergebnisse sprechen dafür, dass das Behandeln fehlender Werte als *systembedingt fehlend*, wie es im NEPS-K9-Mathematiktest gehandhabt wird, zu unverfälschten Ergebnissen führt.

Ergebnisse zu H3: Trennschärfe der Aufgaben und H4: Qualität der Distraktoren

Analysen zur Evaluation der Hypothesen H3 und H4 wurden im technischen Bericht für den NEPS-K9-Mathematiktest durchgeführt (Duchhardt & Gerdes, 2013). Für die Analyse zu Hypothese 3 „H3: Die Aufgaben sind trennscharf“ wurden die punktbiseralen

Korrelationen zwischen den Antwortkategorien und den Personenfähigkeiten berechnet. Die kleinste gefundene Korrelation betrug $r = .34$, die höchste Korrelation betrug $r = .54$. Der Mittelwert der gefundenen punktbiseralen Korrelationen lag bei $r = .44$. Insgesamt wurden diese Werte als zufriedenstellend interpretiert (Duchhardt & Gerdes, 2013). Zur Prüfung der Hypothese 4 „H4: Die Qualität der Distraktoren ist angemessen.“ wurde berechnet, ob die Distraktoren öfter von Schülerinnen und Schülern mit einer niedrigeren Fähigkeit gewählt wurden als von Schülerinnen und Schülern mit höheren mathematischen Kompetenzsschätzungen. Dafür wurden die punktbiseralen Korrelationen zwischen den Distraktoren und der Personenfähigkeit berechnet. Eine hohe positive Korrelation weist auf eventuelle Uneindeutigkeiten gegenüber der korrekten Antwortmöglichkeit hin. Eine negative oder Null-Korrelation weist hingegen auf eine gute Passung hin. Zwei Distraktoren wiesen eine positive Korrelation mit der Personenfähigkeit auf (0.08 und 0.1). Aus theoretischen Gesichtspunkten war es jedoch wünschenswert, diese Distraktoren in die Analysen einzuschließen. Ein weiterer Distraktor wies eine Korrelation von 0 auf, was als unkritisch eingestuft wurde. Die restlichen Korrelationen lagen unter dem Wert 0 und wurden als angemessen eingestuft (Duchhardt & Gerdes, 2013; Pohl & Carstensen, 2012).

Ergebnisse zu H5: Interne Konsistenz der Aufgaben

Die Berechnung der internen Konsistenz ergab einen Cronbach-α-Koeffizienten von $\alpha = .80$. Nach Bortz und Döring (2006) kann dieser Wert als zufriedenstellend interpretiert werden.

4.2.3 Diskussion und Fazit *Bewertung*

Ob die Gewichtung der Aufgabenformate und der Umgang mit fehlenden Werten zu unverfälschten Testergebnissen führen, wurde von Haberkorn et al. (2013), Pohl und Carstensen (2013) und Pohl et al. (2014) untersucht und veröffentlicht. Pohl und Carstensen (2013) kamen zu dem Schluss, dass bezüglich der Gewichtung der Aufgabenformate bei Partial Credit Aufgaben die im NEPS genutzte Bewertungsregel, welche jede gelöste Subaufgabe mit einem halben Punkt bewertet, zu den besten Itemkennwerten führt. Haberkorn et al. (2013) und Pohl et al. (2014) fanden heraus, dass das Ignorieren feh-

lender Werte, wie es in der NEPS-Studie angewandt wird, zu einer guten beziehungsweise besseren Modellpassung führt als andere Arten des Umgangs mit fehlenden Werten. Das Review der technischen und methodischen Dokumente zeigte auf, dass Fehlern in der Eingabe und Kodierung der Aufgaben durch hohe Standardisierung und Aufbereitung der Daten vorgebeugt wird. Der technische Bericht des NEPS-K9-Mathematiktests wies zusätzlich auf gute Trennschärfen der Items und eine gute Qualität der Distraktoren hin. Lediglich zwei Distraktoren korrelierten leicht positiv mit der Personenfähigkeit. Des Weiteren wiesen die Testaufgaben eine zufriedenstellende interne Konsistenz auf.

Im NEPS wurden keine ausführlichen Informationen zur Eingabe und Aufbereitung der Daten veröffentlicht. Es ist anzunehmen, dass nach der scannerunterstützten Dateneingabe und bei der Datenaufbereitung eine Prüfung auf Integrität stattgefunden hat. Es gibt jedoch keine öffentlichen Dokumente, die diesen Vorgang und die Ergebnisse einer solchen Prüfung beschreiben. Auch gibt es keine detaillierten Angaben zur Prozedur und Kontrolle des Scorings. Aufgrund des Fehlens solcher Dokumente wurde die Hypothese H2 anhand von logischer Argumentation basierend auf den durchgeführten Prozeduren gestützt. Fehler in der Eingabe und Kodierung der Testwerte konnten daher nicht vollständig ausgeschlossen werden. Da solche Fehler durch die beschriebenen Prozeduren sehr unwahrscheinlich sind, werden die Auswertungen als Hinweis für eine Validität der Schlussfolgerung betrachtet. Eine Veröffentlichung der oben genannten Qualitätsprüfungen ist trotzdem sehr wünschenswert.

Insgesamt konnten Hinweise für die Validität der Schlussfolgerung *Bewertung* gefunden werden. Die durchführten Analysen führten zu Ergebnissen, welche die Hypothesen H1 bis H5 bestätigen. Das Fazit für die Schlussfolgerung *Bewertung* ist, dass die Rohwerte in den NEPS-Mathematikaufgaben zu Testergebnissen führen, die repräsentativ für die Zieldomäne Mathematische Kompetenz sind.

4.3 Skalierung

Für die Schlussfolgerung *Skalierung* wurde die Argumentationsstruktur in Kapitel 2.3 entwickelt und beschrieben. Eine Darstellung der Argumente, Annahmen und Hypothesen wird in Kapitel 2.3.2, Abbildung 18 gezeigt. Nachfolgend wird die Methode für die Evaluation der Hypothesen für die Schlussfolgerung *Skalierung* beschrieben.

Im Ergebnisteil wird die Passung des IRT-Modells zu den Daten ausgewertet, indem der NEPS-K9-Mathematiktest auf Eindimensionalität, spezifische Objektivität, Rasch-Homogenität, lokale stochastische Unabhängigkeit und auf Abweichung der beobachteten Antwortwahrscheinlichkeit von der Modellvorhersage getestet wird. Abschließend werden die Befunde zusammengefasst und diskutiert.

4.3.1 Methode

Das Argument der Schlussfolgerung *Skalierung* „Das beobachtete Testergebnis führt zu Fähigkeitsschätzern, welche die Zieldomäne Mathematische Kompetenz widerspiegeln." basiert auf der Annahme, dass das verwendete Modell zu den Daten passt. Die Modellgleichung des Skalierungsmodells kann aus theoretischen Annahmen hergeleitet werden. Aus diesem Grund ist es nötig, diese Annahmen bei der Verwendung des Modells zu prüfen (Strobl, 2012). Im Folgenden wird die Methode für die fünf Hypothesen dargelegt, welche die Annahme stützen.

Spezifische Objektivität

Die spezifische Objektivität gewährleistet, dass Aussagen basierend auf einem Vergleich der Fähigkeiten unabhängig von den Aufgaben sind, anhand derer die Personen verglichen werden. Ist eine Aufgabe für zwei Personengruppen mit gleicher Fähigkeit unterschiedlich schwer, so spricht man von DIF. Für die Testung der spezifischen Objektivität können die Aufgaben daher auf DIF getestet werden. Aufgaben, die DIF ausweisen, messen in der Regel eine weitere Kompetenz, die sich für die Personengruppen unterscheidet (Strobl, 2012). Auswertungen zu DIF für den NEPS-K9-Mathematiktest sind im technischen Bericht für die NEPS-Tests gefordert (Pohl & Carstensen, 2012) und im technischen Bericht für den NEPS-K9-Mathematiktest dargelegt worden (Duchhardt & Gerdes, 2013). Für die Auswertung der Hypothese H1 wurde daher ein Review der technischen Berichte vorgenommen. Zusätzlich wurden für die Subgruppen, die einen DIF-Effekt vermuten lassen, grafische Modelltests für eine Einschätzung der Abweichungen von der Personenhomogenität durchgeführt. Dafür wurden die Itemparameter für die Subgruppen getrennt geschätzt und in einem Streudiagramm dargestellt. Bei perfekter Personenhomogenität müssten die Parameter auf der Winkelhalbierenden liegen (Rost, 1996).

Lokale stochastische Unabhängigkeit (LSU)

Eine weitere Annahme des verwendeten IRT-Modells ist die lokale stochastische Unabhängigkeit (LSU) der Aufgaben. Für die Prüfung der LSU wurde der PRT-Index von Huyhn und Ferrara (1995) berechnet. Dieser beruht darauf, dass Zusammenhänge zwischen den manifesten Variablen (den Aufgaben) allein durch die Ausprägung des latenten Merkmals (der Fähigkeit) bedingt sind. Eine Korrelation zwischen den Aufgaben wird bei lokaler stochastischer Unabhängigkeit der Aufgaben nur durch Unterschiede in der Ausprägung des latenten Merkmals verursacht. Bei einer Herauspartialisierung des Einflusses der Fähigkeit θ aus der Korrelation zwischen den manifesten Variablen kann kein Zusammenhang mehr zwischen diesen bestehen. Für die Prüfung der LSU der Aufgaben wurden daher die partiellen Inter-Item-Korrelationen über alle Itempaare mit dem Datensatz der NEPS-Hauptuntersuchung berechnet. Der Mittelwert dieser Korrelationen wird als PRT-Index verwendet (Huyhn & Ferrara, 1995).

Trennschärfen und Abweichung der beobachteten Antwortwahrscheinlichkeit

Im technischen Bericht für die NEPS-Tests (Pohl & Carstensen, 2012) ist auch die Evaluation der Trennschärfen im Sinne von Steigungsparametern und der Abweichung der beobachteten Antwortwahrscheinlichkeit von der mit dem Modell vorhergesagten Wahrscheinlichkeit gefordert worden. Diese Auswertungen sind folglich im technischen Bericht für den NEPS-K9-Mathematiktest dargelegt worden (Duchhardt & Gerdes, 2013). Für die Evaluation dieser zwei Hypothesen wurde ein Review der genannten Dokumente durchgeführt.

4.3.2 Ergebnisse

Ergebnisse zu H1: spezifische Objektivität

Im Folgenden sollen die Ergebnisse zur Hypothese„H1: Für den Test kann spezifische Objektivität festgestellt werden.“ zusammengetragen werden. Für die Testung der spezifischen Objektivität werden die Aufgaben auf DIF getestet. Im technischen Bericht

für die NEPS-Tests wurde vorgeschrieben, wie auf DIF kontrolliert werden soll (Pohl & Carstensen, 2012). DIF solle als der Unterschied zwischen den geschätzten Schwierigkeiten zweier Gruppen unter Kontrolle der mittleren Gruppenunterschiede operationalisiert werden (Lord, 1980). Dabei sollen Haupteffekte der Gruppenvariable sowie Interaktionseffekte von Item und Gruppe für die NEPS-Tests modelliert werden. Zusätzlich gaben die Autoren vor, das Modell, welches Item- und Gruppeneffekte zulässt, mit einem Modell zu vergleichen, welches nur Haupteffekte der Gruppenvariable erlaubt. Die Modellgütekriterien AIC und BIC sollen dabei für die Evaluation des Modellfits herangezogen werden. Dabei berücksichtigt der BIC-Wert die Anzahl der geschätzten Parameter und verhindert so eine Überparametrisierung (Pohl & Carstensen, 2012). Sowohl für den AIC-Wert als auch für den BIC-Wert gilt, dass ein kleinerer Index auf eine bessere Modellpassung hinweist (Rost, 1996).

Für die NEPS-Tests wurde DIF für die Gruppen Geschlecht, Migrationshintergrund, Schulform und die Anzahl der Bücher zu Hause (als Proxy für sozioökonomischen Status) sowie Personen mit fehlenden Antworten berechnet. Die Schülerinnen und Schüler haben dann einen Migrationshintergrund, wenn entweder sie oder mindestens eines ihrer Elternteile in einem anderen Land geboren wurden. Für die Anzahl der Bücher zu Hause wurde eine dichotome Variable mit den Kategorien unter 100 Bücher und über 100 Bücher erstellt, da diese Unterteilung zu annehmbaren Gruppengrößen führt. Des Weiteren wurde die Variable Schulform in die Kategorien Besuch eines Gymnasiums und Besuch einer anderen Schulform unterteilt. Diese Unterteilung wurde gewählt, da das Konzept des Gymnasiums in allen Bundesländern ähnlich ist, wohingegen sich die übrigen Schulformen zwischen den Bundesländern unterscheiden (Pohl & Carstensen, 2012). Im technischen Bericht für den NEPS-K9-Mathematiktest wurden die Ergebnisse bezüglich des DIF für die Variablen Geschlecht, Schulform, Migrationshintergrund und Anzahl Bücher zu Hause dargelegt. Für die Variable Geschlecht wurde nur eine Aufgabe mit einem nennenswerten aber nicht bedeutsamen DIF gefunden (DIF = 0.48 logits). Die übrigen Items zeigten keinen DIF. Zwischen Schülerinnen und Schülern mit wenigen und vielen Büchern Zuhause wurde kein DIF gefunden (maximaler DIF = 0.30 logits). Bei einem Vergleich der Gruppe von Personen, die keine gültige Antwort auf die Bücherfrage gaben, mit den anderen beiden Gruppen wurde für fünf Aufgaben ein nennenswerter aber nicht bedeutsamer DIF gefunden (maximaler DIF = 0.518 logits). Für den Vergleich von Schülerinnen und Schülern, die ein Gymnasium besuchen, mit denen anderer Schulformen wurden drei Aufgaben mit einem nennenswerten aber nicht bedeutsamen

DIF gefunden (maximaler DIF = 0.49). Die übrigen Aufgaben zeigten für die Variable Schulform keine Auffälligkeiten. Für die Variable Migrationshintergrund wurde kein nennenswerter DIF gefunden (maximaler DIF = 0.37 logits). Zusätzlich wurden Modellgütekriterien von den Modellen, die DIF für die Variablen Geschlecht, Schulform, Migrationshintergrund und Anzahl Bücher zu Hause berücksichtigen, mit den Modellgütekriterien der Modelle, die nur einen Haupteffekt zulassen, verglichen. Für alle vier Variablen wurde jeweils ein kleinerer AIC-Wert für die DIF-Modelle gefunden. Kleinere BIC-Werte, wurden für die DIF-Modelle mit jeweils den Variablen Geschlecht und Schultyp sowie für die Haupteffekt-Modelle mit jeweils den Variablen Anzahl Bücher und Migrationshintergrund gefunden.
Die Abbildungen 25 und 26 visualisieren die Personenhomogenität für die beiden Subgruppen. Für die Jungen wurden die Itemparameter bei einer getrennten Skalierung etwas leichter geschätzt als für die Mädchen. Ein stärkerer Effekt wurde für die Subgruppe Schulform nachgewiesen. Für Gymnasiastinnen und Gymnasiasten wurden die Itemparameter bei einer getrennten Skalierung deutlich einfacher geschätzt als für Schülerinnen und Schüler, die andere Schulformen besuchten. Es kann also nicht ausgeschlossen werden, dass für die Subgruppen des Tests, das heißt Jungen und Mädchen sowie Schülerinnen und Schüler eines Gymnasiums und Schülerinnen und Schüler anderer Schulformen, unterschiedliche Fähigkeiten zur Bearbeitung des Tests benötigt werden. Für die Variable Geschlecht ist dieser Effekt als sehr gering, für die Variable Schulform als etwas größer einzuschätzen. Ein DIF-Effekt für die Variablen Anzahl der Bücher zu Hause und Migrationshintergrund ist jedoch unwahrscheinlich.

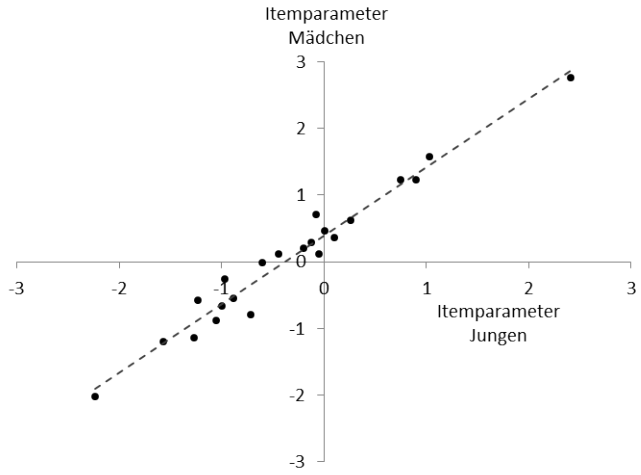


Abbildung 25: Vergleich der Itemparameter bei einer getrennten Skalierung des Tests für Jungen und Mädchen

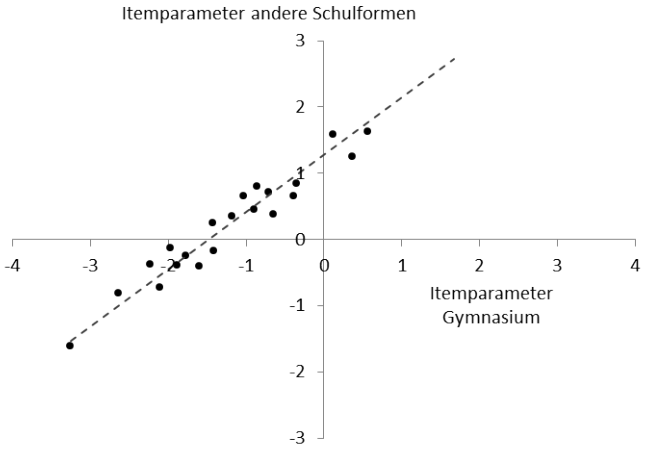


Abbildung 26: Vergleich der Itemparameter bei einer getrennten Skalierung des Tests für Schülerinnen und Schüler des Gymnasiums und anderen Schulformen

Ergebnisse zu H2: LSU der Aufgaben

Für die Hypothese 2 „H2: Die Aufgaben des Tests sind lokal stochastisch unabhängig.“ werden die partiellen Inter-Item-Korrelationen für den NEPS-K9-Mathematiktest in der Abbildung 27 dargestellt. Die niedrigste partielle Korrelation liegt bei r = -.116 und die höchste bei r = .139. Der PRT-Index beträgt r = .040. Es handelt sich also um nur sehr schwache Zusammenhänge zwischen den Residuen der Aufgaben. Daher kann davon ausgegangen werden, dass die Zusammenhänge der Aufgaben hauptsächlich durch die Ausprägung des latenten Merkmals bedingt werden.

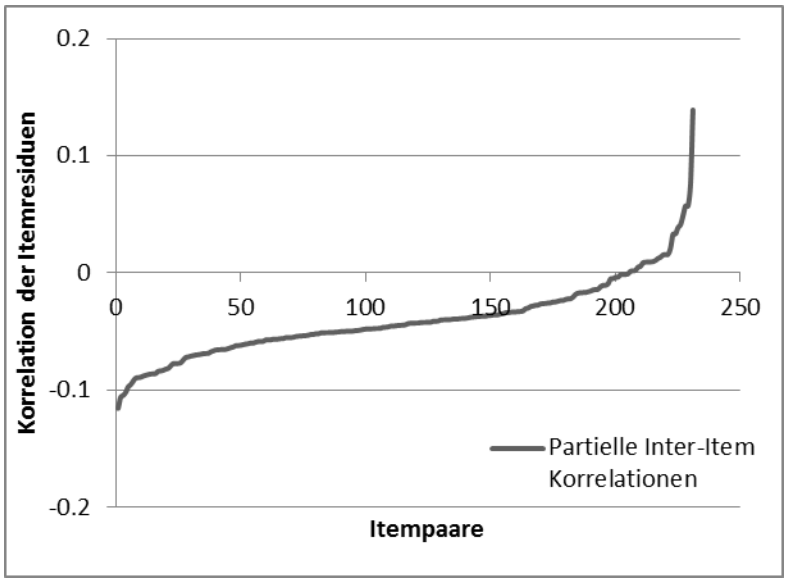


Abbildung 27: Partielle Inter-Item-Korrelationen der NEPS-K9-Mathematikaufgaben

Ergebnisse zu H3: Trennschärfen

Entsprechend der Hypothese 3 „H3: Die Aufgaben haben die gleichen Trennschärfen.“ wurde die Rasch-Homogenität für den NEPS-Mathematiktest im technischen Bericht für den NEPS-K9-Mathematiktest analysiert, indem die Trennschärfe (als Steigungsparameter) für jedes Item in einem 2PL-Modell geschätzt wurde (Duchhardt & Gerdes, 2013). Von Rasch-Homogenität könne ausgegangen werden, wenn die Trennschärfen aus dem 2PL-Modell für alle Items gleich sind (Pohl & Carstensen, 2012). Anschließend wurde die Passung des 1PL-Modells beurteilt, indem die Modellgütekriterien AIC und

BIC dieses Modells mit denen eines 2PL-Modells verglichen wurden (Pohl & Carstensen, 2012). Bei der Skalierung des 2PL-Modells wurden Trennschärfen zwischen 0.58 und 1.54 gefunden. Im Vergleich der Modellinformationskriterien AIC und BIC für das 1PL-Modell und für das 2PL-Modell wurden niedrigere Werte für das 2PL-Modell gefunden (1PL: AIC=379817; BIC=380120; 2PL: AIC=378100 BIC= 378547) (Duchhardt & Gerdes, 2013). Die Berechnungen zeigten, dass die Steigungsparameter für NEPS-Mathematikaufgaben nicht für alle Items gleich sind. Die Hypothese konnte somit nicht bestätigt werden.

Ergebnisse zu H4: Abweichung der beobachteten Antwortwahrscheinlichkeit von der mit dem Modell vorhergesagten Antwortwahrscheinlichkeit

Die Evaluation der Hypothese 4 „Die beobachtete Antwortwahrscheinlichkeit weicht nicht signifikant von der mit dem Modell vorhergesagten Antwortwahrscheinlichkeit ab“ wurde ebenfalls im technischen Bericht der NEPS-Tests gefordert (Pohl & Carstensen, 2012). Der Mean Square (MNSQ) und der Weighted Mean Squares (WMNSQ) sind auf Residuen basierte Fit-Statistiken und beschreiben die Abweichung der beobachteten Wahrscheinlichkeit einer korrekten Antwort bei einer bestimmten Fähigkeit zur vom Modell vorhergesagten Wahrscheinlichkeit. Ein WMNSQ über beziehungsweise unter dem Wert 1 gibt einen Hinweis darauf, ob die Aufgabe weniger trennscharf beziehungsweise trennschärfer ist als erwartet (Wu, 2013). Zusätzlich soll dieser Wert in standard-normalverteilte Maße transformiert werden, wobei die Varianz und der Mittelwert berücksichtigt werden. Dieser transformierte Wert wird in Conquest mit t bezeichnet (Wu et al., 2007; Pohl & Carstensen, 2012). Die Autoren gaben außerdem vor, die modellimplizierten Iteminformationsfunktionen zusätzlich visuell mit empirisch geschätzten Iteminformationsfunktionen zu vergleichen. Dafür sollen laut Autoren die theoretischen ICC, die nach der Schätzung der Itemparameter mit der Wahrscheinlichkeitsfunktion des Rasch-Modells geschätzt werden, der auf den Daten basierten ICC gegenübergestellt werden (Pohl & Carstensen, 2012). Ist die auf den Daten basierte ICC flacher als die geschätzte ICC, so ist die Aufgabe weniger trennscharf als theoretisch erwartet (Wu, 2013). Der Modellfit des 1PL-Modells wurde im technischen Bericht des NEPS-K9-Mathematiktests untersucht (Duchhardt & Gerdes, 2013). Die gefundenen WMNSQ-Werte lagen zwischen 0.91 und 1.1 und wurden als zufriedenstellend bewertet. Dennoch zeigten drei Aufgaben einen t-Wert über |10| (-11.7, 10.4, 13.5) (Duchhardt &

Gerdes, 2013). Bei der Betrachtung der ICCs zeigte sich, dass die Aufgaben mit den beiden höchsten t-Werten jeweils eine akzeptable aber etwas flache Form haben und die Aufgabe mit dem niedrigsten t-Wert eine akzeptable aber etwas steile Form aufweist. Die übrigen ICCs zeigten eine gute Passung der Aufgaben (Duchhardt & Gerdes, 2013).

4.3.3 Diskussion und Fazit *Skalierung*

Für die Überprüfung der spezifischen Objektivität wurden die Aufgaben auf DIF getestet. Keine der Aufgaben wies einen bedeutsamen DIF auf. Zusätzlich wurden Modellgütekriterien von den Modellen, welche jeweils die DIF-Variablen berücksichtigen, mit den Modellgütekriterien der Modelle, die nur einen Haupteffekt zulassen, verglichen. Für die Variablen Geschlecht und Schulform wurden bessere BIC-Werte für die jeweiligen DIF-Modelle gefunden und die grafischen Modelltests wiesen auf leichte Personenheterogenität für die Subgruppe Geschlecht und etwas stärkere Personenheterogenität für die Subgruppe Schulform.
Die partiellen Inter-Item-Korrelationen für die Auswertung der LSU zeigten nur sehr schwache Zusammenhänge zwischen den Residuen der Aufgaben. Es kann daher davon ausgegangen werden, dass die Zusammenhänge der Aufgaben hauptsächlich durch die Ausprägung des latenten Merkmals bedingt werden.
Die Prüfung der Trennschärfen im Sinne von Steigungsparametern zeigte, dass sich die Items nicht ausschließlich durch ihre Schwierigkeit unterscheiden. Bei der Skalierung des 2PL-Modells wurden Trennschärfen zwischen 0.58 und 1.54 gefunden. Niedrigere Werte der Modellinformationskriterien AIC und BIC wurden für das 2PL-Modell aufgezeigt (Duchhardt & Gerdes, 2013). Somit unterscheiden sich die Trennschärfe der Aufgaben für Personen mit unterschiedlichen Fähigkeiten. Die Hypothese konnte nicht bestätigt werden.
Bezüglich der Abweichung der beobachteten Antwortwahrscheinlichkeit von der mit dem Modell vorhergesagten Wahrscheinlichkeit wurden zufriedenstellende WMNSQ-Werte gefunden. Alle ICCs zeigten mindesten akzeptable Passungen.

Für die Variablen Schulform und Geschlecht wurde jeweils das Modell von den Modellgütekriterien favorisiert, welches einen DIF-Effekt für die Variablen zulässt, was auf eine Verletzung der Annahme der Eindimensionalität und der spezifischen Objektivität für diese Gruppen hinweist. Jedoch zeigten die grafischen Modelltests, dass die Personenheterogenität zumindest für die Subgruppe Geschlecht nur sehr gering ist. Auch zeigten die einzelnen NEPS-Aufgaben keine bedeutsamen DIF-Werte für diese Subgruppen. Es konnte nicht ausgeschlossen werden, dass der Test neben der mathematischen Fähigkeit auch Variablen wie Vertrautheit mit den Aufgabenformaten, welche sich eventuell zwischen den Schulformen unterscheidet, sowie Variablen wie Interesse oder Bekanntheit mit den Themen, welche sich eventuell für die Variable Geschlecht unterscheiden, erfasst.

Jedoch ist dieser Effekt nicht als sehr groß einzuschätzen.

Die LSU des NEPS-Mathematiktests wurde in dieser Studie auf andere Weise kontrolliert als im technischen Bericht des NEPS beschrieben. Bei der im technischen Bericht beschriebenen Vorgehensweisen wird die LSU vor allem im Hinblick auf Item-Bundles berechnet. Im NEPS-Mathematiktest kommt jedoch nur ein Item-Bundle vor und es wurde von keinem großen Einfluss auf die LSU ausgegangen. Im technischen Bericht für den NEPS-K9-Mathematiktest wurde daher auf eine solche Berechnung verzichtet (Duchhardt & Gerdes, 2013). Aus diesem Grund wurde für diese Studie der PRT-Index gewählt.

Die bessere Passung des 2Pl-Modells lässt sich unter anderem auch dadurch erklären, dass diese durch die freien Schätzungen der Trennschärfeparameter eine größere Flexibilität aufweist. Durch die Skalierung mit dem 1PL-Modell wurden die Aufgaben jedoch so behandelt, als ob diese sich lediglich durch ihre Schwierigkeit unterschieden. Unterschiede in der Trennschärfe wurden nicht zugelassen. Dies könnte zur Folge haben, dass die Itemparameter nicht präzise geschätzt wurden.

Insgesamt konnten Hinweise für die Schlussfolgerung *Skalierung* gefunden werden. Die Ergebnisse des technischen Berichts wiesen auf die spezifische Objektivität des Tests für die meisten Subgruppen hin sowie darauf, dass die beobachtete Antwortwahrscheinlichkeit nicht signifikant von der Modellvorhersage abweicht. Die Analysen zeigten außerdem die lokale stochastische Unabhängigkeit des Tests auf. Jedoch konnte auch eine Einschränkung der Validität der Schlussfolgerung *Skalierung* aufgedeckt werden. Die Ergebnisse des technischen Berichts zeigten, dass sich die Trennschärfen für die Aufgaben unterscheiden und konnten die Eindimensionalität und spezifische Objektivität für die Subgruppen Geschlecht und Schulform nicht bestätigen.

Die Schlussfolgerung der *Skalierung*, dass das beobachtete Testergebnis zu Fähigkeitsschätzern führt, welche die Zieldomäne Mathematische Kompetenz widerspiegeln, konnte nicht vollständig unterstützt werden. Bei der Interpretation der Testergebnisse müssen eventuelle Unterschiede in der Parameterschätzung für die Variablen Geschlecht und Schulform sowie eventuelle Ungenauigkeiten bei der Schätzung der Itemparameter durch die Fixierung der Trennschärfen bei der 1PL-Skalierung berücksichtigt werden.

4.4 Generalisierung

Die Schlussfolgerung *Generalisierung* wurde in Kapitel 2.4 entwickelt. Eine Übersicht des Argumentes, der Annahmen und Hypothesen dieser Schlussfolgerung wurde in Kapitel 2.4.4, Abbildung 19 gegeben. In diesem Abschnitt wird die Methode zur Überprüfung der Hypothesen beschrieben. Außerdem werden die Ergebnisse zur Standardisierung der Durchführungsbedingungen dargelegt. Dazu wird die Auswahl und Schulung der Testleiterinnen und Testleiter, eine Begutachtung des Testmanuals und eine Einschätzung der Einhaltung der Vorgaben durch die Testleiterinnen und Testleiter dargelegt. Des Weiteren wird eine Analyse der Messgenauigkeit vorgestellt, welche eine Analyse der Testinformationsfunktion und des Zusammenhangs der Personenfähigkeiten mit den Standardfehlern sowie ein Maß zur Reliabilität beinhaltet. Das Unterkapitel Generalisierung endet mit einer Zusammenfassung und Diskussion der Ergebnisse.

4.4.1 Methode

Das Argument der Schlussfolgerung *Generalisierung* „Die Fähigkeiten auf der latenten Skala sind angemessene Schätzer für erwartete Ergebnisse in parallelen Messungen.“ basiert auf zwei Annahmen und zwei Hypothesen (vgl. Kapitel 2.4). Nachfolgend wird die Methode zur Evaluation der Hypothesen beschrieben.

Durchführungsbedingungen

Für die Auswertung der Hypothese H1: „Die Testung wurde sorgfältig anhand von angemessenen, standardisierten Prozeduren durchgeführt.“ wurde ein Review des methodischen Berichts durchgeführt. Die Maßnahmen bezüglich einer Standardisierung der Durchführungsbedingungen (Auswahl und Schulung der Testleiterinnen und Testleiter sowie Erstellung eines Testmanuals) für den NEPS-K9-Mathematiktest sind vor der Durchführung der Studie beschlossen worden und der Prozess der Testleitergewinnung und Schulung ist im Methodenbericht dokumentiert worden (IEA Data Processing and Research Center, 2013). Diese Maßnahmen wurden anhand von Kriterien aus der Literatur geprüft.

Messgenauigkeit

Für die Evaluation der Reliabilität der individuellen Fähigkeitsschätzer wurden verschiedene Berechnungen durchgeführt. Jedes Item hat eine Informationsfunktion $I_i(\Theta)$. Diese Funktion dient als Index, wie viel Information das Item zur Fähigkeitsmessung beiträgt und wie groß die Unsicherheit bezüglich der Vorhersage der Itemantwort ist. Die Iteminformationsfunktion für dichotome Items wird im Folgenden dargestellt:

$$I_i(\Theta) = \frac{[P_i'(\theta)]^2}{P_i(\theta)(1 - P_1(\theta)).}$$

Wobei $P_i(\theta)$ die Wahrscheinlichkeit darstellt, das Item zu lösen. Für polytome Items kann die Information der Antwort in einer Kategorie (k) eines Items (i) als Teil der Iteminformation basierend auf der Kategorie definiert werden (Bock, 1972; Muraki, 1993):

$$I_{ik}(\Theta) = P_{ik}(\theta)I_i(\theta).$$

Diese Gleichung kann auch Itemkategorieinformationsfunktion (IKIF) genannt werden. Die Iteminformationsfunktion für polytome Items kann dementsprechend als Summe der IKIF beschrieben werden:

$$I_i(\Theta) = \sum_{k=1}^{m_i} I_{ik}(\theta).$$

Alle Iteminformationsfunktionen summieren sich zur Testinformationsfunktion auf, welche wiederum den Beitrag des gesamten Tests zur Fähigkeitsmessung zeigt:

$$I(\Theta) = \sum_{i=1}^{n} I_i(\Theta).$$

Der Standardfehler der Schätzung oder der Messung spiegelt die Varianz der latenten Fähigkeitsschätzung wieder und ist eine Umkehrung der Testinformationsfunktion (Invertierung der Quadratwurzel der Testinformation):

$$SE(\hat{\Theta}) = \frac{1}{\sqrt{I(\Theta)}}.$$

$I(\Theta)$ ist die Testinformation und $SE(\hat{\Theta})$ ist der Standardfehler der Schätzung für das Testinstrument. Je höher die Testinformation ist, desto niedriger ist der Standardfehler der Schätzung und desto weniger fehlerbehaftet ist die Fähigkeitsmessung (Fan & Sun, 2013).

Die Testinformationsfunktion und die Standardfehler der Personenschätzer wurden mit der Software Conquest berechnet. Der Zusammenhang der Personenfähigkeit mit der Messgenauigkeit wurde anschließend visuell dargestellt.
Ein Gesamtmaß für die Genauigkeit der Personenschätzer wurde von der Software Conquest mit der EAP/PV-Reliabilität angegeben. Dieser Wert ist im Gegensatz zu den Standardfehlern der Personenparameterschätzungen populationsabhängig und gibt an, wie messgenau der Test bezüglich der Population ist. Die EAP/PV-Reliabilität wurde berechnet, indem die erklärte Varianz gemäß des geschätzten Modells durch die totale Personenvarianz geteilt wurde (Wu, 2013). Diese Berechnungen sind bereits im technischen Bericht von Duchhardt und Gerdes (2013) durchgeführt worden.

4.4.2 Ergebnisse

Ergebnisse zur Testdurchführung

Im Folgenden werden die Informationen zu den Durchführungsbedingungen des NEPS-K9-Mathematiktests aus dem methodischen Bericht (IEA Data Processing and Research Center, 2013) anhand von Kriterien zur standardisierten Testdurchführung der Standards for Educational and Psychological Testing 2014 bewertet. Die American Educational Research Association, American Psychological Association, National Council on Measurement in Education (2014) gaben in den Standards for Educational and Psychological Testing vor, welche Kriterien bei der Testdurchführung eingehalten werden sollten, um unter anderem konstruktirrelevante Varianz in den Testscores zu vermeiden. So müssen (1) die Testleiterinnen und Testleiter eine angemessene Schulung erhalten, eine Übersicht über den Testprozess gewinnen und eine Dokumentation des Prozesses

erhalten, sodass sie fähig sind, die Testung angemessen durchzuführen und die Wichtigkeit der Einhaltung der Vorgaben zu verstehen. (2) Testentwicklerinnen und Testentwickler sollten die standardisierten Durchführungsbedingungen spezifizieren, welche für die Interpretation der Testergebnisse von Belang sind. Die Spezifikationen bezüglich der Instruktionen für Testpersonen, Zeitvorgaben, die Art der Aufgabenpräsentation und Antworten sowie Materialien und Ausstattung sollten genau definiert werden. (3) Änderungen oder Störungen der standardisierten Durchführung sollten dokumentiert und berichtet werden, sodass diese bei der Interpretation und Nutzung der Daten berücksichtigt werden können. (4) Auch die Art der räumlichen Testumgebung sollte spezifiziert werden. Die Testumgebung sollte beispielsweise nicht zu laut, zu dunkel oder zu klein sein. (5) Die Instruktionen für die Testpersonen sollten klar verständlich sein und alle wichtigen Informationen zur Bearbeitung der Tests beinhalten. (6) Möglichkeiten zur Täuschung durch Testpersonen sollte vorgebeugt werden.

Insgesamt wurden 370 Testleiterinnen und Testleiter eingesetzt. Diese waren entweder dem Erhebungsinstitut angehörig oder Studierende ausgewählter Studiengänge. Eine im Vorfeld durchgeführte Schulung informierte alle Testleiterinnen und Testleiter über Studie, den Ablauf der Erhebungsvorbereitung sowie Richtlinien zur Erhebungsdurchführung, die Erhebungsmaterialien, die Einverständniserklärungen, die Schülerlistenführung, den Ablauf des Testtages und die besonderen Aufgaben der Testleiterinnen und Testleiter. Außerdem wurde eine praktische Übungsphase durchgeführt (IEA Data Processing and Research Center, 2013). Die Durchführungsbedingungen wurden klar spezifiziert. Die Erhebungssitzungen fanden im Klassenverband statt. Das Material bestand aus dem Test in Papierform und einem Stift. Zusätzlich durften die Schülerinnen und Schüler einen Taschenrechner benutzen. Die Aufgaben wurden schriftlich präsentiert. Die Bearbeitungszeit für den NEPS-K9-Mathematiktest lag bei 28 Minuten. Abweichungen von den Durchführungsvorgaben und Probleme bei der Testung wurden durch die Testleiterinnen und Testleiter in einem Testsitzungsprotokoll festgehalten. Die Bearbeitungszeiten laut Testsitzungsprotokoll wurden im Methodenbericht (IEA Data Processing and Research Center, 2013) veröffentlicht. In ca. 720 der 985 Testsitzungen wurde die Bearbeitungszeit von 28 Minuten präzise eingehalten. In ca. 90 Testsitzungen wurde die Bearbeitungszeit um eine Minute überschritten, in ca. 20 Testsitzungen um 2 Minuten und in zwei Testsitzungen um fünf Minuten. Damit wurde die Bearbeitungszeit in ca. 12% der Testsitzungen leicht überschritten. In den übrigen Testsitzungen wurde die Bearbeitungszeit von 28 Minuten mit maximal zehn Minuten unterschritten. Die Art

der Testumgebung wurde in den veröffentlichen Dokumenten nicht genau spezifiziert. Es ist lediglich bekannt, dass die Testungen an einem von der Schule gewählten Vormittag und im Klassenverband stattfanden. Die Instruktion der Aufgaben wurde in den bisher veröffentlichen Dokumenten nicht weiter beschrieben. Unter anderem zur Vorbeugung von Täuschungen war in den meisten Testsitzungen (89,4%) neben der Testleiterin oder dem Testleiter auch eine Aufsichtslehrkraft anwesend.

Insgesamt kann von einer standardisierten Durchführung gesprochen werden. Einige Informationen über die Durchführungsbedingungen sind bisher noch nicht veröffentlicht worden.

Ergebnisse zur Reliabilität der Fähigkeitsmessungen

Die Ergebnisse zu Hypothese 2 „Die individuellen Fähigkeitsmessungen sind reliabel." sind sukzessiv in mehreren Abbildungen zusammengefasst. In Abbildung 28 wird die Testinformationsfunktion (TIF) des Tests gezeigt. Die horizontale Achse repräsentiert die latente Fähigkeitsskala in logits und die vertikale Achse die Testinformation. In dem Bereich, in dem die TIF ihre höchsten Werte hat, werden mit dem Test die meisten Informationen erfasst, besteht jedoch auch die meiste Varianz. Die Kurve ist sehr leicht nach links versetzt. Die meisten Informationen wurden im Fähigkeitsbereich um 0 logits erfasst. In einem Bereich kleiner als -3.6 und größer als 3.3 wurden nur noch wenige Informationen mit dem Test generiert.

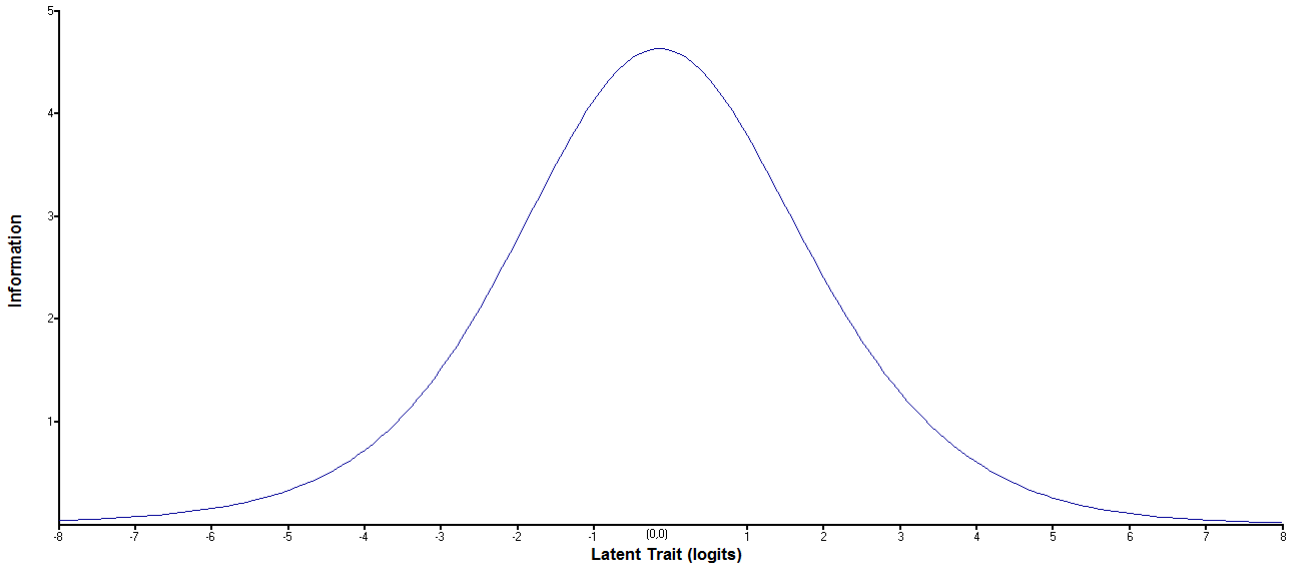


Abbildung 28: Testinformationsfunktion

Abbildung 29 zeigt den Zusammenhang zwischen der Personenfähigkeit und der Messgenauigkeit. In dem Bereich, in dem die Standardfehler am niedrigsten ausfallen, sind die Personenfähigkeitsschätzer am präzisesten. Im Durchschnitt beträgt der Standardfehler $SE = 0.54$. Der kleinste Standardfehler liegt bei $SE = 0.46$ und der höchste Standardfehler bei $SE = 2.31$. Für die meisten Schülerinnen und Schüler (91%) wurde ein WLE zwischen ($-2 \leq \theta \leq +2$) gefunden. In diesem Bereich liegt der Standardfehler im Durchschnitt bei $SE = 0.5$. In den Randbereichen fällt der Standardfehler deutlich höher aus. In diesen Bereich fallen allerdings nur wenige Testpersonen. Insgesamt wurde für 12% der Schülerinnen und Schüler ein Fähigkeitsschätzer mit eine Messfehler über $SE = 0.6$ gefunden. Aufgrund der begrenzten Messungen pro Person (22 Aufgaben) ist dieser Wert als angemessen einzuschätzen. Für die Interpretation der Testwerte ist allerdings zu beachten, dass bei einem $SE = 0.6$ das 95%-Konfidenzintervall einen Bereich von 1.2 logits um den Fähigkeitsschätzer umfasst.

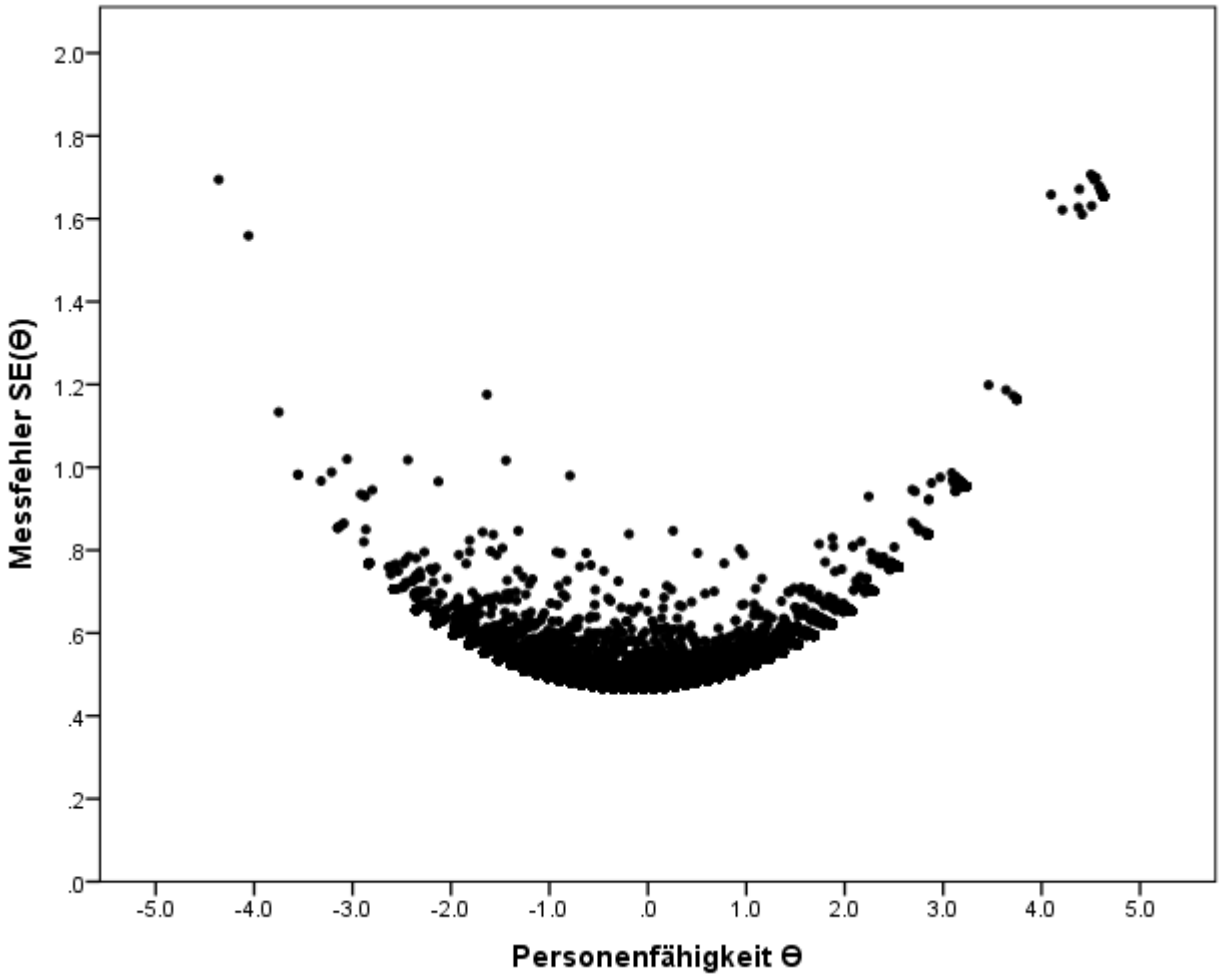


Abbildung 29: Zusammenhang der Personenfähigkeit und der Messgenauigkeit

Insgesamt konnte gezeigt werden, dass der Test in einem mittleren Fähigkeitsbereich (-2 $\leq \theta \leq +2$) eine annehmbare Messgenauigkeit hat, in den Randbereichen jedoch höhere Standardfehler aufweist.

Im technischen Bericht für den NEPS-K9-Mathematiktest wurde die EAP/PV-Reliabilität angegeben, welche bei der Skalierung des NEPS-Mathematiktests mit dem Programm Conquest berechnet worden war. Es wurde eine EAP/PV-Reliabilität von *EAP/PV reliability* =.811 gezeigt. Dieser Wert kann wie Cronbachs Alpha interpretiert werden und deutet auf exakte Schätzer der Personenfähigkeit hin (Duchhardt & Gerdes, 2013; Pohl & Carstensen, 2012).

4.4.3 Diskussion und Fazit *Generalisierung*

Zusammenfassend kann festgehalten werden, dass die Durchführungsbedingungen den Anforderungen der Standards for Educational and Psychological Testing weitgehend entsprachen. Einige Informationen zur Testdurchführung bezüglich der Instruktion und der Testumgebung sowie genauere Angaben zur Störung von Testsitzungen fehlten jedoch. Die vorgegebene Testzeit wurde von fast allen Testleiterinnen und Testleitern eingehalten. Insgesamt sprechen die durchgeführten Standardisierungen und die gefundene Reliabilität dafür, dass auch die Instruktion und die Testumgebungen angemessen standardisiert wurden. Außerdem zeigten die Ergebnisse, dass der Test im mittleren Fähigkeitsbereich (-2 $\leq \theta \leq +2$) eine angemessene Messgenauigkeit aufweist und in den Randbereichen höhere Messfehler hat. Im sehr hohen und sehr niedrigen mathematischen Fähigkeitsbereich sind die Schätzungen jedoch weniger präzise. Die EAP/PV-Reliabilität ist zufriedenstellend.

Für die Interpretation der Testergebnisse als generalisierte Fähigkeitsschätzer ist eine Untersuchung der fehlenden Informationen von Belang. So sollten die Standardisierung der Testumgebung und die Verständlichkeit und Standardisierungen der Instruktionen untersucht werden. Außerdem wäre es wünschenswert, wenn im Rahmen eines Qualitätsmonitorings zusätzlich stichprobenartig die Einhaltung der Durchführungsbedingungen überprüft worden wäre, wobei geschulte Beobachterinnen und Beobachter unangekündigt den Testablauf überwacht und dokumentiert hätten. Es ist nicht bekannt, ob eine solche Qualitätsprüfung für die Durchführung des NEPS-K9-Tests stattgefunden hat, jedoch

würde eine solche Untersuchung stichhaltige Evidenz für die Einhaltung der Durchführungsbedingungen liefern. Auch ein Review des nicht veröffentlichten Testmanuals auf die Eindeutigkeit und Vollständigkeit der Instruktionen würde wichtige Evidenz für die Validität der Schlussfolgerung liefern. Des Weiteren ist die Messgenauigkeit des Tests in Anbetracht der Anzahl eingesetzter Aufgaben zwar zufriedenstellend, jedoch beinhaltet das 95%-Konfidenzintervall minimal 1.8 logits und im Durchschnitt über alle Schülerinnen und Schüler 2.1 logits. Diese Streuung muss bei der Interpretation der Testwerte berücksichtigt werden.

Insgesamt konnten Hinweise für die Validität der Schlussfolgerung *Generalisierung* gefunden werden. Die Durchführungsbedingungen waren standardisiert und entsprachen den Standards for Educational and Psychological Testing. Die Analysen der zweiten Hypothese ergaben, dass die Reliabilität des Tests als zufriedenstellend eingestuft werden kann. So weisen die Testinformationsfunktion und die Standardfehler auf eine angemessene Messgenauigkeit des Tests für die meisten Schülerinnen und Schüler hin. Trotz der zufriedenstellenden Werte wurden aufgrund der Testlänge relativ große Konfidenzintervalle gefunden, die die mögliche Testwertinterpretation einschränken. Als Fazit kann geschlossen werden, dass Fähigkeiten auf der latenten Skala im mittleren Fähigkeitsbereich angemessene Schätzer für erwartete Ergebnisse über parallele Messungen sind. Eine gewisse Unsicherheit der Fähigkeitsschätzer muss jedoch aufgrund der fehlenden Informationen zur Standardisierung und der relativ großen Konfidenzintervalle berücksichtigt werden.

4.5 Konstruktbezug

Für die Schlussfolgerung *Konstruktbezug* wurde die Argumentationsstruktur in Kapitel 2.5 entwickelt und beschrieben. Eine Darstellung des Argumentes, der Annahmen und der Hypothesen wird in Kapitel 2.5.2 in Abbildung 20 gezeigt. Nachfolgend wird die Methode für die Evaluation dieser Hypothesen beschrieben. Im Ergebnisteil wird die Analyse der dimensionalen Struktur des NEPS-Mathematiktests beschrieben. Außerdem wird in diesem Abschnitt dargelegt, inwiefern sich der NEPS-Mathematiktest dimensional von anderen in NEPS gemessenen Kompetenzen abgrenzen lässt. Der Abschnitt schließt mit einer Zusammenfassung und Diskussion der gefundenen Ergebnisse ab.

4.5.1 Methode

Das Argument der Schlussfolgerung *Konstruktbezug* „Vom (generalisierten) Testergebnis des NEPS-Mathematiktests für die 9. Klassenstufe lässt sich auf das Konstrukt der mathematischen Kompetenz, wie es im NEPS definiert wird, schließen.“ basiert auf einer Annahme und vier Hypothesen (vgl. Kapitel 2.5). Nachfolgend wird die Methode zur Evaluation der Hypothesen beschrieben.

Dimensionale Struktur des NEPS-K9-Mathematiktests

Im Rahmen der dimensionalen Analysen wurden die theoretischen Annahmen über die Dimensionalität des NEPS-Mathematiktests empirisch überprüft. Für die mathematische Kompetenz im NEPS wird zwar Eindimensionalität angenommen, dennoch wurden im Rahmenkonzept Teildimensionen definiert. Pohl und Carstensen (2012) gaben im technischen Bericht zur Skalierung der Kompetenzdaten vor, wie die Annahme der Eindimensionalität zu überprüfen sei. Für den NEPS-K9-Mathematiktest sind diese Berechnungen bereits durchgeführt und im technischen Bericht veröffentlicht worden (Duchhardt & Gerdes, 2013). Aus diesem Grund wurde für die Beantwortung der Forschungsfrage ein Review des technischen Berichts für den NEPS-K9-Mathematiktest durchgeführt. Die dritte Hypothese „Für eine mehrdimensionale Skalierung der mathematischen, naturwissenschaftlichen, ICT- und Lesekompetenzen auf separaten Dimensionen im NEPS wird eine bessere Passung gefunden als bei einer eindimensionalen Skalierung der Kompetenzen auf einer gemeinsamen Dimension.“ wurde untersucht, indem die Modellgütekriterien AIC, BIC und CAIC einer eindimensionalen Skalierung der Mathematik- und Naturwissenschaftskompetenzen, der Mathematik- und Lesekompetenzen sowie der Mathematik- und ICT-Kompetenzen mit denen einer zweidimensionalen Skalierung der mathematischen Kompetenz und der naturwissenschaftlichen Kompetenz, der mathematischen Kompetenz und der Lesekompetenz sowie der mathematischen Kompetenz und der ICT-Kompetenz verglichen wurden. Der CAIC-Wert stellt eine Korrektur des AIC-Wertes dar, die auch bei größeren Stichprobenumfängen konsistent bleibt (Rost, 1996). Für diese Berechnungen wurden die Daten der NEPS-Haupterhebung im Jahr 2010 verwendet. Dabei wurden die Daten der Schülerinnen und Schüler für die Berechnung der Modelle berücksichtigt, welche Aufgaben des Mathematiktests und des jeweils anderen Kompetenztests bearbeitet hatten. Für die Evaluation der vierten Hypo-

these wurden in den zweidimensionalen Modellen außerdem die latenten Korrelationen zwischen der Mathematik- und der Naturwissenschaftskompetenz respektive der Kompetenz Leseverständnis berechnet und verglichen. Für die Evaluation dieser Hypothesen wurden die Daten der NEPS-Haupterhebung verwendet.

4.5.2 Ergebnisse

Ergebnisse zu H1: Zusammenhänge innerhalb der Teildimensionen

Duchhardt und Gerdes (2013) prüften entsprechend der Hypothese 1 „Innerhalb der NEPS-Teildimensionen können sehr hohe Zusammenhänge gefunden werden.“ die Korrelationen zwischen den vier Inhaltsbereichen des NEPS. Diese wurden in einem vierdimensionalen Modell berechnet. Dabei wurde jede Aufgabe einem Inhaltsbereich zugeordnet und für jeden Inhaltsbereich wurde eine eigene Dimension modelliert (*between-item-multidimensionality*). Im technischen Bericht für die Auswertung der Kompetenztests im NEPS wurden sehr hohe Korrelationen zwischen den Dimensionen gefordert (Carstensen, 2013), um von einer Eindimensionalität des Tests ausgehen zu können (Pohl & Carstensen, 2012). Die Ergebnisse der Berechnungen werden in Tabelle 14 dargestellt. Die Varianzen für die Dimensionen befinden sich auf der Diagonalen und die Korrelationen zwischen den Inhaltsbereichen befinden sich unterhalb der Diagonalen. Insgesamt wurden hohe Korrelationen zwischen $.906 \leq r \leq .967$ gefunden.

Die gefundenen Zusammenhänge erreichten jedoch nicht die im NEPS geforderte Höhe der Korrelation von $r > .95$ (Pohl & Carstensen, 2012). Aus diesem Grund konnte diese Hypothese nicht bestätigt werden.

Tabelle 14: Ergebnisse der vierdimensionalen Skalierung des NEPS-K9-Mathematiktests aus Duchhardt & Gerdes, 2013

	Dimension 1	Dimension 2	Dimension 3	Dimension 4
Quantität (7 Aufgaben)	1.220			
Raum und Form (6 Aufgaben)	0.925	1.180		
Veränderung und Beziehungen (6 Aufgaben)	0.965	0.942	1.466	
Daten und Zufall (3 Aufgaben)	0.967	0.906	0.946	1.109

Ergebnisse zu H2: Dimensionalität des Tests

Zur Prüfung der Hypothese 2: „Für den NEPS-Mathematiktest ist eine eindimensionale Skalierung einer mehrdimensionalen Skalierung vorzuziehen.“ lässt sich die Arbeit von Duchhardt und Gerdes (2013) anführen. Die Autoren verglichen im technischen Bericht das vierdimensionale Modell, in welchem jeder Inhaltsbereich auf einer eigenen Dimension lädt (siehe Tabelle 14), mit einem eindimensionalen Modell, in welchem alle Mathematikaufgaben auf einer gemeinsamen Dimension laden, anhand der Modellgütekriterien AIC und BIC. Für das vierdimensionale Modell wurden mit einem AIC-Wert von 380 456 und einem BIC-Wert von 380 729 niedrigere Modellgütekriterien gefunden als für das eindimensionale Modell (AIC = 380 619, BIC = 380 823). Die Ergebnisse weisen darauf hin, dass der NEPS-K9-Mathematiktest keine eindimensionale Struktur hat. Die Hypothese H2 konnte daher nicht bestätigt werden.

Ergebnisse zu H3: Kompetenzspezifische Skalierung

Für die Hypothese 3 „Für eine mehrdimensionale Skalierung der mathematischen, naturwissenschaftlichen, Lese- und ICT-Kompetenzen auf separaten Dimensionen im NEPS wird eine bessere Passung gefunden als bei einer eindimensionalen Skalierung der Kompetenzen auf einer gemeinsamen Dimension.“ wurden jeweils drei ein- und zweidimensionale Modellierungen durchgeführt. Im Folgenden werden die Ergebnisse von jeweils einer eindimensionalen Skalierung und einer zweidimensionalen Modellierung des NEPS-Mathematiktests und des NEPS-Naturwissenschaftstests (1), des NEPS-Mathematiktests und des NEPS-Lesetests (2) sowie des NEPS-Mathematiktests und des ICT-Tests (3) vorgestellt. In Tabelle 15 werden die Modellgütekriterien AIC, BIC und CAIC, Devianz und Parameter der ein- und zweidimensionalen Skalierungen der mathematischen und naturwissenschaftlichen NEPS-Tests gezeigt. Für die Skalierung der mathematischen und naturwissenschaftlichen Kompetenzen auf jeweils einer separaten Dimension wurden bessere Modellgütekriterien gefunden als für die Skalierung beider Tests auf einer gemeinsamen Dimension.

Die Ergebnisse der ein- und zweidimensionalen Skalierung des NEPS-Mathematiktests und des NEPS-ICT-Tests werden in Tabelle 16 dargestellt. Für die zweidimensionale Modellierung der Kompetenzen wurden niedrigere AIC-, BIC- und CAIC-Werte gefunden

Tabelle 15: Modellgütekriterien für die ein- und zweidimensionale Skalierung der mathematischen und naturwissenschaftlichen Kompetenzmessungen im NEPS

	Parameter	Devianz	N	AIC	BIC	CAIC
Mathematik-NaWi eindimensional	78	942155	14462	942311	942902	942980
Mathematik-NaWi zweidimensional	81	939198	14462	939360	939974	940055

als für die Modellierung der Tests auf einer gemeinsamen Dimension.

Tabelle 16: Modellgütekriterien für die ein- und zweidimensionale Skalierung des Mathematik- und ICT-Kompetenztests im NEPS

	Parameter	Devianz	N	AIC	BIC	CAIC
Mathematik-ICT eindimensional	82	996860	14471	997024	997645	997727
Mathematik-ICT zweidimensional	85	992713	14471	992883	993527	993612

Tabelle 17 zeigt die Ergebnisse der ein- und zweidimensionalen Skalierung des NEPS-Mathematiktests und des NEPS-Lesetests. Für die Skalierung des Mathematik- und Lesetests auf jeweils einer separaten Dimension wurden zu niedrigeren AIC-, BIC- und CAIC-Werten gefunden als für die Skalierung auf einer gemeinsamen Dimension.

Der NEPS-Mathematiktest lässt sich dimensional von den naturwissenschaftlichen, Lese- und ICT-Kompetenztests des NEPS abgrenzen. Die Hypothese H3 konnte bestätigt werden.

Tabelle 17: Modellgütekriterien für die ein- und zweidimensionale Skalierung des Mathematik- und Lesekompetenztests im NEPS

	Parameter	Devianz	N	AIC	BIC	CAIC
Mathematik-Lesen eindimensional	66	715420	13410	715552	716047	716113
Mathematik-Lesen zweidimensional	69	708148	13410	708286	708804	708873

Ergebnisse zu H4: Zusammenhang von mathematischer Kompetenz mit naturwissenschaftlicher Kompetenz beziehungsweise Lesekompetenz in NEPS

Die Ergebnisse zu Hypothese 4 „Die mathematische Kompetenz im NEPS hängt höher mit der naturwissenschaftlichen Kompetenz zusammen als mit der Lesekompetenz.“ werden im Folgenden dargestellt. Der Zusammenhang zwischen der mathematischen Kompetenzmessung und der naturwissenschaftlichen Kompetenzmessung im NEPS wird in Abbildung 30 gezeigt. Es wurde eine Korrelation von $r = .82$ gefunden. Die latente Korrelation zwischen der mathematischen Kompetenz im NEPS und der Lesekompetenz im NEPS beträgt $r = .69$ (siehe Abbildung 31).

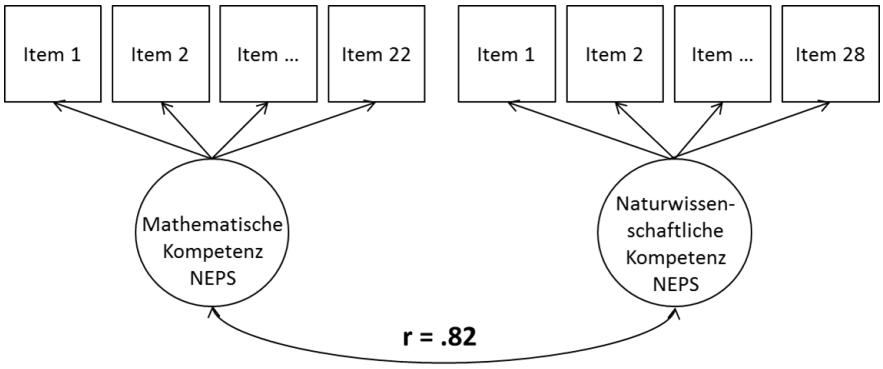


Abbildung 30: Latente Korrelation zwischen der mathematischen und der naturwissenschaftlichen Kompetenz im NEPS

Zusammenfassend kann geschlossen werden, dass die naturwissenschaftliche Kompetenz

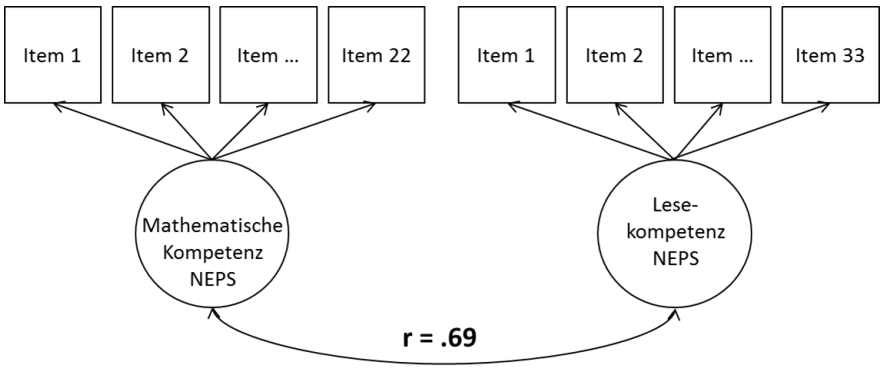


Abbildung 31: Latente Korrelation zwischen der mathematischen Kompetenz und der Lesekompetenz im NEPS

deutlich höher mit der mathematischen Kompetenz zusammenhängt als die Lesekompetenz. Die Hypothese konnte somit bestätigt werden.

4.5.3 Diskussion und Fazit *Konstruktbezug*

Insgesamt konnten Hinweise für die Validität der Schlussfolgerung *Konstruktbezug* gefunden werden. So bestätigten die Ergebnisse, dass sich der NEPS-Mathematiktest von den NEPS-Kompetenztests in den Naturwissenschaften, in Lesen und in ICT abgrenzen lässt. Außerdem wurden die Erwartungen bezüglich des Zusammenhangs zwischen der mathematischen Kompetenz im NEPS sowie der naturwissenschaftlichen Kompetenz und der Lesekompetenz durch die gefundenen Korrelationen belegt.

Jedoch konnten auch Hinweise für eine Einschränkung der Validität der Schlussfolgerung *Konstruktbezug* gefunden werden. So wurden zwar hohe Korrelationen zwischen den Inhaltsbereichen des NEPS-Mathematiktests gefunden, jedoch entsprachen diese nicht der erwarteten Höhe von $r > .95$. Außerdem wurden für eine vierdimensionale Skalierung des NEPS-Tests mit einer eigenen Dimension für jeden Inhaltsbereich bessere Modellgütekriterien gefunden als für eine eindimensionale Skalierung des NEPS-Tests mit einer einzigen Dimension mathematischer Kompetenz (Duchhardt & Gerdes, 2013). Die Hypothesen zur Dimensionalität des NEPS-K9-Mathematiktests konnten damit nicht bestätigt werden.

Obwohl bessere Modellgütekriterien für eine vierdimensionale Modellierung des NEPS-

K9-Mathematiktests gefunden wurden, konnte diese Struktur mit den bei Duchhardt und Gerdes (2013) durchgeführten Berechnungen nicht bewiesen werden. Die Modellgütekriterien AIC, BIC und CAIC können Modelle miteinander vergleichen, jedoch nicht ein 'richtiges' Modell identifizieren. Ein weiterer Nachteil dieser Modellgütekriterien ist, dass es keine klaren Kriterien für die Größe des Unterschiedes zwischen den Werten der Modelle gibt, um von einer besseren Modellpassung eines Modells ausgehen zu können (Rost, 2004).

Des Weiteren wurde bei der Berechnung des vierdimensionalen Modells von einer Between-Item-Dimensionality der Inhaltsbereiche, basierend auf der Rahmenkonzeption des NEPS-Mathematiktests, ausgegangen. Jede Aufgabe kann demnach nur einem Inhaltsbereich zugeordnet werden. Diese Zugehörigkeit der Aufgaben zu den Inhaltsbereichen im NEPS wurde jedoch nicht empirisch bestätigt. Zum einen konnten die Aufgaben durch die Expertinnen und Experten auch in sehr ähnlich definierte Inhaltsbereiche der PISA-Rahmenkonzeption und in leicht unterschiedlich strukturierte Inhaltsbereiche der LV-Rahmenkonzeption eingeordnet werden (vgl. Kapitel 4.1). Überdies wäre auch eine Dimensionalität des Mathematiktests beruhend auf den sechs Prozessen denkbar. Die Einordnung in die Prozesse liegt jedoch nicht vor und ein solches Modell konnte den anderen Modellen nicht gegenübergestellt werden. Diese Überlegungen haben zur Folge, dass für den NEPS-Mathematiktest auch durchaus eine andere Strukturierung und Operationalisierung der Inhaltsbereiche denkbar wäre als die in der NEPS-Rahmenkonzeption vorgesehene. Eine mehrdimensionale Skalierung hätte jedoch zur Folge, dass für die Fähigkeitsschätzer der Personen in den Dimensionen noch weniger Messungen pro Person zur Verfügung ständen und die Standardfehler dementsprechend steigen würden. Somit würde zwar das Konstrukt mathematischer Kompetenz durch die Modellierung besser abgebildet, die Personenfähigkeitsschätzer wären jedoch ungenauer. Des Weiteren suggerieren die hohen Korrelationen zwischen den Inhaltsbereichen, dass für die erfolgreiche Bearbeitung der Aufgaben auch eine gemeinsame Fähigkeit benötigt wird. Es ist zu vermuten, dass die Aufgaben aus dem NEPS-K9-Mathematiktest, neben einer gemeinsamen mathematischen Fähigkeit, spezifische Fähigkeit für die Bearbeitung erfordern. Aufgrund der begrenzten Aufgabenanzahl ließen sich diese spezifischen Fähigkeiten jedoch nicht reliabel abbilden. Für die Interpretation der NEPS-Mathematikdaten ist die Gewichtung der eventuellen spezifischen Fähigkeiten und deren Interaktion nicht unbedingt von Interesse. Das für die Auswertung vorgegebene eindimensionale Modell (Pohl & Carstensen, 2012) repräsentiert vor allem die gemeinsamen mathematischen Fähigkeiten, die zur Lösung der komplexen Aufgaben benötigt werden. Bei einer Interpretati-

on der Testergebnisse als deskriptives Leistungsmaß wäre trotz eventueller Mehrdimensionalität auch ein eindimensionales Modell angemessen (vgl. Hartig & Höhler, 2010). Bei der Testwertinterpretation ist daher darauf hinzuweisen, dass die eindimensionalen Fähigkeitsschätzer die Kompetenzstruktur des NEPS-K9-Mathematiktests zwar nicht vollständig repräsentieren, jedoch für eine Interpretation der Testwerte im Sinne einer Reduktion der Komplexität des Konstruktes angemessen sind.

Weitere Studien zur tatsächlichen Dimensionalität des Tests, beispielsweise durch die Testung eines möglichen g-Faktormodells durch Faktorenanalysen oder die Spezifizierung eines 2PL-Modells mit Within-Item-Dimensionalität (Robitzsch, 2009), könnten Aufschluss über die empirische Struktur des NEPS-Tests geben.

Das Fazit für die Schlussfolgerung *Konstruktbezug* ist, dass vom (generalisierten) Testergebnis des NEPS-Mathematiktests für die 9. Klassenstufe nicht auf das Konstrukt der mathematischen Kompetenz, wie es im NEPS definiert wird, geschlossen werden kann. Die Eindimensionalität des NEPS-Mathematiktests ließ sich in den Daten nicht bestätigen. Aus diesem Grund sind lediglich Interpretationen der Testwerte als Reduktion einer komplexeren Kompetenzstruktur angemessen.

4.6 Extrapolation

Die Argumentationsstruktur der Schlussfolgerung *Extrapolation* wurde in Kapitel 2.6 entwickelt und beschrieben. Eine Darstellung des Argumentes, der Annahmen und der Hypothesen wird in Kapitel 2.6.2 in Abbildung 21 gezeigt. Im anschließenden Abschnitt wird die Methode für die Evaluation dieser Hypothesen für die Schlussfolgerung *Extrapolation* dargelegt. Im Ergebnisteil wird die Prüfung des Zusammenhanges der Testleistung im NEPS-Mathematiktest mit Kriterien mathematischer Kompetenz beschrieben. Dafür wird erst die Analyse des Zusammenhanges der Testleistung mit Noten unterschiedlicher Fächer dargelegt. Anschließend wird der Zusammenhang zwischen der mathematischen Kompetenz im NEPS mit den mathematischen Kompetenzen in PISA und im LV aufgezeigt. In einem letzten Schritt wird die Berechnung des Zusammenhanges der NEPS-Mathematikergebnisse mit den kognitiven Fähigkeiten sowie metakognitiven Fähigkeiten der Schülerinnen und Schüler beschrieben. Der Abschnitt schließt mit einer Zusammenfassung und Diskussion der gefundenen Ergebnisse ab.

4.6.1 Methode

Das Argument der Schlussfolgerung *Extrapolation* „Die Kompetenz der Zieldomäne, wie sie mit dem NEPS-Test erfasst wird, ist ein Indikator für die Leistung in der Zieldomäne Mathematische Kompetenz.“ basiert auf einer Annahme und fünf Hypothesen (vgl. Kapitel 2.6). Nachfolgend wird die Methode zur Evaluation der fünf Hypothesen beschrieben.

Zusammenhang mit Kriterien mathematischer Kompetenz

Für die Untersuchung der ersten Hypothese „Die Leistung der Schülerinnen und Schüler im NEPS-Mathematiktest hängt stärker mit der Mathematiknote zusammen als mit Zeugnisnoten anderer Fächer.“ wurden die Daten der Validierungsstudie aus dem Jahr 2012 verwendet. Die NEPS-Mathematik WLE's der Schülerinnen und Schüler sind im *scientific use file* enthalten. Sie sind bereits mit einem eindimensionalen IRT-Modell geschätzt worden (Duchhardt & Gerdes, 2013). Es wurden die 14 523 Schülerinnen und Schüler, die den NEPS-Mathematiktest bearbeitet haben und für die ein Mathematik-WLE vorgelegen hat, in die folgende Berechnung einbezogen. Die Schulnoten waren ebenfalls Bestandteil des *scientific use files*. Die Mathematik- und die Deutschnote aus dem letzten Jahreszeugnis wurden mit einem Papierfragebogen erfasst. Insgesamt fehlten von den 14 523 Teilnehmern die Angaben von 706 Schülerinnen und Schülern zur Mathematiknote und von 640 Schülerinnen und Schülern zur Deutschnote. Die Schulnoten der Schülerinnen und Schüler wurden am Klassenmittelwert zentriert, um klassenspezifische Unterschiede in der Notengebung zu berücksichtigen (siehe hierzu z.B. Baumert, Trautwein und Artelt, 2003b). Unter Berücksichtigung der Cluster-Struktur wurden die Korrelationen zwischen den WLE's und den zentrierten Schulnoten mit Hilfe von MPLUS berechnet. Dabei wurden die fehlenden Noten imputiert.
Für einen Vergleich mit Noten anderer Fächer wurde der Zusammenhang zusätzlich mit dem Datensatz der Validierungsstudie berechnet. In dieser Studie wurden die Schulnoten in den Fächern Mathematik, Deutsch, Biologie, Chemie, Physik und Naturwissenschaften mit einer Trackingliste erfasst. Die Information über die Noten stammte von den jeweiligen Lehrerinnen und Lehrern. In die Berechnung wurden die 1330 Schülerinnen und Schüler einbezogen, welche den NEPS-Mathematiktest bearbeitet hatten. Für diese Schülerinnen und Schüler wurde der Mathematik-WLE mit einem eindimensio-

nalen PCM-Modell mit der Software Conquest nach dem in Kapitel 3.3 beschriebenen Vorgehen geschätzt. Da in den unterschiedlichen Schulen und Klassen verschiedene Notensysteme vorlagen, wurden die Noten zuerst einheitlich auf Werte von 1 (sehr gut) bis 6 (ungenügend) kodiert. Anschließend wurden die Schulnoten der Schülerinnen und Schüler am Klassenmittelwert zentriert. Insgesamt fehlten von 6 der 1 330 Testpersonen die Angaben zur Mathematiknote und zur Deutschnote. Von 325 Personen fehlten die Angaben zur Biologienote, von 44 die Angaben zur Chemienote und von 85 die Angaben zur Physiknote. Die fehlenden Werte in den naturwissenschaftlichen Noten entstanden teilweise dadurch, dass in den Klassen einige Fächer nicht unterrichtet wurden. Diese fehlenden Werte wurden bei der Berechnung der Korrelationen zwischen den WLE's und den zentrierten Schulnoten unter Berücksichtigung der Cluster-Struktur mit Hilfe von MPLUS imputiert.

Die zweite Hypothese „Die mathematische Kompetenzmessung im NEPS hängt stark mit den mathematischen Kompetenzwerten, gemessen durch LV und PISA, zusammen." und die dritte Hypothese „Die mathematische Kompetenz im NEPS hängt deutlich niedriger mit den naturwissenschaftlichen Kompetenzwerten, gemessen durch PISA und LV, zusammen." wurden durch mehrere Berechnungen analysiert. Zum einen wurden die manifesten Korrelationen zwischen den NEPS-Mathematik-WLEs und den Mathematik-WLEs aus PISA und LV (für die Berechnung der WLEs siehe Kapitel 3.3) unter Berücksichtigung der Clusterstruktur mit dem Programm MPLUS berechnet. Für die visuelle Prüfung der Zusammenhänge zwischen den Studien und mit den Kompetenzstufen aus PISA und LV wurden zusätzlich Streudiagramme erstellt, in welchen die Grenzen der PISA- und LV-Kompetenzstufen eingezeichnet sind. Für eine Interpretation der Größenordnung der gefundenen Korrelationen wurden zusätzlich die manifesten Korrelationen zwischen den NEPS-WLEs und den Fähigkeitsschätzern für die naturwissenschaftliche Kompetenz in PISA und LV unter Berücksichtigung der Clusterstruktur berechnet.
Zudem wurden die latenten Korrelationen zwischen den Studien berechnet, indem zwei dreidimensionale IRT-Modelle skaliert wurden. Im ersten Modell wurde der NEPS-Mathematiktest auf der ersten Dimension, der LV-Mathematiktest auf der zweiten Dimension und der LV-Naturwissenschaftstest auf der dritten Dimension modelliert. Anschließend wurden die latenten Korrelationen zwischen den drei Dimensionen verglichen. In einem zweiten Modell wurde der NEPS-Mathematiktest auf der ersten Dimension, der PISA-Mathematiktest auf der zweiten Dimension und der PISA-Naturwissenschaftstest auf der dritten Dimension modelliert. Auch hier wurden die Korrelationen zwischen den

drei Dimensionen gegenübergestellt. Für die Evaluation dieser Hypothesen wurden die Daten aus der Validierungsstudie aus dem Jahr 2012 verwendet. Für das erste dreidimensionale Modell wurden diejenigen Schülerinnen und Schüler in den Berechnungen berücksichtigt, die jeweils Aufgaben des NEPS-Mathematiktests und des LV-Mathematiktests beantwortet hatten. Für das zweite dreidimensionale Modell wurden die Schülerinnen und Schüler berücksichtigt, die jeweils Aufgaben des NEPS-Mathematiktests und des PISA-Mathematiktests bearbeitet hatten.

Für die Untersuchung der vierten Hypothese „Die Testergebnisse im NEPS-K9-Mathematiktest hängen mit dem Kriterium kognitive Fähigkeiten zusammen." wurden sowohl die Daten der Haupterhebung als auch die Daten der Validierungsstichprobe verwendet. In der Haupterhebung wurde für die Erfassung fluider kognitiver Leistungspotenziale ein Matrizen-Test zum schlussfolgernden Denken eingesetzt (Universität Bamberg, 2012). Die Antworten der Schülerinnen und Schüler sowie der Summenwert des Tests waren im *scientific use file* integriert. Für die Berechnung des Zusammenhangs mit der Mathematikleistung wurde die Korrelation zwischen dem Mathematik-WLE und dem Summenwert des Matrizentests unter Berücksichtigung der Cluster-Struktur mit dem Programm MPLUS berechnet. Dabei wurden die 14 523 Schülerinnen und Schüler in die Berechnung einbezogen, welche am NEPS-Test teilgenommen hatten. Für 1 106 Schülerinnen und Schüler lagen keine Werte für den Matrizen-Test vor. Diese Werte wurden bei der Berechnung der Korrelation imputiert.
In der Validierungsstudie wurde die kognitive Fähigkeit mit dem Subtest des Berliner Test zur Erfassung fluider und kristalliner Intelligenz (BEFKI) zur Erfassung figuraler Aspekte gemessen. Anhand der Vorgaben für diesen Test wurde ein Summenwert für die Schülerinnen und Schüler gebildet. Unter Berücksichtigung der Cluster-Struktur wurde mit dem Programm MPLUS die Korrelation des Summenwertes mit dem WLE für die Testpersonen berechnet. Dabei wurden nur die Daten der Schülerinnen und Schüler berücksichtigt, welche die Aufgaben aus dem NEPS-Mathematiktest bearbeitet hatten. Von diesen 1330 Schülerinnen und Schülern hatten 26 den BEFKI nicht bearbeitet. Die Summenwerte dieser Schülerinnen und Schüler wurden bei der Berechnung der Korrelationen imputiert.

Die fünfte Hypothese „Die Testergebnisse im NEPS-K9-Mathematiktest hängen mit dem Kriterium prozedurale Metakognition zusammen." wurde anhand der Daten aus der Haupterhebung ausgewertet. In der Haupterhebung ist die deklarative Metakognition mit einem Strategiewissenstest gemessen worden. Der Test besteht aus acht Szenarios,

welche unterschiedliche Schul- und Freizeitaktivitäten beschreiben (Lockl, 2013). Sowohl die Antworten der Schülerinnen und Schüler als auch ein mittlerer Testwert der deklarativen Metakognition waren im *scientific use file* integriert. Der Zusammenhang zwischen dem Mathematik-WLE und dem Testwert der deklarativen Metakognition wurde mit der Software MPLUS unter Berücksichtigung der Cluster-Struktur berechnet.

4.6.2 Ergebnisse

Ergebnisse zu H1: Zusammenhang mit Zeugnisnoten

Die Ergebnisse für die Evaluation der Hypothese 1: „Die Leistung der Schülerinnen und Schüler im NEPS-Mathematiktest hängt stärker mit der Mathematiknote zusammen als mit Zeugnisnoten anderer Fächer.“ werden in den Tabellen 18 und 19 dargestellt. Die Höhe der gefundenen Korrelationen mit der Mathematiknote von r = -.28 für die Haupterhebung und r = -.34 für die Validierungsstudie haben eine ähnliche Höhe wie die in der PISA-2000-Studie gefundene Korrelation von r = -.32 zwischen dem nationalen, curriculumnahen Test und den Mathematiknoten (Deutsches PISA-Konsortium, 2003). Die Korrelationsmuster zeigen außerdem, dass die Mathematikleistung deutlich höher mit der Mathematiknote zusammenhängt als mit den Noten andere Fächer. Mit der Deutschnote wurde in beiden Studien die niedrigste Korrelation gefunden.

Tabelle 18: Korrelation der Mathematikleistung im NEPS-Test mit den Schuljahresnoten in Mathematik und Deutsch in der Haupterhebung

	N	Mathematik WLE	p
Mathematiknote	14523	-.28	< 0.01
Deutschnote	14523	-.01	0.05

Tabelle 19: Korrelation der Mathematikleistung im NEPS-Test mit den Schuljahresnoten in Mathematik, Deutsch, Biologie, Chemie und Physik in der Validierungsstudie

	N	Mathematik WLE	p
Mathematiknote	1330	-.32	< 0.01
Deutschnote	1330	-.18	< 0.01
Biologienote	1330	-.20	< 0.01
Chemienote	1330	-.22	< 0.01
Physiknote	1330	-.28	< 0.01

Ergebnisse zu H2 und H3: Zusammenhang mit mathematischer und naturwissenschaftlicher Kompetenz in PISA und LV

Die manifeste Korrelation zwischen den Mathematik-WLEs aus NEPS und PISA unter Berücksichtigung der Clusterstruktur beträgt $r = .69$ ($SE = 0.02$). Der Zusammenhang der Mathematik-WLEs aus NEPS und LV ist mit einer Korrelation von $r = .72$ ($SE = 0.02$) in einer vergleichbaren Höhe. In Abbildung 32 und 33 werden die manifesten Zusammenhänge zwischen der mathematischen Kompetenz sowie den Kompetenzstufen in PISA und LV und der mathematischen Kompetenz im NEPS visuell dargestellt. Die Ergebnisse zeigen, dass Schülerinnen und Schüler mit höheren Kompetenzwerten im NEPS in der Regel auch eine höhere Kompetenzstufe in PISA beziehungsweise im LV erreichten. Es wurde jedoch auch eine deutliche Streuung der Kompetenzwerte über die Kompetenzstufen sichtbar. Diese Abweichungen sind jedoch unter anderem den Reliabilitäten der Tests und der Anzahl der Kompetenzstufen geschuldet. Größere Abweichungen wurden vor allem in den Randbereichen deutlich, in welchen die mathematischen Fähigkeitsschätzer ungenauer sind.

Die Mathematik-WLEs aus dem NEPS korrelieren mit den Naturwissenschaft-WLEs aus PISA mit $r= .53$ ($SE = 0.02$) und den Naturwissenschafts-WLEs aus dem LV $r= .57$ ($SE = 0.04$) auf manifester Ebene deutlich niedriger.

Die latenten Zusammenhänge zwischen den mathematischen und naturwissenschaftli-

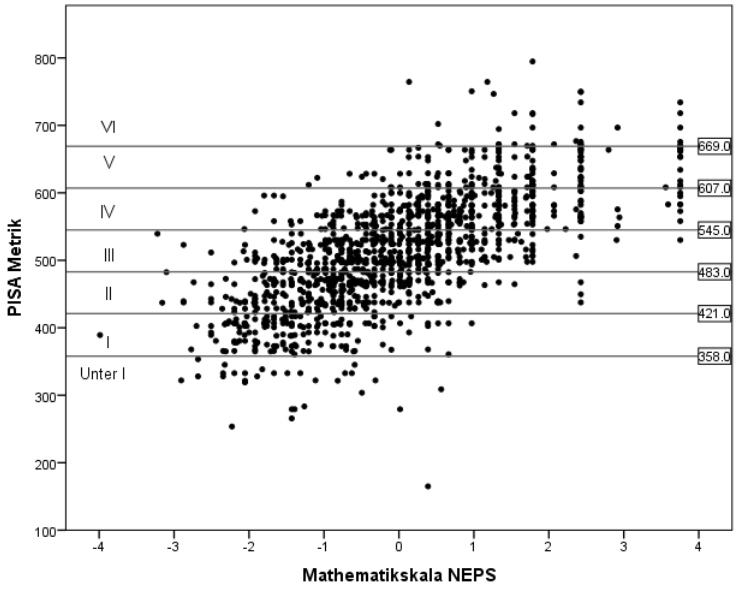


Abbildung 32: Zusammenhang zwischen der mathematischen Kompetenz im NEPS und der mathematischen Kompetenz in PISA

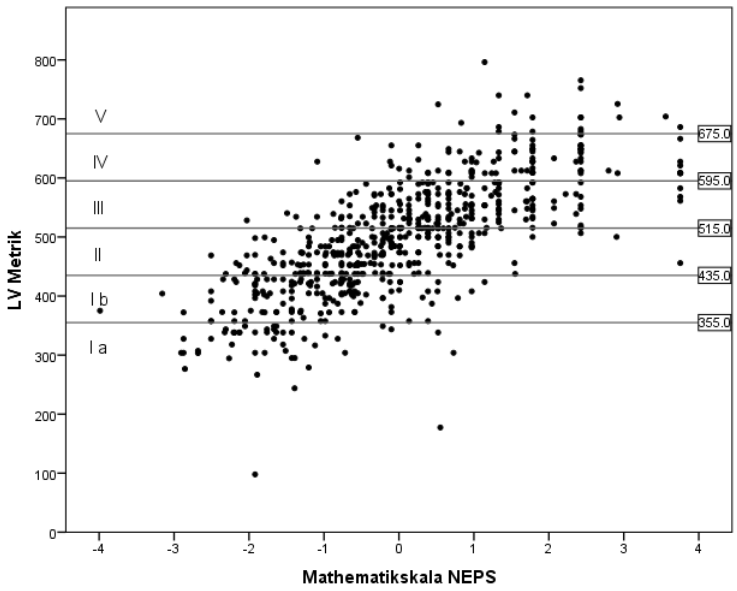


Abbildung 33: Zusammenhang zwischen der mathematischen Kompetenz im NEPS und der mathematischen im LV

chen Kompetenzen im NEPS, in PISA und im LV werden in Abbildung 34 gezeigt. Zwischen den mathematischen Kompetenztests aus beiden Studien wurde auch auf la-

tenter Ebene mit einer Korrelation von $r = .89$ ein deutlich höherer Zusammenhang gefunden als zwischen der mathematischen Kompetenz aus NEPS und der naturwissenschaftlichen Kompetenz aus PISA ($r = .72$).
Die Korrelationen zwischen dem NEPS-Mathematiktest und dem Mathematiktest sowie dem Naturwissenschaftstest aus dem LV werden in Abbildung 35 dargestellt. Auch hier wurde mit einer Korrelation von $r = .91$ ein höherer Zusammenhang zwischen den Mathematiktests beider Studien gefunden als zwischen dem NEPS-Mathematitest und dem LV-Naturwissenschaftstest ($r = .71$).

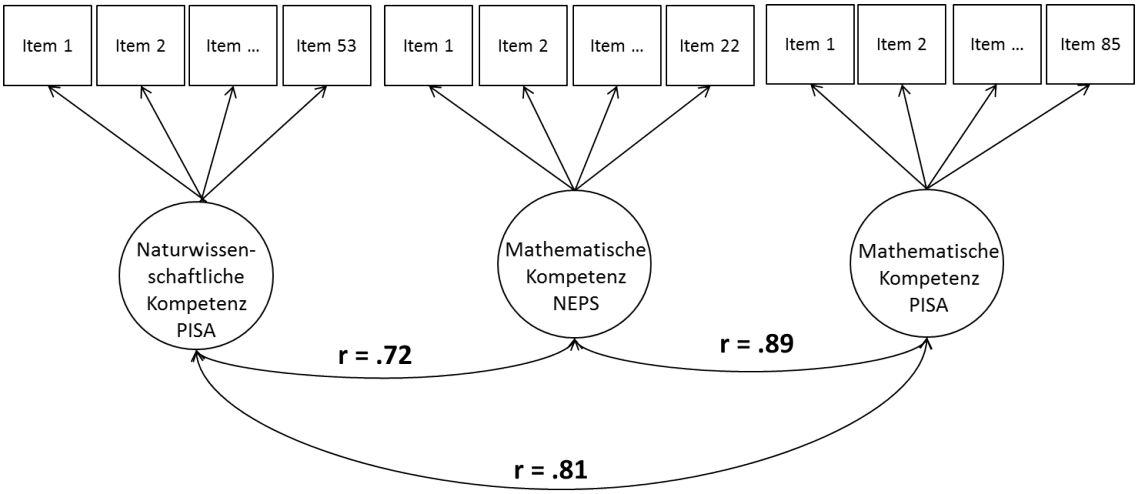


Abbildung 34: Latente Korrelation zwischen der mathematischen Kompetenz im NEPS und der mathematischen beziehungsweise naturwissenschaftlichen Kompetenz in PISA

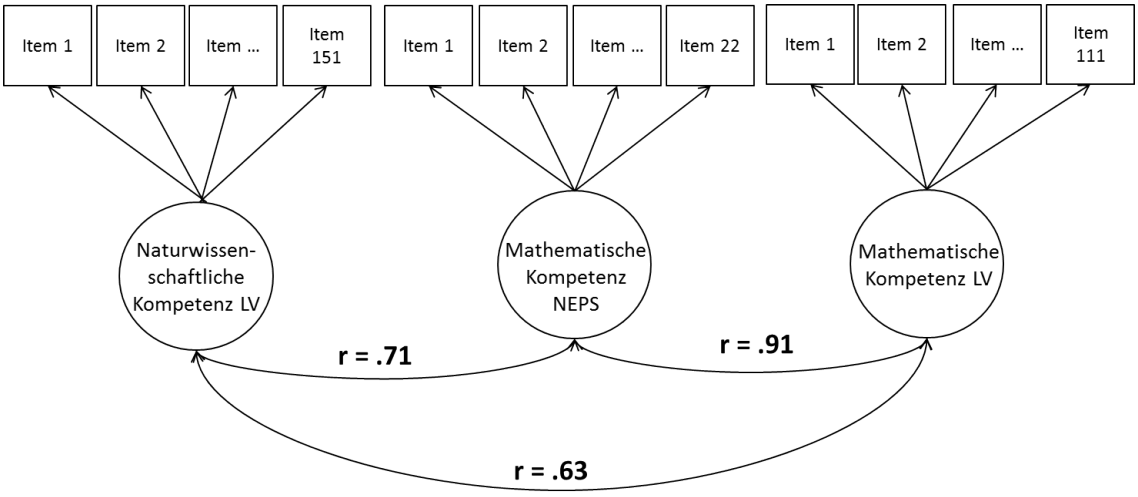


Abbildung 35: Latente Korrelation zwischen der mathematischen Kompetenz im NEPS und der mathematischen beziehungsweise naturwissenschaftlichen Kompetenz im LV

Die Hypothese über die Zusammenhänge der mathematischen Kompetenzmessung im NEPS mit den mathematischen und naturwissenschaftlichen Kompetenzmessungen im LV und in PISA wurde durch diese Ergebnisse bestätigt.

Ergebnisse zu H4: Zusammenhang mit kognitiven Fähigkeiten

Für die Evaluation der Hypothese „Die Testergebnisse im NEPS-K9-Mathematiktest hängen mit dem Kriterium kognitive Fähigkeiten zusammen." werden Ergebnisse aus sowohl der Haupterhebung als auch der Validierungsstudie vorgestellt. Die Korrelation der Mathematikleistung in der Haupterhebung mit dem Summenwert des Tests zum schlussfolgernden Denken beträgt $r = .49$ ($SE = 0.01$).
Die Korrelation der Mathematikleistung aus der Validierungsstudie mit dem Summenwert des Subtests zum figuralen Denken aus dem BEFKI beträgt $r = .53$ ($SE = 0.02$). Diese Werte zeigen jeweils einen substantiellen Zusammenhang der im NEPS gemessenen Mathematikleistung mit dem schlussfolgernden beziehungsweise figuralen Denken auf, weisen jedoch auch auf Unterschiede zwischen den Konstrukten hin.

Ergebnisse zu H5: Zusammenhang mit Metakognition

Die Ergebnisse zur Hypothese „Die Testergebnisse im NEPS-K9-Mathematiktest hängen mit dem Kriterium deklarative Metakognition zusammen." werden im Folgenden berichtet.
Die Korrelation zwischen dem Summenwert der deklarativen Metakognition und dem Mathematik-WLE beträgt $r = .27$ ($SE = 0.02$). Dieser Wert lässt auf einen Zusammenhang der mit dem NEPS-Test gemessenen Mathematikleistung mit der deklarativen Metakognition schließen, weist jedoch auch auf deutliche Unterschiede zwischen den Konstrukten hin.

4.6.3 Diskussion und Fazit *Extrapolation*

Insgesamt konnten Hinweise für die Validität der Schlussfolgerung *Extrapolation* gefunden werden. So wurden insgesamt höhere Korrelationen der Mathematikleistung im

NEPS mit der Mathematiknote gefunden als mit Noten anderer Fächer. Außerdem wurden starke manifeste und latente Zusammenhänge mit den Mathematikleistungen aus den Studien PISA und LV aufgezeigt. Des Weiteren entsprechen die Zusammenhänge mit kognitiver Fähigkeit und Metakognition den Erwartungen.

Dass die Mathematikleistung insgesamt höhere Korrelationen mit den Noten in der Validierungsstudie zeigt als mit den Noten in der Hauterhebung, könnte an Unterschieden in den Stichproben und/oder in der unterschiedlichem Erhebung der Noten liegen. In der Validierungsstudie wurden die Noten von den Lehrerinnen und Lehrern angegeben und in der Haupterhebung wurden die Schülerinnen und Schüler hinsichtlich ihrer Noten befragt. In den Mittelwerten und in der Verteilung der Testwerte wurde deutlich, dass die Validierungsstudie im NEPS-Mathematiktest leistungsstärker war als die Stichprobe der Haupterhebung. Jedoch zeigte ein Vergleich der Mathematik- und Deutschnoten zwischen den Stichproben, dass die Mittelwerte der Noten aus der Haupterhebung und der Validierungsstudie gleich hoch sind (siehe Anhang, 5.2.2). Aufgrund der positiven Selektion der Validierungsstichprobe (mehr Gymnasien etc.) liegt es nahe, dass diese Schülerinnen und Schüler auch im Mathematikunterricht leistungsstärker waren. Es ist daher zu vermuten, dass die Schülerinnen und Schüler der Haupterhebung eventuell höhere Noten angaben als sie tatsächlich auf dem letzten Zeugnis erreicht hatten oder dass sie eine im Vergleich vorteilhaftere Bewertung ihrer Lehrerinnen und Lehrer erhalten hatten.

Der Einsatz der Mathematikmessungen aus PISA und dem LV als Kriterium mathematischer Kompetenz kann insofern gerechtfertigt werden, als die kriteriale Interpretation der Mathematikleistung dieser Tests Inhalt der im NEPS definierten Zieldomäne ist. Kritisch kann jedoch angemerkt werden, dass es nur wenige Studien gibt, welche diese kriterialen Interpretationen für die PISA- und LV-Messungen validiert haben (vgl. Kapitel 1.2). Die Aussagekraft dieser Hinweise für Validität muss also im Verhältnis zur Validität der Testwertinterpretationen aus dem LV und aus PISA interpretiert werden. Des Weiteren ist die exakte Höhe der gefundenen Zusammenhänge aufgrund der Unterschiede zwischen der Validierungsstichprobe und der Haupterhebungsstichprobe nur für die bei dieser Berechnung verwendete Validierungsstichprobe aussagekräftig. Jedoch ist zu erwarten, dass die Interpretation der Ergebnisse gleichbleibend ist. Denn trotz eines eventuellen Unterschiedes in der Höhe des Zusammenhanges mit der Stichprobe der Haupterhebung haben die Abweichungen beider Stichproben von der Normalverteilung in ihrer Verteilung der Mathematikleistungen die gleichen Tendenzen (vgl. Anhang,

5.2.2).

Die Höhe des Zusammenhanges der Metakognition mit dem NEPS-Mathematiktest liegt deutlich unter den Befunden aus anderen Studien (Lingel et al., 2014; Schneider & Artelt, 2010, $r > .38$). Dies lässt sich unter anderem dadurch erklären, dass im NEPS im Gegendatz zu den anderen Studien kein mathematikspezifisches Strategiewissen erfasst wurde. Im NEPS-Test zur Erfassung der deklarativen Metakognition haben nur fünf von den acht eingesetzten Szenarios einen schulischen Kontext. Zwei dieser fünf schulischen Szenarios beziehen sich wiederum auf die Domäne Lesen. Zum anderen handelt es sich bei der Korrelation aus der Studie von Lingel et al. (2014) um eine latente Korrelation. Im NEPS wurde jedoch für die deklarative Metakognition ein mittlerer Testwert gebildet, welcher mit dem Mathematik-WLE korreliert wurde. Aus diesem Grund ist von einer Fehlerbehaftung des Wertes auszugehen.

Insgesamt konnten Hinweise für die Schlussfolgerung der *Extrapolation*, dass von der im NEPS gemessenen mathematischen Kompetenz auf die Kompetenz in der Zieldomäne geschlossen werden kann, gefunden werden. In dieser Studie wurden jedoch lediglich konkurrente Kriterien für die Zieldomäne erfasst. Weitere Analysen mit Kriterien, welche die Zieldomäne besser repräsentieren, sind jedoch für die Validierung der Schlussfolgerung notwendig. Beispielsweise böte sich in Zukunft im NEPS die Möglichkeit, die mathematische Kompetenzmessung in NEPS-K9 als möglichen Prädiktor für Bildungserfolg, erreichten sozioökonomischen Status, Wahl des Ausbildungszweiges etc. zu untersuchen.

5 Gesamtdiskussion

Nach Entwicklung des Interpretation/Use Argument (IUA) und der Bildung des Validitätsargumentes stellt sich die Frage, was die Ergebnisse aus dem Validitätsargument für die Nutzung des NEPS-K9-Mathematiktest bedeuten und wie diese Ergebnisse einzuordnen sind. In diesem Kapitel soll nach einer kurzen Zusammenfassung ein Fazit aus dem Validitätsargument geschlossen werden. Im Anschluss daran werden das Interpretation/Use Argument (IUA) und das Validitätsargument reflektiert und Implikationen für Nutzerinnen und Nutzer der Testdaten formuliert.

5.1 Validität der Testwertinterpretation des NEPS-K9-Mathematiktests: Ein Fazit

5.1.1 Zusammenfassung des Validitätsarguments

In dieser Arbeit wurde ein IUA für den NEPS-K9-Mathematiktest basierend auf dem Argument Based Approach von Kane (1990, 2001, 2006, 2008, 2012, 2013) entwickelt. Die zuverlässige Messung der Mathematikkompetenz am Ende der Sekundarstufe ist von Belang, da es sich hierbei um eine Schlüsselkompetenz zum Ende der Pflichtschulzeit handelt. Der Argument Based Approach wurde für die Validierung der Testwertinterpretation gewählt, da er auf einem modernen Validitätskonzept beruht und da durch die Anwendung des Ansatzes eine transparente und logische Argumentationskette zur Evaluierung der Testwertinterpretationen aufgebaut wird (Kane, 2013). Des Weiteren schafft diese Arbeit ein praxisbezogenes Beispiel zur Anwendung des Validierungsansatzes. Kanes theoretische Beschreibungen des Ansatzes (1990, 2001, 2006, 2008, 2012, 2013) dienten als Basis für die Entwicklung des Modells. Kane (2013) betonte jedoch, dass

seine Beschreibungen der Schlussfolgerungen nicht ausschöpfend seien und dass die Argumentationskette an den Test und die zu validierende Testwertinterpretation angepasst werden müsse. Das hier entwickelte IUA wurde daher für den NEPS-K9-Mathematiktest adaptiert und enthält neben den drei von Kane vorgeschlagenen Schlussfolgerungen für eine deskriptive Testwertinterpretation drei zusätzliche Schlussfolgerungen. In dieser Studie wurde die von Kane vorgeschlagene Schlussfolgerung *Entscheidung* nicht ausgewertet. Mit dem NEPS-K9-Mathematiktest werden keine konkreten Entscheidungsregeln für die Testnutzung ermöglicht, welche Konsequenzen für Testpersonen hätten. Aus diesem Grund bleibt die Testwertinterpretation auf deskriptiver Ebene. Die Testwertinterpretation, die mit dem NEPS-K9-Mathematiktest untersucht wurde, ist daher: „Die Unterschiede in der beobachteten Testleistung geben die Unterschiede in der mathematischen Fähigkeit wieder, wie diese im Rahmenkonzept definiert wird." (vgl. Kapitel 2). Die Schlussfolgerungen *Domänenbeschreibung*, *Bewertung*, *Skalierung*, *Generalisierung*, *Konstruktbezug* und *Extrapolation* wurden in Kapitel 2 entwickelt. Dabei wurden für jede Schlussfolgerung die Argumente, Annahmen und Hypothesen aus Literatur abgeleitet. Jede Schlussfolgerung wurde anschließend in einer Abbildung mit ihren Argumenten, Annahmen und Hypothesen visuell dargestellt.
In Kapitel 4 wurde das sogenannte Validitätsargument erstellt. Dafür wurden alle Schlussfolgerungen des IUA evaluiert, indem die Hypothesen und damit auch die Annahmen und Argumente ausgewertet wurden. Die Ergebnisse wurden diskutiert und Einschränkungen der Schlussfolgerungen und Interpretationen wurden aufgedeckt.
Im Folgenden werden die Ergebnisse zu der Evaluation der Schlussfolgerungen noch einmal kurz zusammengefasst. Tabelle 20 zeigt alle Schlussfolgerungen, Argumente, Annahmen und Hypothesen sowie die gefundenen Hinweise für Validität und die gefundenen Einschränkungen für die Testwertinterpretationen. Bei der Auswertung der Schlussfolgerung *Domänenbeschreibung* wurde gezeigt, dass die Zieldomäne für den NEPS-K9-Mathematiktest in Bezug auf die Zieldomänen von LV und PISA eingegrenzt werden muss. Der NEPS-K9-Mathematiktest soll zwar Mathematical Literacy basierend auf der Definition von PISA erfassen, jedoch beinhaltet die Zieldomäne des NEPS vor allem persönliche Kontexte und nicht den Prozess „mathematische Hilfsmittel verwenden". Aktive Komponenten von Prozessen gehören ebenfalls nicht zur Zieldomäne. Der NEPS-Test soll zwar auch curriculare Inhalte basierend auf dem LV-Rahmenkonzept erfassen, jedoch wird auch dieses nur mit Einschränkungen getan. So wird der Anforderungsbereich aus dem LV „Verallgemeinern und reflektieren" nicht erfasst und gibt es kaum Aufgaben im NEPS, die den LV-Inhaltsbereich „Raum und Form" abdecken. Insgesamt deckt

der NEPS-K9-Mathematiktest zwar wichtige Teilkomponenten des internationalen Mathematical Literacy Konzeptes aus PISA 2012 und wichtige curriculare Komponenten aus dem LV 2012 ab, jedoch erfasst der Test vermutlich nicht die gleiche Mathematical Literacy und die gleichen curricularen Aspekte wie PISA und LV. Schon mit seiner eingeschränkten Testzeit und Itemzahl kann der Test diese Konzepte nicht mit einer vergleichbaren Genauigkeit messen. Die Zieldomäne Mathematische Kompetenz hat daher zwar große Ähnlichkeit zu den Zieldomänen aus PISA und dem LV, ist jedoch nicht die gleiche. Auch konnten aus den Testwerten keine Rückschlüsse auf Teilkomponenten von PISA und dem LV gezogen werden, da diese durch die NEPS-Aufgaben nicht ausreichend repräsentiert werden. Für eine Annahme der Schlussfolgerung *Domänenbeschreibung* muss die Zieldomäne enger definiert werden. So müssen die aktiven Komponenten der Prozesse beispielsweise aus der Definition der Zieldomäne ausgeschlossen werden.

Für die *Bewertung* wurde gezeigt, dass alle Hypothesen bestätigt werden konnten. Es konnte gezeigt werden, dass die Bewertungskriterien die Bildung korrekter Testergebnisse erlauben, dass die Aufgabenkodierung bestimmungsgemäß und den Kodieranweisungen entsprechend durchgeführt wurde und dass die psychometrische Qualität des Testinstrumentes gewährleistet ist. Die Rohwerte der NEPS-Mathematikaufgaben führen somit zu Testergebnissen, die repräsentativ für die Zieldomäne Mathematische Kompetenz sind.

Die Evaluation der Schlussfolgerung *Skalierung* zeigte, dass die Interpretation der Testwerte angepasst werden muss. Nicht alle Voraussetzungen für das verwendete IRT-Modell konnten bestätigt werden. So zeigte zwar keine der NEPS-Aufgaben einen bedeutsamen DIF, jedoch wurden für Geschlecht und Schulform bessere Modellgütekriterien für ein Modell gefunden, welches neben Haupteffekten auch DIF zulässt. Für diese beiden Subgruppen konnte daher nicht ausgeschlossen werden, dass neben der Variable mathematische Fähigkeit auch eine andere Variable gemessen wird, die zu Unterschieden in den Lösungswahrscheinlichkeiten innerhalb der Subgruppen bei gleicher mathematischer Fähigkeit führt. Außerdem wurden unterschiedliche Trennschärfen im Sinne von Steigungsparametern für die Aufgaben gefunden. Die lokale stochastische Unabhängigkeit sowie die Passung der beobachteten Wahrscheinlichkeiten zu den vom Modell vorhergesagten Wahrscheinlichkeiten konnten jedoch bestätigt werden. Insgesamt muss mit leichten Ungenauigkeiten in den Schätzungen der Personenfähigkeiten und Itemparameter gerechnet werden. Die Interpretation dieser Schlussfolgerung muss also angepasst werden, indem nicht von einer Widerspiegelung der wahren mathematischen Fähigkeit durch die Testergebnisse ausgegangen wird, sondern von einer Annäherung an die wahre mathematische

Kompetenz in der Zieldomäne.

Für die Schlussfolgerung *Generalisierung* konnte dargelegt werden, dass die Durchführungsbedingungen der Messung standardisiert waren. Zwar fehlten präzise Informationen bezüglich der Instruktion, der Testumgebung, Störungen in den Testsitzungen und des Qualitätsmonitoring, doch entsprechen die Durchführungsbedingungen weitgehend den Standards for Educational and Psychological Testing (Standards) von 2014. Auch konnte eine für die meisten Schülerinnen und Schüler angemessene Messgenauigkeit des Tests gezeigt werden. Für Testpersonen mit sehr hoher oder sehr niedriger mathematischen Fähigkeit wurden große Abweichungen gefunden. Es konnte daher geschlussfolgert werden, dass die Fähigkeiten auf der latenten Skala im mittleren Fähigkeitsbereich ($-2 \leq \theta \leq +2$) angemessene Schätzer für erwartete Ergebnisse in parallelen Messungen sind. Bei der Testwertinterpretation sollten die relativ weiten Konfidenzintervalle für die Personenfähigkeitsschätzer jedoch berücksichtigt werden. Die Fähigkeiten in den Randbereichen müssen mit Vorsicht interpretiert werden, da es hier zu größeren Abweichungen kommen kann. Die Generalisierung gilt jedoch nur, wie bereits in Kapitel 2.4 beschrieben, für die standardisierten Bedingungen. Ob die Testergebnisse sich auch auf das Konstrukt mathematischer Kompetenz zurückführen lassen und ob sie die mathematische Kompetenz in der Zieldomäne repräsentieren, wurde in den letzten beiden Schlussfolgerungen ausgewertet.

Das Konstrukt mathematischer Kompetenz, auf welchem der NEPS-K9-Mathematiktest basiert, wird im Rahmenkonzept beschrieben. Die Auswertung der Schlussfolgerung *Konstruktbezug* zeigte, dass sich die im Rahmenkonzept beschriebene dimensionale Struktur nicht vollständig bestätigen lässt. Das mathematische Konstrukt des NEPS ließ sich zwar von anderen NEPS-Konstrukten abgrenzen, jedoch konnte die Eindimensionalität des Tests nicht bestätigt werden. Die hohen Korrelationen zwischen den Inhaltsbereichen weisen auf eine übergreifende mathematische Kompetenz, die zum Lösen aller Aufgaben benötigt wird, hin. Jedoch muss davon ausgegangen werden, dass bestimmte Aufgaben auch spezifische Fähigkeiten erfordern. Als Konsequenz muss die Interpretation der Schlussfolgerung angepasst werden. Ein vereinfachtes Konstrukt mathematischer Kompetenz kann durch die Testwerte angemessen repräsentiert werden. Mit den bisherigen Auswertungen konnten noch keine soliden Schlussfolgerungen für die tatsächliche empirische Struktur des Konstruktes gemacht werden.

Die Evaluation der Schlussfolgerung *Extrapolation* zeigte, dass die mathematische Kom-

petenz, gemessen mit dem NEPS-K9-Mathematiktest, erwartungsgemäß mit den Kriterien Schulnoten, (meta-)kognitive Fähigkeit und mathematische Fähigkeit im LV und in PISA zusammenhängt. Die Kriterien Schulnoten und (meta-)kognitive Fähigkeit sind Kriterien, die im Zusammenhang mit der Zieldomäne stehen, diese jedoch nicht repräsentieren. Die mathematischen Fähigkeiten aus dem LV und aus PISA repräsentieren Teile der Zieldomäne, jedoch muss die Zieldomäne des NEPS enger definiert werden (siehe *Domänenbeschreibung*). Dementsprechend fehlten für die Auswertung des IUA Kriterien, die die Zieldomäne Mathematische Kompetenz im Alltag des NEPS widerspiegeln. Aus diesem Grund konnte lediglich die Interpretation validiert werden, dass die mit dem NEPS-K9-Mathematiktest erfasste Kompetenz ein Indikator für mathematische Kompetenz im Alltag ist. Diese Interpretation konnte jedoch bestätigt werden.

Tabelle 20: Validitätsargument

Schlussfolgerung	zugrundeliegendes Argument	zugrundeliegende Annahmen	Hypothesen	Hinweise für Validität	Einschränkungen
Domänenbeschreibung	Die Rohwerte im NEPS-Mathematiktest spiegeln relevante und repräsentative Elemente für die Zieldomäne Mathematische Kompetenz wider.	1. Die Aufgaben des NEPS-Mathematiktests decken die für die Zieldomäne relevanten Teilkompetenzen angemessen ab.	1. In den Aufgaben des NEPS-K9-Tests lassen sich die relevanten Domänen mathematischer Kompetenz aus dem Rahmenkonzept von PISA und dem des LV identifizieren.	Identifikation relevanter Teilkompetenzen aus PISA und LV in den NEPS-Aufgaben.	Keine Abdeckung des PISA-Prozesses „mathematische Hilfsmittel verwenden“ und LV-Anforderungsbereiches „Verallgemeinern und Reflektieren“ durch die NEPS-Aufgaben.
			2. Die kognitiven und inhaltlichen Teilbereiche werden in NEPS und dem LV bzw. NEPS und PISA auf ähnliche Weise operationalisiert.	Ähnliche Operationalisierung der inhaltlichen Teilkompetenzen im NEPS wie in PISA und im LV.	Kein Vergleich mit den Prozessen möglich.
			3. Die Gewichtung der Komponenten aus den PISA- und LV-Rahmenkonzepten im NEPS-K9-Mathematiktest unterscheidet sich nicht signifikant von der Gewichtung in den Mathematiktests aus den Haupterhebungen PISA 2012 und LV 2012.	Keine Unterschiede in der Gewichtung der inhaltlichen Teilkompetenezen zu der Gewichtung in PISA 2012 und LV 2012 und zu der Gewichtung der Anforderungsbereiche in PISA 2012.	Unterschiedliche Gewichtung der PISA-Kontexte. Kein Vergleich der Gewichtung der LV-Anforderungsbereiche und der LV- und PISA-Prozesse möglich.
Bewertung	Die Rohwerte in den NEPS-Mathematikaufgaben führen zu Testergebnissen, die repräsentativ für die Zieldomäne Mathematische Kompetenz sind.	1. Die Aufgabenkodierung wurde bestimmungsgemäß und den Kodieranweisungen entsprechend durchgeführt.	1. Fehler in der Eingabe und Kodierung der Testhefte aller Schülerinnen und Schüler können ausgeschlossen werden.	Hohe Standardisierung der Eingabe und Kodierung der Testhefte.	Keine präzisen, veröffentlichten Informationen zur Prüfung der Integrität der Daten sowie zur Qualitätsprüfung des Scorings.

Fortsetzung auf der nächsten Seite

Schlussfolgerung	zugrundeliegendes Argument	zugrundeliegende Annahmen	Hypothesen	Hinweise für Validität	Einschränkungen
		2. Die Bewertungskriterien des NEPS-Mathematiktests für die 9. Klassenstufe ermöglichen die Bildung korrekter Testergebnisse.	2. Die unterschiedlichen Gewichtungen der Aufgabenformate und der Umgang mit fehlenden Werten führen zu unverfälschten Testergebnissen.	Gewichtung der Aufgabenformate und Umgang mit fehlenden Werten im NEPS führt zu unverfälschten Testergebnissen.	
		3. Eine hohe psychometrische Qualität des Testinstruments ist gewährleistet.	3. Die Aufgaben sind trennscharf.	Angemessene Trennschärfen der Aufgaben.	
			4. Die Qualität der Distraktoren ist angemessen.	Angemessene Trennschärfen der Distraktoren.	
			5. Die Aufgaben sind intern konsistent.	Zufriedenstellendes Cronbachs-α.	
Skalierung	Das Testergebnis führt zu Fähigkeitsschätzern, welche die mathematische Kompetenz der Schülerinnen und Schüler widerspiegeln.	1. Das verwendete IRT-Modell passt zu den Daten.	1. Für den Test kann spezifische Objektivität festgestellt werden.	Keine bedeutamen DIF-Werte für NEPS-Aufgaben.	Bessere Modellgütekriterien für DIF-Modell für die Variablen Geschlecht und Schulform.
			2. Die Aufgaben des Tests sind lokal stochastisch unabhängig.	Sehr niedrige partielle Inter-Item-Korrelationen.	
			3. Die Aufgaben haben die gleichen Trennschärfen.		Unterschiedliche Trennschärfen und bessere Modellgütekriterien bei einer zwei 2PL-Modellierung.
			4. Die beobachtete Antwortwahrscheinlichkeit weicht nicht signifikant von der mit dem Modell vorhergesagten Wahrscheinlichkeit ab.	Zufriedenstellende WMNSQ-Werte.	

Fortsetzung auf der nächsten Seite

Schlussfolgerung	zugrundeliegendes Argument	zugrundeliegende Annahmen	Hypothesen	Hinweise für Validität	Einschränkungen
Generalisierbarkeit	Die Fähigkeiten auf der latenten Skala sind angemessene Schätzer für erwartete Ergebnisse in parallelen Messungen.	1. Die Durchführungsbedingungen der Messung sind standardisiert.	1. Die Testung wurde sorgfältig anhand von angemessenen, standardisierten Prozeduren durchgeführt.	Durchführungsbedingungen entsprechen zum großen Teil den Standards for Educational and Psychological Testing	Fehlende Informationen bezüglich der Instruktion, der Testumgebung, Störungen der Testsitzungen und des Qualitätsmonitorings der Durchführung.
		2. Die Messgenauigkeit des Tests ist angemessen.	2. Die individuellen Fähigkeitsmessungen sind reliabel.	Hinweise für angemessene Messungen im mittleren Fähigkeitsbereich und angemessene EAP/PV-Reliabilität.	Die Messungen in den Randbereichen fallen ungenau aus.
Konstruktbezug	Die generalisierten Ergebnisse im NEPS-Mathematiktest für die neunte Klasse sind auf das Konstrukt der mathematischen Kompetenz zurückzuführen.	1. Die angenommene dimensionale Struktur des NEPS-Mathematikkonstrukts lässt sich analytisch bestätigen.	1. Innerhalb der NEPS-Teildimensionen können sehr hohe Zusammenhänge gefunden werden.		Zusammenhänge zwischen den Teildimensionen sind niedriger als erwartet.
			2. Für den NEPS-Mathematiktest ist eine eindimensionale Skalierung einer mehrdimensionalen Skalierung vorzuziehen.		Bessere Modellgütekriterien für eine mehrdimensionale Skalierung.

Fortsetzung auf der nächsten Seite

Schlussfolgerung	zugrundeliegendes Argument	zugrundeliegende Annahmen	Hypothesen	Hinweise für Validität	Einschränkungen
			3. Für eine mehrdimensionale Skalierung der mathematischen, naturwissenschaftlichen, Lese- und ICT-Kompetenzen aus dem NEPS auf separaten Dimensionen wird eine bessere Passung gefunden als bei einer eindimensionalen Skalierung der Kompetenzen auf einer gemeinsamen Dimension.	Bessere Modellgütekriterien für eine mehrdimensionale Skalierung mathematischer, naturwissenschaftlicher, Lese- und ICT-Kompetenzen aus NEPS auf jeweils einer eigenen Dimension.	
			4.Die mathematische Kompetenz im NEPS hängt stärker mit der naturwissenschaftlichen Kompetenz zusammen als mit der Kompetenz Lesen.	Erwartungsgemäße Zusammnehänge des NEPS-Mathematiktests mit der naturwissenschaftlichen und der Lesekompetenz.	
Extrapolation	Die Kompetenz der Zieldomäne, wie sie mit dem NEPS-Test erfasst wird, ist ein Indikator für die Leistung in der Zieldomäne Mathematische Kompetenz.	1. Die Testleistung im NEPS-Mathematiktest für die neunte Klasse hängt mit der mathematischen Kompetenz in der Zieldomäne zusammen.	1. Die Leistung der Schülerinnen und Schüler im NEPS-Mathematiktest hängt stärker mit der Mathematiknote zusammen als mit Zeugnisnoten anderer Fächer.	Höhere Korrelation des NEPS-Tests mit der Mathematiknote als mit Noten anderer Fächer.	Die hier genutzten Kriterien sind nur eingeschränkt für die Leistung der mathematischen Kompetenz in der Zieldomäne zu interpretieren.
			2. Die mathematische Kompetenz im NEPS hängt stark mit den mathematischen Kompetenzwerten, gemessen durch PISA und den LV, zusammen.	Starke Zusammenhänge der mathematischen Kompetenz aus dem NEPS mit der mathematischen Kompetenz aus PISA und dem LV.	

Fortsetzung auf der nächsten Seite

Schlussfolgerung	zugrundeliegendes Argument	zugrundeliegende Annahmen	Hypothesen	Hinweise für Validität	Einschränkungen
			3. Die mathematische Kompetenz im NEPS hängt deutlich weniger mit den naturwissenschaftlichen Kompetenzwerten, gemessen durch PISA und den LV, zusammen.	Deutlich geringere Zusammenhänge zwischen der mathematischen Kompetenz aus NEPS und der naturwissenschaftlichen Kompetenz aus PISA und dem LV.	
			4. Die Testergebnisse im NEPS-K9-Mathematiktest hängen mit dem Kriterium kognitive Fähigkeiten zusammen.	Erwartungsgemäßer Zusammenhang der mathematischen Kompetenz aus NEPS mit der kognititven Fähigkeit.	
			5. Die Testergebnisse im NEPS-K9-Mathematiktest hängen mit dem Kriterium deklarative Metakognition zusammen.	Erwartungsgemäßer Zusammenhang der mathematischen Kompetenz aus NEPS mit der Metakognition.	

5.1.2 Diskussion des IUA und des Validitätsarguments

Fazit zum IUA und zum Validitätsargument

Was ist das Fazit des Validitätsarguments? Sind die Testwertinterpretationen des NEPS-K9-Mathematiktests valide? Kane (2013) schrieb, dass eine beabsichtigte Testwertinterpretation nach sorgfältiger Prüfung der Schlüssigkeit und der Plausibilität der Schlussfolgerungen und Annahmen als valide angenommen werden könne. Dies geschehe jedoch unter Berücksichtigung der Möglichkeit, dass neue Auswertungen zu einer Neubewertung der Konklusion führen können (Kane, 2013). Auch die Standards von 2014 gaben an, dass der Validierungsprozess niemals vollendet sei, da es immer zusätzliche Informationen gebe, die gesammelt werden können, um den Test und seine Schlussfolgerungen noch besser verstehen zu können. Jedoch könne bei genügend unterstützender und vertretbarer Evidenz eine zusammenfassende Beurteilung hinsichtlich der Validität einer Testwertinterpretation getroffen werden (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 2014, S.22). Ein solides Validitätsargument integriere dabei diverse Validitätsbeweise in einer schlüssigen Darstellung des Grades, indem die Beweise und Theorien die beabsichtigte Testwertinterpretation stützen. Dementsprechend ist die zusammenfassende, deskriptive Testwertinterpretation, die für den NEPS-K9-Mathematiktest in dieser Arbeit evaluiert werden sollte „Die Unterschiede in der beobachteten Testleistung geben die Unterschiede in der mathematischen Fähigkeit wieder, wie diese im Rahmenkonzept definiert wird.“ zu dem Grad valide, in dem das Validitätsargument die Schlussfolgerungen des IUA unterstützt. In der Zusammenfassung wurden alle Validitätsbeweise und auch alle Einschränkungen noch einmal dargelegt. Vom ursprünglichen IUA mussten einige Schlussfolgerungen angepasst werden und es wurde weiterer Forschungsbedarf aufgedeckt. Der Grad, zu welchem das IUA valide ist, wird im Folgenden mit Hilfe der Abbildung 36 verdeutlicht. Der schwarze Kreis stellt die mit dem NEPS-Test möglichen Testwertinterpretationen dar. Der graue Kreis beinhaltet die Zieldomäne Mathematische Kompetenz, wie sie im NEPS-Rahmenkonzept definiert ist und welche damit die beabsichtigte Testwertinterpretation repräsentiert. Der Bereich 1, in dem sich die beiden Kreise überschneiden, beinhaltet die Evidenzen, welche eine Interpretation der Testergebnisse als Repräsentation der Zieldomäne unterstützen. In diesen Bereich fallen die hohen gefundenen Übereinstimmungen zwischen dem NEPS-Test und den Rahmenkonzepten aus dem LV und aus PISA, die Evidenzen für die Standardisierung der Testergebnisse und die

den Erwartungen entsprechenden Zusammenhänge mit anderen Variablen. Der Bereich 2, welcher sich außerhalb des grauen Kreises befindet, beinhaltet die Evidenzen, die auf Konstruktinvarianz deuten. In diesem Bereich werden Interpretationen der Testwerte außerhalb der Zieldomäne unterstützt. Der Hinweis auf DIF sowie Ungenauigkeiten in den oberen und unteren Fähigkeitsbereichen lässt sich hier einordnen. Im Bereich 3, welcher sich außerhalb des schwarzen Kreises befindet, können die Evidenzen eingeordnet werden, die auf Einschränkungen der Interpretation der Testwerte als Repräsentation der Zieldomäne hinweisen. In diesen Bereich fallen die Evidenzen für die unvollständige Abdeckung der Teilkompetenzen aus PISA und dem LV. Durch die fehlenden Analysen mit Variablen, welche die Zieldomäne repräsentieren, ist die Größe dieses Bereiches nicht vollständig abzuschätzen.

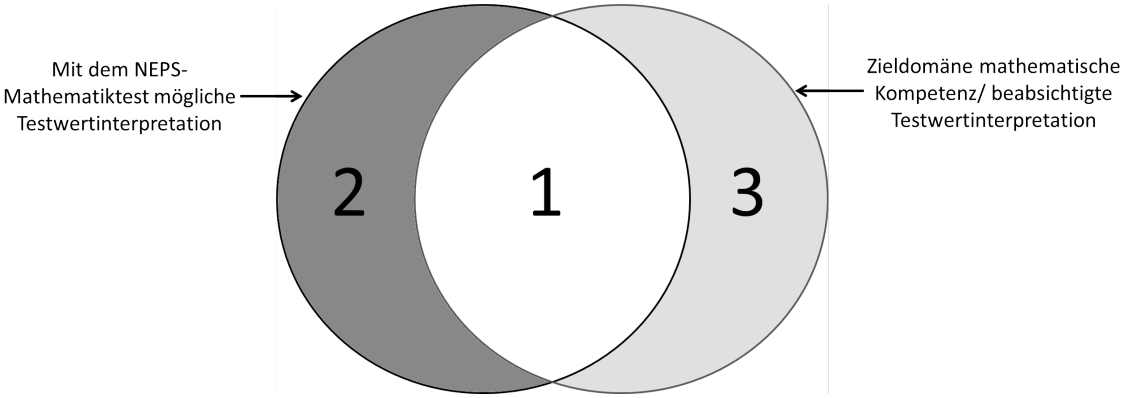


Abbildung 36: Übereinstimmung der Evidenzen mit der beabsichtigten Testwertinterpretation

Eine vollständige Überlappung der beiden Kreise würde die Testwertinterpretation, dass die Unterschiede in der beobachteten Testleistung Unterschiede in der Zieldomäne Mathematische Kompetenz wiedergeben, vollständig unterstützen. Wie Tabelle 20 jedoch zeigt, werden auch Evidenzen, die in den Bereich 2 und 3 fallen, gefunden. Es kann daher nicht ausgeschlossen werden, dass Unterschiede in der beobachteten Testleistung auf Unterschiede durch konstruktirrelevante Varianz in Bereich 2 zurückzuführen sind. Des Weiteren kann nicht auf den Bereich 3 der Zieldomäne geschlossen werden. Durch die Formulierung des IUA wurde der graue Kreis, also die beabsichtigte Testwertinterpretation oder Zieldomäne, definiert. Das Validitätsargument macht den schwarzen Kreis, also die tatsächlich möglichen Interpretationen, sichtbar. Da sich der schwarze Kreis jedoch nicht vollständig mit dem grauen Kreis deckt, sondern Hinweise für die Bereiche 2 und 3 gefunden wurden, ist folgende einschränkende Testwertinterpretation, basierend

auf dem hier entwickelten IUA und Validitätsargument, möglich:

„Die Unterschiede in der Rangfolge der beobachteten Testleistung deuten auf Unterschiede in der Rangfolge der Ausprägung der mathematischen Kompetenz, wie diese im Rahmenkonzept definiert wird, hin.“ .

Tabelle 21 zeigt ein aktualisiertes IUA mit den Argumenten, welche durch das Validitätsargument gestützt werden und welche die neu formulierte Testwertinterpretation stützen.

In den Standards wird beschrieben, dass beinahe in allen Tests Elemente fehlen, die einige potenzielle Nutzerinnen und Nutzer als wichtig erachten, und dafür Elemente beinhaltet sind, die potenzielle Nutzerinnen und Nutzer als ungeeignet erachten. Validierung umfasst die sorgfältige Prüfung von möglichen Verzerrungen, welche aus der inadäquaten Repräsentation des Konstruktes und Aspekten der Messung entstehen (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 2014). In dieser Arbeit wurden die Testwertinterpretationen auf Annahmen untersucht, die nicht bestätigt werden können und die zu fälschlichen Schlussfolgerungen aus den Testergebnissen führen können. Die Testwertinterpretationen wurden anschließend so umformuliert, dass unverfälschte Aussagen anhand der Testergebnisse möglich sind.

Tabelle 21: Das IUA für den NEPS-K9-Mathematiktest angepasst an das Validitätsargument

Schlussfolgerung	zugrundeliegendes Argument	zugrundeliegende Annahmen
Domänenbeschreibung	Die Rohwerte im NEPS-Mathematiktest spiegeln relevante und repräsentative Elemente für die (angepasste) Zieldomäne Mathematische Kompetenz wider.	Die Aufgaben des NEPS-Mathematiktests decken die für die Zieldomäne relevanten Teilkompetenzen angemessen ab.
Bewertung	Die beobachteten Rohwerte in den NEPS-Mathematikaufgaben führen zu Testergebnissen, die repräsentativ für die Zieldomäne Mathematische Kompetenz sind.	Die Bewertungskriterien des NEPS-Mathematiktests für die 9. Klassenstufe ermöglichen die Bildung korrekter Testergebnisse. Die Aufgabenkodierung wurde bestimmungsgemäß und den Kodieranweisungen entsprechend durchgeführt. Eine hohe psychometrische Qualität des Testinstruments ist gewährleistet.
Skalierung	Das Testergebnis führt zu Fähigkeitsschätzern, welche sich der mathematischen Kompetenz der Schülerinnen und Schüler annähern.	Das verwendete Skalierungsmodell eignet sich für die Schätzung der mathematischen Fähigkeiten als Annäherung.

Fortsetzung auf der nächsten Seite

Schlussfolgerung	zugrundeliegendes Argument	zugrundeliegende Annahmen
Generalisierbarkeit	Die Fähigkeiten im mittleren Bereich der latenten Skala sind angemessene Schätzer für erwartete Ergebnisse über parallele Messungen.	Die Durchführungsbedingungen der Messung sind standardisiert. Die Messgenauigkeit des Tests ist im mittleren Fähigkeitsbereich angemessen.
Konstruktbezug	Die generalisierten Ergebnisse im NEPS-Mathematiktest für die neunte Klasse können durch eine Reduktion einer komplexeren Kompetenzstruktur als eindimensionales Konstrukt dargestellt werden.	Die eindimensionale Struktur des NEPS-Mathematikkonstrukts eignet sich für die Darstellung der mathematischen Kompetenz.
Extrapolation	Die Kompetenz der Zieldomäne, wie sie mit dem NEPS-Test erfasst wird, ist ein Indikator für die Leistung in der Zieldomäne Mathematische Kompetenz.	Die Testleistung im NEPS-Mathematiktest für die neunte Klasse hängt mit Indikatoren mathematischer Kompetenz in der Zieldomäne zusammen.

Reflexion des IUA und des Validitätsarguments

Das IUA verbindet die Rohwerte des NEPS-K9-Mathematiktests mit den beabsichtigten Interpretationen als mathematische Kompetenz in der Zieldomäne. Die Schlussfolgerungen *Domänenbeschreibung*, *Bewertung*, *Skalierung*, *Generalisierung*, *Konstruktbezug* und *Extrapolation* stellen die einzelnen Interpretationsschritte dar, die für die Validierung notwendig sind. Jede Schlussfolgerung basiert auf einem Argument, welches die Schlussfolgerung begründet. Das Argument wird wiederum durch Annahmen gestützt. Aus den Annahmen lassen sich konkrete Hypothesen ableiten, die im Rahmen der Evaluation der Schlussfolgerung geprüft werden müssen (siehe auch Kapitel 2). Die Validität der Testwertinterpretation ist von dem Validitätsargument abhängig, welches die Evaluation des IUA beinhaltet. Das IUA mit seinen Argumenten, Annahmen und Hypothesen ist die Grundlage für die beabsichtigten Interpretationen und Auswertungen. Es ist daher sehr wichtig, dass neben einer Sammlung von Evidenzen für die Annahmen und Hypothesen des IUA im Validitätsargument auch das IUA und das Validitätsargument selbst reflektiert werden. In diesem Abschnitt sollen daher der Prozess der Konstruktion sowie die Formulierung des IUA anhand von Kriterien nach Kane (2013) sowie die Inhalte des Validitätsargumentes anhand von Kriterien der Standards aus 2014 (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 2014) evaluiert werden.

Reflexion des Entwicklungsprozesse des IUA. Im Folgenden soll die Entwicklung des IUA und des Validitätsarguments für den NEPS-K9-Mathematiktest reflektiert werden. Kane (2001) schlug vier Kriterien vor, anhand derer das IUA entwickelt werden solle. Nach Kane sei nämlich die Entwicklung eines IUA einfach zu beschreiben, jedoch in der Praxis nicht unbedingt einfach durchzuführen.

1. **Deutliche und explizite Beschreibung des IUA.**
   Nach Kane (2001) müsse das IUA zuerst deutlich und präzise beschrieben werden. Oft würden die beabsichtigten Testwertinterpretationen nur sehr allgemein formuliert. Die Interpretationen in der Zieldomäne oder Entscheidungen würden zwar in einigen Fällen deutlich beschrieben, aber die meisten Schlussfolgerungen und Annahmen zwischen den Rohwerten und der Interpretation beziehungsweise Entscheidung in der Zieldomäne würden oftmals nicht explizit definiert.

Auch für den NEPS-K9-Mathematiktest waren viele Schlussfolgerungen und Annahmen nicht explizit formuliert. Die Zieldomäne Mathematische Kompetenz wird im Rahmenkonzept des Tests mit dem Mathematical Literacy Konzept, basierend auf PISA und einer gewissen curricularen Basierung, angelehnt an den LV, beschrieben. Des Weiteren werden im Rahmenkonzept Inhalte und Prozesse beschrieben, die mit dem Test gemessen werden sollen und die Teil der Kompetenz in der Zieldomäne sind. Viele Schlussfolgerungen und Annahmen, welche die Testwerte mit der Zieldomäne verbinden, wurden implizit angenommen beziehungsweise nicht weiter spezifiziert, obwohl einige Untersuchungen zur *Bewertung*, *Skalierung* und *Generalisierung* durchaus durchgeführt wurden. Die beabsichtigten Testwertinterpretationen sollten nach Kane (2001) deutlich und öffentlich gemacht werden, indem ein Netzwerk von Schlussfolgerungen formuliert wird. In diesem würden die Rohwerte den Startpunkt bilden und die Aussagen, Vorhersagen, Entscheidungen etc. zugehörig zur Interpretation und Testnutzung die Schlussfolgerungen. Der Schritt der Ausformulierung des IUA gehöre laut Kane (2001) in die Phase der Testentwicklung und werde im optimalen Fall von den Testentwicklerinnen und Testentwicklern vorgenommen. Für den NEPS-K9-Mathematiktest wurden die Testwertinterpretationen jedoch wie in vielen Tests nicht während der Testentwicklung präzise ausformuliert. Daher wurden die Schlussfolgerungen für den NEPS-K9-Mathematiktest in dieser Arbeit erst nach der Phase der Testentwicklung aus dem Konstrukt des NEPS-Tests logisch abgeleitet und ausformuliert. In einem zweiten Schritt wurden die benötigten Annahmen wiederum aus den Schlussfolgerungen hergeleitet. Dementsprechend wurde zwar der erste Schritt des Prozesses nicht in der nach Kane (2001) definierten Reihenfolge durchgeführt, jedoch wurde eine deutliche und explizite Beschreibung des IUAs in dieser Arbeit erstellt.

2. **Bildung einer vorläufigen Version des Validitätsargumentes.**
   Im Argument Based Approach ist die anfängliche Entwicklung eines Validitätsargumentes eher konfirmierender Natur. Das Ziel bei der Entwicklung des Tests und der Bildung des IUA im ersten Schritt ist es, den Test und das IUA so zu gestalten, dass sie miteinander vereinbar sind. Auf diese Weise wird nicht nur eine logische Grundlage für die beabsichtigten Interpretationen geschaffen, sondern auch vorläufige Evidenz, welche diese Interpretationen stützt (Kane, 2001). Wenn der Test zum Beispiel mathematische Kompetenz erfassen soll, kann eine besonders sorgfältige

Beschreibung der Zieldomäne und eine gut fundierte Entwicklung der Mathematikaufgaben gewährleisten, dass diese die Zieldomäne abdecken und somit als Stützung für die Schlussfolgerung Domänenbeschreibung interpretiert werden können. Außerdem kann in dieser Phase konstruktirrelevanter Varianz vorgebeugt werden (Kane, 2004). Eine angemessene Formulierung der Mathematikaufgaben kann beispielsweise verhindern, dass die Lesefähigkeit die Lösungswahrscheinlichkeit der Aufgaben beeinflusst. Einige der Schlussfolgerungen und Annahmen können auf diese Weise bereits so gut fundiert sein, dass keine weitere Stützung notwendig ist. Andere Schlussfolgerungen und Annahmen können sehr bedenklich sein und eine genaue Prüfung sowie umfangreiche empirische Evidenz benötigen. Zwischen diesen beiden Extremen liegen Schlussfolgerungen und Annahmen, die unterschiedliche Grade und Arten der Evidenz benötigen (Kane, 2004).
Wie bereits beschrieben, wurde bei der Entwicklung des NEPS-K9-Mathematiktests kein IUA entwickelt, welches eine Konsistenz der entwickelten Testung und des IUA hergestellt und begründet hätte. Dementsprechend wurde auch kein vorläufiges Validitätsargument gebildet. Dennoch wurden Studien durchgeführt, die einige Eigenschaften des Tests und der Testprozedur untersuchten und rechtfertigten. Ein Beispiel hierfür ist die Gewichtung der Aufgabenformate. Diese wurden in der Testentwicklungsphase geprüft (siehe Kapitel 4.2). Nach der Bildung des IUA in dieser Studie wurde daher die bereits bestehende Evidenz für das IUA gesammelt und kritisch betrachtet. Einige Annahmen wurden durch die durchgeführten Studien bereits so fundiert, dass keine weiteren Untersuchungen mehr nötig waren. Für andere Schlussfolgerungen bestand nur wenig oder gar keine Evidenz.
Insgesamt wurde zu Beginn des Validierungsprozesses alle bereits bestehende Validitätsevidenz für das IUA gesammelt und so weiterer Untersuchungsbedarf für das IUA aufgedeckt.

3. **Empirische und/oder logische Evaluation der problematischeren Annahmen.**
   Das IUA kann als Folge der Validitätsevidenz abgelehnt oder verbessert werden. Die schwächsten Annahmen brauchen in der Regel die meiste Stützung, haben aber den meisten Informationsgehalt. Dennoch ist es auch sinnvoll, Schlussfolgerungen zu untersuchen, die einfach zu testen sind (Kane, 2004). Gefundene Unstimmigkeiten können korrigiert werden, indem die Interpretation und/oder die Messung angepasst wird.

In dieser Studie wurden für alle Schlussfolgerungen Validitäsevidenzen aus bereits veröffentlichten Untersuchungen gesammelt. Bei unzureichender Evidenz wurden neue Analysen durchgeführt. Dabei wurden die umfangreichsten Untersuchungen für die Schlussfolgerungen (*Konstruktbezug* und *Extrapolation*) durchgeführt.

4. **Neuformulierung des IUA und des Validitätsarguments.**
   Nach der Evaluation der problematischen Annahmen kann die Notwendigkeit entstehen, das IUA neu zu formulieren. Das angepasste IUA muss gegebenenfalls anschließend neu evaluiert werden. Diese Schritte sollen nach Kane (2001) solange wiederholt werden, bis alle Schlussfolgerungen des IUA plausibel sind oder bis das IUA abgelehnt wird.
   Bei der Evaluation der Schlussfolgerungen und Annahmen des NEPS-Mathematiktests wurden einige Unstimmigkeiten aufgedeckt. Da der Test bereits entwickelt und eingesetzt wurde, war es nicht mehr möglich den Test beziehungsweise die Testdurchführung anzupassen. Der Argument Based Approach lässt jedoch Raum für eine Formulierung von Einschränkungen oder Ausnahmen für Interpretationen. Dementsprechend wurden das IUA und das Validitätsarument an die gefundene Evidenz angepasst.

Das IUA für den NEPS-K9-Mathematiktest wurde in einem iterativen Prozess entwickelt. Die von Kane (2001, 2004) vorgegeben Kriterien für den Prozess der Entwicklung konnten dabei nicht vollständig eingehalten werden, da der Ansatz des Argument Based Approach erst nach der Testentwicklung und -durchführung angewendet wurde. Eine Anwendung des Ansatzes während des Testentwicklungsprozesses hätte Abstimmungsbedarf des Tests mit der beabsichtigten Interpretation frühzeitig aufgedeckt und eine Anpassung des Tests in dieser Phase ermöglicht, was zu einer höheren Übereinstimmung der beabsichtigten Testwertinterpretation mit der möglichen Testwertinterpretation geführt hätte. Dennoch wurden alle Kritrien nach Kane (2001, 2004), wenn auch nicht in der vorgeschlagenen Reihenfolge, eingehalten und das IUA wurde an das Validitätsargument angepasst. Damit ist die Entwicklung des IUA als angemessen einzuschätzen.

Evaluation der Formulierung des IUA anhand von Qualitätskriterien. Kane (2012) definierte drei Kriterien für die angemessene Formulierung des IUA. Im Folgenden soll das in dieser Arbeit entwickelte IUA anhand dieser drei Kriterien evaluiert werden.

1. **Deutliche Formulierung des IUA**
   Das Netzwerk der Schlussfolgerungen, Annahmen und Stützungen sollte detailliert genug beschrieben werden, sodass die Grundüberlegungen und Begründungen für die beabsichtigten Interpretationen und Nutzungen deutlich werden. Die explizite Formulierung des IUA hat mehrere Funktionen. Zum einen bietet es einen Leitfaden für die Testentwicklung, da das IUA die Annahmen verdeutlicht, die der Test erfüllen muss. Zum anderen bildet das IUA einen Rahmen für die Bildung des Validitätsarguments (Kane, 2012).
   Bei der Entwicklung des IUA wurde deutlich, dass sich aus den Schlussfolgerungen und Argumenten keine derart präzisen Annahmen herleiten ließen, dass die benötigte Stützung für diese Annahmen direkt logisch und nachvollziehbar abgeleitet werden konnte. Aus diesem Grund wurde ein zusätzlicher Zwischenschritt eingeführt. Aus den Annahmen wurden für den Test konkrete Hypothesen abgeleitet, welche für die Stützung der Annahme gelten müssten. Die Auswertung dieser Hypothesen konnte anschließend logisch hergeleitet werden. Das Netzwerk an Schlussfolgerungen, Annahmen und Stützungen wurde durch das Einfügen von Hypothesen sehr detailliert beschrieben. Die Formulierung des IUA für den NEPS-K9-Mathematiktest kann somit als deutlich und explizit eingeschätzt werden.

2. **Stimmigkeit des IUA**
   Die Begründungen des IUA von den Rohwerten bis zu den Schlussfolgerungen in der Zieldomäne und eventuellen Entscheidungen sollte logisch und nachvollziehbar sein (Kane, 2012).
   Auch die logische Nachvollziehbarkeit wurde durch das Ableiten von konkreten Hypothesen aus den Annahmen sichergestellt. Auf diese Weise wurden die Grundüberlegungen für die Interpretationen, Hypothesen und gewählten Stützungen deutlich gemacht und die Argumentationskette transparent gestaltet.

3. **Plausibilität der Schlussfolgerungen und Annahmen**
   Die Annahmen sollten glaubwürdig und überzeugend sein. Einige Annahmen können dabei vorausgesetzt werden, andere müssen durch sorgfältige Dokumentation und Analysen von Abläufen gestützt werden. Wiederum andere Annahmen benötigen empirische Evidenz (Kane, 2012). So kann beispielsweise bei einer sorgfältigen Standardisierung von Testbedingungen davon ausgegangen werden, dass die Ergebnisse nicht durch unterschiedliche Bedingungen beeinflusst werden. Annahmen über Stichproben können durch eine sorgfältige Dokumentation der Stichproben-

ziehung und eine Analyse dieses Prozesses gestützt werden. Für die Interpretation der Testergebnisse bezüglich der Zieldomäne wird in der Regel empirische Evidenz benötigt.
Um die Schlussfolgerungen und Annahmen plausibel zu gestalten, wurden diese anhand von Literatur zur Testentwicklung und anhand von Literatur zum Argument Based Approach entwickelt. Einige Annahmen wurden direkt aus dem Rahmenkonzept des NEPS-K9-Mathematiktests abgeleitet. Auf diese Weise wurde die theoretische Fundierung der Schlussfolgerungen und Annahmen sichergestellt. Zusätzlich wurden für die Entwicklung der Evidenzen neben der Fachliteratur auch Studien zur Evaluation ähnlicher Annahmen herangezogen. Das IUA für den NEPS-K9-Mathematiktest kann also als plausibel eingeschätzt werden.

Das IUA wurde transparent gestaltet. Alle Hypothesen, Annahmen und Schlussfolgerungen wurden logisch hergeleitet und nachvollziehbar dargestellt (vgl. Kapitel 2). Auf diese Weise wird deutlich, welche Interpretationen mit dem NEPS-K9-Mathematiktest evaluiert wurden und warum. Das Validitätsargument deckt wiederum auf, welche Interpretationen möglich sind und an welchen Stellen Einschränkungen in der Interpretation nötig sind. Eine Zusammenfassung der möglichen Interpretationen und Einschränkungen ist in Tabelle 20 (Seite 181) dargestellt. Somit entspricht die Formulierung des IUA den Qualitätskriterien nach Kane.

Evaluation des IUA und des Validitätsarguments anhand der Standards for Educational and Psychological Testing. Im Folgenden soll für das in dieser Studie entwickelte Validitätsargument geprüft werden, inwiefern es den Standards for Educational and Psychological Testing (Standards) von 2014 (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 2014) entspricht. Die Standards beschreiben Validierung wie folgt: „Validation can be viewed as a process of constructing and evaluating arguments for and against the intended interpretation of testscores and their relevance to the proposed use“ (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 2014, S. 11). Dieser Ansatz basiert, wie bereits in Kapitel 1.1.3 beschrieben, ebenfalls auf dem Bilden und Evaluieren von Argumenten. Die Art der benötigten Validitätsevidenz ist von den beabsichtigten Testwertinterpretationen abhängig. In den Standards von 2014 werden fünf Bereiche vorgeschlagen, in denen Evidenz für die Validität von Testwertinterpretationen gesammelt werden kön-

nen. Nachfolgend wird evaluiert, inwiefern diese Bereiche durch das Validitätsargument des NEPS-K9-Mathematiktests abgedeckt wurden.

1. **Evidenz basiert auf dem Testinhalt**
   Dieser Bereich beinhaltet Untersuchungen zum Zusammenhang des Testinhaltes und des Konstruktes. Diese Untersuchungen können logische oder empirische Analysen zur Angemessenheit der Repräsentation der Zieldomäne durch den Testinhalt oder die Relevanz der Zieldomäne für die Testwertinterpretation sein. Diese Analysen können unter anderem Expertenreviews zur Passung des Tests zum Konstrukt beinhalten (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 2014). Solche Untersuchungen wurden in der Schlussfolgerung *Domänenbeschreibung* (vgl. Kapitel 4.1) durchgeführt. Dort wurde die Repräsentation der Zieldomäne durch die Untersuchung der Passung der NEPS-Mathematikaufgaben mit den Rahmenkonzepten aus dem LV und aus PISA untersucht, da diese Rahmenkonzepte Teile der Zieldomäne darstellen. Es konnte eine angemessene Passung gefunden werden.

2. **Evidenz basiert auf Antwortprozessen**
   Die Art der Evidenz beinhaltet empirische und theoretische Untersuchungen der Passung zwischen dem Konstrukt und der tatsächlich geforderten Leistung beziehungsweise der geforderten Antwortprozesse für die Bearbeitung der Aufgaben. Evidenzen können zum Beispiel durch die Analyse der individuellen Antworten gesammelt werden, indem Testpersonen zu ihren Antworten befragt werden, indem die Entstehung der Antwort beobachtet und dokumentiert wird oder indem Zusammenhänge des Tests beziehungsweise von Testteilen mit anderen Variablen untersucht werden. Auch eine Analyse der Aufgaben durch Expertinnen und Experten ist möglich. So kann die Einschätzung der benötigten Prozesse durch die Expertinnen und Experten mit der beabsichtigten Interpretation abgeglichen werden. Eine solche Analyse durch Expertinnen und Experten wurde in der Schlussfolgerung *Domänenbeschreibung* (vgl. Kapitel 4.1) durchgeführt. Die Expertinnen und Experten ordneten die Items den kognitiven Prozessen aus PISA und dem LV zu, um zu untersuchen, ob die NEPS-Aufgaben die Prozesse der Zieldomäne abdecken. Die Ergebnisse zeigten, dass die NEPS-Aufgaben die Prozesse aus PISA und dem LV zwar größtenteils abdecken, die Zieldomäne für den NEPS-Test jedoch enger definiert werden muss, da aktive Komponenten der Prozesse nicht durch die

Aufgaben erfasst werden.

3. **Evidenz basiert auf der internen Struktur**
   Untersuchungen der internen Struktur eines Tests sollen aufzeigen, inwiefern die Zusammenhänge zwischen Testaufgaben und Testkomponenten das Konstrukt bestätigen, auf welchem die Testwertinterpretationen basieren. Dabei hängen die durchzuführenden Analysen von der beabsichtigten Testinterpretation ab. Untersuchungen zur internen Struktur des Tests wurden in dieser Studie in der Schlussfolgerung *Konstruktbezug* abgedeckt (vgl. Kapitel 4.5). Hier wurden die Dimensionalität des NEPS-K9-Mathematiktests und die Zusammenhänge der Mathematikaufgaben mit anderen Testdomänen aus dem NEPS untersucht. Insgesamt ließ sich der NEPS-K9-Mathematiktest von den anderen im NEPS verwendeten Kompetenztests abgrenzen. Die Eindimensionalität ließ sich jedoch nicht bestätigen. Für die Skalierung der Kompetenzdaten als Reduktion eines komplexeren Konstruktes ist die Eindimensionalität jedoch angemessen. Nach den Standards von 2014 gehört auch die Untersuchung von DIF zu dieser Kategorie. Im IUA für den NEPS-Test wird DIF in der Schlussfolgerung *Skalierung* (vgl. Kapitel 4.3) untersucht. Die einzelnen Mathematikaufgaben zeigten keine DIF-Effekte, jedoch wurden bessere Modellgütekriterien für die Variablen Geschlecht und Schulform für die Modelle gefunden, die DIF zulassen.

4. **Evidenz basiert auf Beziehungen zu anderen Variablen**
   In vielen Fällen impliziert die beabsichtigte Interpretation eines Tests Zusammenhänge mit anderen Variablen. Dementsprechend bilden Analysen der Testwerte mit externen Variablen Validitätsevidenz. Analysen aus dieser Kategorie wurden in der Schlussfolgerung *Extrapolation* evaluiert (vgl. Kapitel 4.6). Dort wurden die konvergenten und divergenten Zusammenhänge der NEPS-Mathematiktestwerte mit Werten anderer Tests und Kriterien untersucht. So wurden stärkere Zusammenhänge der NEPS-Mathematiktestwerte mit den Mathematiknoten gefunden als mit Noten anderer Fächer. Außerdem wurden starke Zusammenhänge mit den mathematischen Kompetenzmessungen aus PISA und dem LV gefunden. Des Weiteren konnten erwartungsgemäße Zusammenhänge mit Messungen der kognitiven und metakongnitiven Fähigkeit gezeigt werden.

5. **Evidenz basiert auf Konsequenzen der Testung**
   Einige Testwertinterpretationen haben direkte Konsequenzen für die Testteilneh-

merinnen und Testteilnehmer zur Folge. In diesem Fall sollte der Validierungsprozess die Sammlung von Evidenz bezüglich der Testwertinterpretation für diese Nutzung beinhalten. Andere Konsequenzen können durch Testwertinterpretationen entstehen, die über die ursprünglichen Interpretationen und Nutzungen der Testentwicklerinnen und Testentwickler hinaus gehen. Auch können Konsequenzen folgen, die unbeabsichtigt sind. In diesem IUA wurde jedoch nur eine deskriptive Interpretation der Testwerte validiert. Es werden von den Testentwicklerinnen und Testentwicklern keine darüber hinausgehenden Testnutzungen für den NEPS-K9-Mathematiktest formuliert. Die Daten der Testungen in den NEPS-Haupterhebungen, für die der NEPS-K9-Mathematiktest eingesetzt wird, werden der Wissenschaft in Form von *scientific use files* zur Verfügung gestellt. Die Testnutzerinnen und Testnutzer verwenden also nur die Testwerte und setzen den Test nicht selbst ein. Dementsprechend ist auch für Testnutzerinnen und Testnutzer nur eine deskriptive Interpretation von Testwerten relevant, das heißt es werden keine Entscheidungsregeln angewendet.

Auch steht in den Standards von 2014 geschrieben, dass ein solides Validitätsargument diverse Validitätsbeweise in eine schlüssige Darstellung des Grades integriert, indem die Beweise und Theorien die beabsichtigte Testwertinterpretation stützen. Es umfasst nach den Standards 2014 Evidenz von neuen und bereits berichteten Untersuchungen und kann die Notwendigkeit für eine Überarbeitung der Konstruktdefinition, des Tests oder Aspekten der Messung sowie für weitergehende Untersuchungen aufzeigen.
Auch das hier formulierte IUA integriert Evidenzen für die Validität und Einschränkungen der Testwertinterpretationen. Ebenso wie in den Standards gefordert, werden bereits bestehende Untersuchungen genutzt sowie neue Analysen durchgeführt. Neben der Notwendigkeit für weitere Untersuchungen und eventuelle Hinweise für eine Überarbeitung der Konstruktdefinition (vgl. Kapitel 5.1.3) wird jedoch auch die Testwertinterpretation und somit das IUA durch die Validierung angepasst (vgl. Kapitel 5.1.2). Insgesamt entspricht die Validierung der deskriptiven Testwertinterpretation des NEPS-K9-Mathematiktests den Standards for Educational and Psychological Testing aus 2014.

5.1.3 Implikationen für Testnutzerinnen und Testnutzer

In dieser Arbeit konnte ein Validitätsargument für den NEPS-K9-Mathematiktest gebildet werden. Dabei wurden einige Einschränkungen von Testwertinterpretationen aufgezeigt. Einige Testwertinterpretationen konnten in dieser Arbeit nicht gestützt werden, da noch weiterer Forschungsbedarf besteht (vgl. Tabelle 20). Einige Hypothesen, wie zum Beispiel der prädiktive Zusammenhang der Testergebnisse mit zukünftigen Kriterien wie Schulerfolg etc., konnten nicht aufgestellt und getestet werden, da die Datengrundlage für eine solche Auswertung nicht zur Verfügung stand. Aus dieser Studie ergab sich der Bedarf für weitere Auswertungen und zusätzliche Stützung. Aus diesem Grund können Empfehlungen für weitere Untersuchungen bezüglich einiger Schlussfolgerungen des IUA formuliert werden, um diese zu stärken oder zu erweitern. Auch können Empfehlungen für Untersuchungen abgeleitet werden, die eine Erweiterung der Testwertinterpretation für den NEPS-K9-Mathematiktest ermöglichen.

Implikationen für die Stärkung des IUA

Die Schlussfolgerung *Domänenbeschreibung* könnte zusätzlich gestärkt werden, indem die Operationalisierung der NEPS-Prozesse mit den Prozessen aus dem PISA- und dem LV-Rahmenkonzept verglichen wird. Dafür müssen zuerst die NEPS-Mathematikaufgaben den Prozessen aus dem NEPS-Rahmenkonzept zugeordnet werden. Des Weiteren würde ein Vergleich der Gewichtungen der Prozesse aus dem LV und aus PISA sowie der Kontexte aus PISA und der Anforderungsbereiche aus dem LV mit dem NEPS-Mathematiktest zusätzliche Stützung für die Schlussfolgerung *Domänenbeschreibung* bieten. Diese Analysen basieren auf Expertenreviews der Aufgaben. Zusätzlich wäre eine Analyse von individuellen Antwortprozessen denkbar. Eine solche Analyse, beispielsweise durch die Methode des lauten Denkens, würde es ermöglichen, zu erkennen, ob die kognitiven Prozesse von den Schülerinnen und Schülern tatsächlich zur Bearbeitung der Aufgaben benötigt werden. Auch kompensatorische Prozesse könnten aufgedeckt werden. So würde beispielsweise sichtbar, wenn Schülerinnen und Schüler in einem kognitiven Prozess, der eigentlich zum Lösen der Aufgabe benötigt wird, niedrige Fähigkeiten aufweisen, sich die Lösung der Aufgabe aber durch hohe Fähigkeiten in einem anderen Prozess erschließen können.
Die Schlussfolgerung *Bewertung* könnte weiter gestärkt werden, indem die eingegebe-

nen Daten auf ihre Integrität durch beispielsweise ein Datencleaning geprüft würden. Beziehungsweise es ist davon auszugehen, dass eine solche Prüfung stattgefunden hat. Daher wäre es sehr wünschenswert, wenn die Auswertungen zur Integritätsprüfung veröffentlicht werden würden. Dasselbe gilt für die Prüfung des Scorings der Daten. Auch hier wäre es wünschenswert, wenn die Prozesse und Ergebnisse zur Qualitätsprüfung des Scorings veröffentlicht würden.

Weitere Untersuchungen zu der möglichen Ursache von DIF würden die Schlussfolgerung *Skalierung* weiter untermauern. Dies könnte auch durch die Untersuchung von individuellen Antwortprozessen geschehen. So könnte untersucht werden, ob Mädchen oder Jungen beziehungsweise Schülerinnen und Schüler von Gymnasien oder anderen Schulformen Schwierigkeiten beim Lösen einer Aufgabe zeigen, die unabhängig von der mathematischen Fähigkeit ist. Auch könnten die Mathematikaufgaben von Expertinnen und Experten auf eventuelle außermathematische Schwierigkeiten beziehungsweise Benachteiligungen für diese Subgruppen untersucht werden.

Die Schlussfolgerung *Generalisierbarkeit* könnte durch weitere Analysen zur Standardisierung der Testinstruktion und der Testumgebung sowie Analysen bezüglich eines Qualitätsmonitorings der Testdurchführung und möglicher Störungen gestützt werden. Wie bei der Schlussfolgerung *Bewertung* ist auch hier anzunehmen, dass eine solche Standardisierung und ein Qualitätsmonitoring stattgefunden haben. Eine Veröffentlichung solcher Prozesse würde die Schlussfolgerung transparenter machen und stärken.

Die dimensionale Struktur des Konstruktes mathematischer Kompetenz im NEPS konnte nicht bestätigt werden. Analysen zur tatsächlichen Dimensionalität des Konstruktes durch zum Beispiel explorative Faktorenanalysen sind zwar für die Fragestellung in dieser Arbeit nicht notwendig, würden die Schlussfolgerung jedoch für zukünftige Nutzungen der Testdaten transparenter machen. Es würde deutlicher werden, wie sich das eindimensionale Modell von der empirischen, dimensionalen Struktur unterscheidet und mögliche Konsequenzen für das Ignorieren dieser Struktur und der Interpretation der Testwerte als eindimensionales Konstrukt verdeutlichen.

Die Schlussfolgerung *Extrapolation* könnte erweitert werden, indem die Zusammenhänge der NEPS-Mathematikkompetenz mit direkten Kriterien für die Zieldomäne untersucht würden. Kriterien, die die mathematische Kompetenz im Alltag widerspiegeln und für eine solche Analyse geeignet wären, sind der zukünftige Schulabschluss, der zukünftige Ausbildungserfolg, der zukünftig erreichte sozioökonomische Status, etc.. Auch wäre es möglich, den Zusammenhang mit Ergebnissen aus praktischen Prüfungen zu analysieren, welche die mathematische Kompetenz im Alltag erfassen. Während die auf Stift

und Papier basierten Tests Kontexte durch Beschreibungen in den Aufgabenstimuli zu schaffen versuchen, könnten Schülerinnen und Schüler in realen oder annähernd realen Kontexten beobachtet werden. Die hier genannten Analysen ergäben sich aus dem Validitätsargument und würden die beabsichtigte Testwertinterpretationen zusätzlich fundieren beziehungsweise erweitern.
Es ist zudem denkbar, die in dieser Studie aufgedeckten Einschränkungen der Interpretation durch die Anpassung des Tests und des Testkonstruktes aufzuheben. So wäre es möglich, die Testaufgaben so anzupassen, dass sie die Teilkompetenzen von PISA und dem LV besser abdecken. Auch könnte der Test so angepasst werden, dass dieser keinen DIF für die Subgruppen Geschlecht und Schulform aufweist, dass die Aufgaben die gleichen Trennschärfen im Sinne von Steigungsparametern aufweisen und dass der Test eindimensional ist. Das Rahmenkonzept für den NEPS-K9-Mathematiktest könnte aber auch so verändert werden, dass das mathematische Konstrukt als mehrdimensional beschrieben wird. Da es sich um eine längsschnittliche Untersuchung handelt, in der die mathematischen Testungen zwischen den Kohorten und innerhalb der Kohorte vergleichbar sein sollen, wäre eine Veränderung des Tests und des Testkonstruktes im Laufe der Studie jedoch nicht zielgerichtet.
Eine kritische Reflexion der Schlussfolgerungen, Argumente, Annahmen und Hypothesen des in dieser Arbeit gebildeten IUAs ist sehr wünschenswert. So plädierte Kane (2013) dafür, dass das IUA im Anschluss an die Entwicklung und erste Auswertung extern auf mögliche alternative Interpretationen und versteckte Annahmen untersucht werden solle. Eine solche Analyse könne Schwächen in der Argumentationskette aufdecken und mögliche rivalisierende Hypothesen für die Interpretation der Testwerte aufstellen. Vor der Nutzung neuer Testwertinterpretationen ist es daher zu empfehlen, dass IUA auf diese Weise kritisch zu prüfen und gegebenenfalls Schlussfolgerungen, Argumente, Annahmen oder Hypothesen anzupassen und diese neu zu evaluieren.

Implikationen für die Erweiterung des IUA

Im Falle des NEPS-K9-Mathematiktests ist es nicht vorgesehen, dass dieser Test außerhalb der Studie genutzt wird. Die Testnutzung, inklusive der Umstände der Messung, der Zielgruppe, des Testeinsatzes etc. wird voraussichtlich innerhalb der längsschnittlichen NEPS-Studie konstant bleiben. Die Daten werden der Wissenschaft in Form von *scientific use files* zur Verfügung gestellt. Daher ist es wichtig, dass die Testwertinterpre-

tationen, die mit diesen Daten getätigt werden, valide sind. Laut den Standards aus 2014 hätten die Testentwicklerinnen und Testentwickler sowie Testnutzerinnen und Testnutzer eine gemeinsame Verantwortung für die Validierung der Testwertinterpretationen. Die Testentwicklerinnen und Testentwickler seien für die Evidenz und die Begründungen verantwortlich, die für die von ihnen vorgesehenen Testwertinterpretationen relevant sind. Die Testnutzerinnen uns Testnutzer seien verantwortlich für die Evaluation der Evidenz in dem Rahmen, in welchem der Test beziehungsweise die Testwerte von ihnen genutzt werden soll beziehungsweise sollen. Werde eine Interpretation der Testwerte beabsichtigt, die sich von der vorgesehenen Interpretation der Testentwicklerinnen und Testentwickler unterscheidet, so liege die Verantwortung der Validierung dieser Interpretation bei den Testnutzerinnen und Testnutzern (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 2014, S.13). Im Falle des NEPS-K9-Mathematiktests wurde zwar keine konkrete Testwertinterpretation von den Testentwicklerinnen und Testentwicklern vorgegeben, jedoch wurde in dieser Studie die Testwertinterpretation „Die Unterschiede in der Rangfolge der beobachteten Testleistung deuten auf Unterschiede in der Rangfolge der Ausprägung der mathematischen Kompetenz im Alltag hin“ validiert. Sollten Testnutzerinnen und Testnutzer andere Interpretationen beabsichtigen, wie beispielsweise „Die Unterschiede in der beobachteten Testleistung geben die Unterschiede in der Fähigkeit wieder, mathematische Probleme in Realsituationen lösen zu können“, so müssten diese Interpretationen zuvor validiert werden.

Durch die derzeitige Validierung können die Daten aus dem *scientific use file* für deskriptive Interpretationen, das heißt ohne die Nutzung der Ergebnisse für Entscheidungen, verwendet werden. So können Zusammenhänge zwischen der mathematischen Kompetenz mit anderen Variablen untersucht werden, wie beispielsweise die Zusammenhänge von Interesse oder Motivation mit der mathematischen Fähigkeit im NEPS-Mathematiktest.

Die NEPS-K9-Testergebnisse eignen sich außerdem für eine Interpretation im Rahmen der Kompetenzstufen aus dem LV und aus PISA auf Populationsebene. Die Grundlage hierfür bildet ein Linking mit den Kompetenzskalen aus dem LV und aus PISA. Die großen konzeptionellen Übereinstimmungen, die zwischen den Studien gefunden wurden (vgl. Kapitel 1.2.5) bilden eine erste Voraussetzung für ein Linking (van den Ham et al., 2014). Im Rahmen einer Validierungsstudie wurde ein Linking des dort eingesetzten NEPS-Mathematiktests mit den in der Validierungsstudie eingesetzten Mathematikauf-

gaben aus dem LV und aus PISA mit Hilfe des Equipercentile Equating durchgeführt. Die Ergebnisse zeigten, dass die Linkingwerte, die auf dem NEPS-Test basieren und anhand derer die Linkingfunktion auf die LV- bzw. PISA-Metrik gebracht wurde, zu annähernd gleichen Verteilungen auf den Kompetenzstufen führen wie die LV- bzw. PISA-Werte aus der Validierungsstudie. Außerdem führte das Linking zu stabilen Ergebnissen über Subgruppen (van den Ham, Ehmke, Roppelt & Stanat, angenommen; Ehmke, van den Ham, Sälzer & Heine, in Vorbereitung). Für die Übertragung der in der Linkingstudie berechneten Linkingfunktionen auf die Stichprobe der Haupterhebung aus NEPS-K9 wäre jedoch zu prüfen, inwiefern durch die Unterschiede der Validierungs- und Haupterhebungsstichprobe (vgl. Anhang 5.2.2) zusätzliche Linkingfehler entstehen. Die großen Zusammenhänge zwischen dem NEPS-K9-Mathematiktest und den Mathematiktests aus dem LV und aus PISA sowie die Stabilität über Subgruppen weisen darauf hin, dass eine vorsichtige Interpretation der Verteilung der NEPS-K9-Testergebnisse auf den Kompetenzstufen aus dem LV und aus PISA auf Populationsebene gerechtfertigt ist. Durch eine solche Nutzung des Linkings würde eine Erweiterung der NEPS-Testwertinterpretation möglich und es könnten beispielsweise in der neunten Klassenstufe die Verteilungen der unterschiedlichen Kohorten auf die LV- und PISA-Kompetenzstufen untersucht sowie Trends analysiert werden.

Ein Ziel der NEPS-Studie ist die längsschnittliche Untersuchung der Kompetenzentwicklung im Lebenslauf (Blossfeld & von Maurice, 2011). Die Daten aus dem NEPS-K9-Mathematiktest an sich bieten jedoch nur querschnittliche Informationen. Durch das Multi-Kohorten-Sequenz-Design sollen Testwertinterpretationen zwischen unterschiedlichen Zeitpunkten und unterschiedlichen Kohorten verbunden werden. So werden in unterschiedlichen Kohorten, jedoch zu unterschiedlichen Zeitpunkten, gleiche Tests eingesetzt. Beispielsweise wurde der NEPS-Mathematiktest in der Startkohorte 4 im ersten Erhebungszeitraum im Jahr 2010 eingesetzt. Dieser Test wurde laut Erhebungsplan im Jahr 2014 nochmals in der Startkohorte 3 eingesetzt. Die Schülerinnen und Schüler der Startkohorte 3 befanden sich im Jahr 2014 in der neunten Klassenstufe. In beiden Startkohorten sollten die gleichen Testwertinterpretationen möglich sein.
Ob der NEPS-Test über die verschiedenen Kohorten hinweg generalisierbar ist, kann durch eine Erweiterung der Schlussfolgerung *Generalisierung* untersucht werden. Das Argument, dass die Fähigkeiten auf der latenten Skala angemessene Schätzer für erwartete Ergebnisse über parallele Messungen sind, muss nicht nur innerhalb der Kohorte, sondern auch über die Kohorten hinweg, gelten. Dementsprechend muss angenommen

werden, dass die Durchführungsbedingungen über die Messungen in den Kohorten hinweg standardisiert sind und dass die Fähigkeitsmessungen über die Kohorten hinweg reliabel sind. Die Testergebnisse sollten nicht durch Unterschiede in Umständen der Kohortentestungen beeinflusst werden.
Innerhalb der Kohorten werden die Kompetenzen der Schülerinnen und Schüler zu unterschiedlichen Zeitpunkten mit unterschiedlichen Tests gemessen. So wurden die mathematischen Kompetenzen in der Startkohorte 4, welche sich zum Zeitpunkt der Haupterhebung im Jahr 2010 in der neunten Klassenstufe befand, zu diesem Zeitpunkt mit dem K9-Mathematiktest erfasst. Laut Erhebungsplan sollten die Schülerinnen und Schüler dieser Startkohorte im Jahr 2014 erneut mit einem Mathematiktest getestet werden. Für die längsschnittliche Nutzung der Testwerte müssen die beabsichtigten längsschnittlichen Testwertinterpretationen validiert werden. Folgende zu evaluierende Testwertinterpretation wäre denkbar: „Unterschiede in der beobachteten Testleistung über die Zeit (beispielsweise in den Mathematiktests aus den Jahren 2010 und 2014) spiegeln die Entwicklung in der Zieldomäne Mathematische Kompetenz wider.“. Eine Voraussetzung für diese Testwertinterpretation ist, dass die Unterschiede in der beobachteten Testleistung in den (in den Jahren 2010 und 2014) verwendeten Tests die Unterschiede in der Zieldomäne Mathematische Kompetenz widerspiegeln. Für beide Tests muss also die Schlussfolgerung für letztere Interpretationen durchlaufen werden (siehe Abbildung 37). Für die erweiterte, längsschnittliche Testwertinterpretation müssen zusätzliche Annahmen getestet werden, die eine Brücke zwischen den beiden IUAs der Tests schlagen und eine Verbindung der Interpretationen herstellen. So gilt zum einen für die Schlussfolgerung *Skalierung*, dass die Fähigkeitsschätzer beider Tests so auf einer Metrik skaliert werden müssen, dass korrekte Schätzungen des Fähigkeitszuwachses möglich werden. Zum anderen wird im Rahmenkonzept beschrieben, wie sich die mathematische Kompetenz über die Zeit entwickelt und was die Inhaltsbereiche und Prozesse für die verschiedenen Altersstufen umfassen. Für die Schlussfolgerung *Konstruktbezug* gilt daher, dass sich die angenommene Zeit-Dimension des Konstruktes bestätigen lassen muss.

Für die Nutzerinnen und Nutzer des *scientiftic use files*, welche eine längsschnittliche Testwertinterpretation oder die Nutzung der NEPS-Daten des NEPS-K9-Mathematiktests aus einer anderen Startkohorte als der Startkohorte 3 beabsichtigen, gilt, dass diese das hier entwickelte IUA erweitern und überprüfen müssen.

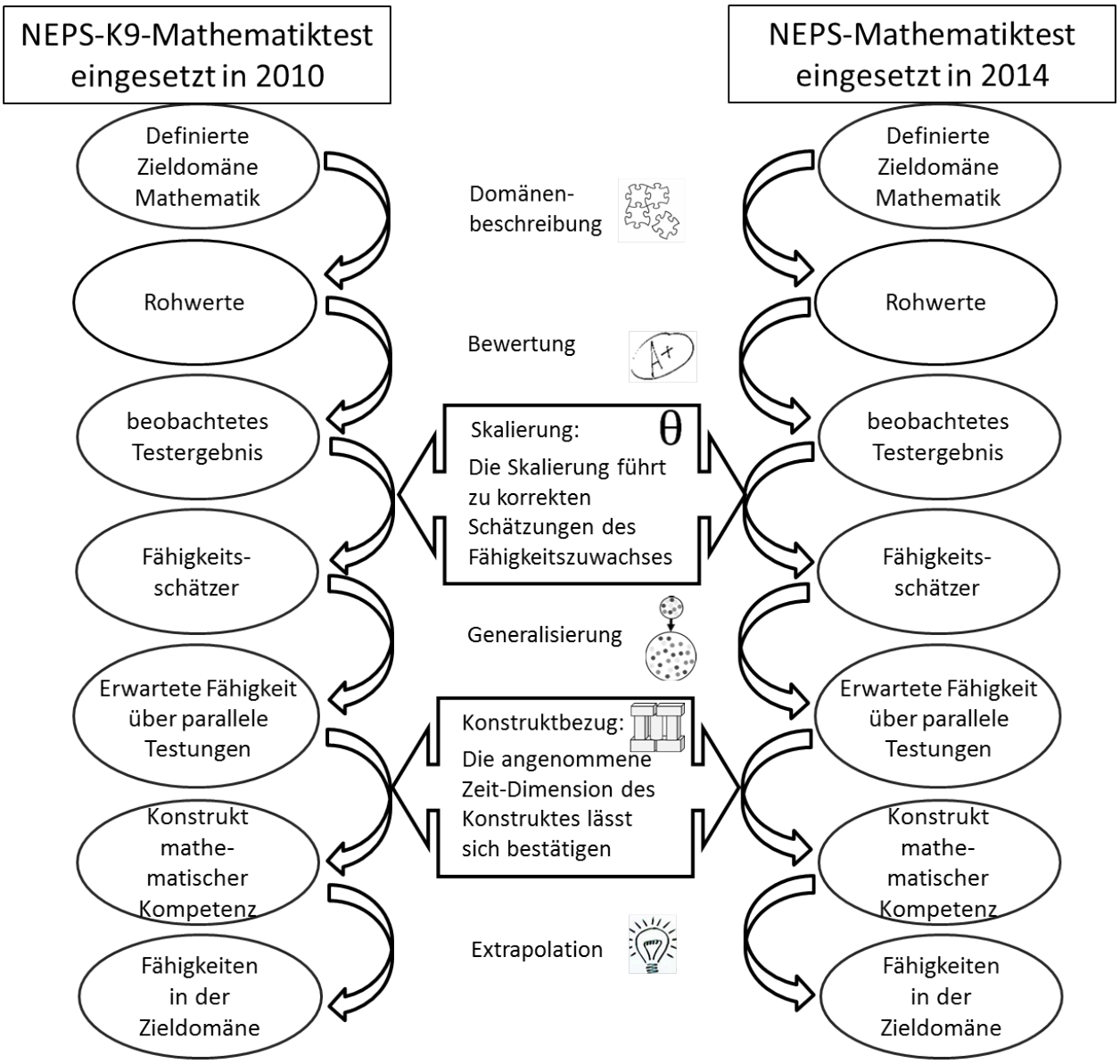


Abbildung 37: Beispiel für die Erweiterung des IUA für die längsschnittliche Interpretation der Testwerte des NEPS-K9-Mathematiktests

5.2 Der Argument Based Approach als Ansatz zur Validierung von Testwertinterpretationen

In dieser Arbeit wurde der Argument Based Approach nach Kane für die Validierung des NEPS-K9-Mathematiktests eingesetzt. Dieser Ansatz wurde aufgrund seiner Vortei-

le gegenüber anderen Validierungsansätzen, wie der „Konstruktvalidität“ nach Messick (1989a, 1994, 1995), dem „Evidence Centered Assessment Design“ von Mislevy (2007); Mislevy et al. (2003), dem „Assessment Use Argument“ von Bachman (2005) und dem Validierungsansatz der Standards (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 2014), verwendet. Jedoch können für den Argument Based Approach auch einige kritische Aspekte aufgedeckt werden. Im Folgenden soll daher der Argument Based Approach reflektiert werden, indem zuerst die Gemeinsamkeiten des Ansatzes mit anderen modernen Validierungsansätzen aufgezeigt werden, die Vorteile des Argument Based Approach gegenüber diesen Ansätzen hervorgehoben werden und die kritischen Aspekte des Ansatze dargestellt werden. Anschließend sollen aus diesen Darlegungen Implikationen für zukünftige Validierungen abgeleitet werden.

5.2.1 Eine kritische Betrachtung des Ansatzes

Es gibt viele verschiedene Validierungsansätze, basierend auf dem modernen Konzept der Konstruktvalidität, jedoch hat sich bisher keiner dieser Ansätze konstant durchgesetzt. Zu diesen Ansätzen gehören neben dem Argument Based Approach von Kane (2013), die „Konstruktvalidität“ nach Messick (1989a, 1994, 1995), das „Evidence Centered Assessment Design“ von Mislevy et al. (2003) und (Mislevy, 2007), das „Assessment Use Argument“ von Bachman (2005) und der Validierungsansatz der Standards (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 2014). Eine Beschreibung dieser Ansätze befindet sich in Kapitel 1.1.2 und 1.1.3. Der Argument Based Approach nach Kane (2013) deckt die wichtigsten Aspekte dieser Ansätze ab. Die sechs Aspekte der Konstruktvalidität nach Messick (1989a, 1994, 1995) lassen sich zum Beispiel in den Schlussfolgerungen von Kane wiederfinden. So können die Aspekte Inhalt und Theorie sowie Prozessmodelle in die Schlussfolgerung *Domänenbeschreibung* und die externen Aspekte in die Schlussfolgerung *Extrapolation* eingeordnet werden. Der Aspekt Struktur kann der Schlussfolgerung *Konstruktbezug*, der Aspekt Generalisierbarkeit der Schlussfolgerung *Generalisierbarkeit* und der Aspekt Konsequenzen der Schlussfolgerung *Extrapolation* zugeordnet werden. Dabei lässt sich im Ansatz der Konstruktvalidität nach Messick kein expliziter Aspekt wiederfinden, der sich mit der Schlussfolgerung *Bewertung* gleichsetzen lässt. In der Matrix der Konstruktvalidität wird im Vergleich zum Argument Based Approach

stärker auf die Testkonsequenzen und Testnutzung eingegangen. Die Analyse der Relevanz und Nützlichkeit des Tests sowie die Wertimplikationen werden explizit in den Validierungsprozess aufgenommen. Dabei verändert sich Messicks Beschreibung darüber, welche Art von Konsequenzen für die Validierung notwendig ist, in seinen Publikationen und dieser Aspekt von Messicks Validitätsansatz bleibt bis heute unklar (Newton & Shaw, 2014).

In den Standards der American Educational Research Association, American Psychological Association, National Council on Measurement in Education (2014) gibt es fünf Arten der Evidenz, welche die Validität der Testwertinterpretationen stützen können (Evidenz basierend auf Testinhalten, auf Antwortprozessen, auf der internen Struktur, auf Beziehungen zu anderen Variablen und auf Testkonsequenzen, vgl. Kapitel 1.1.3). Alle diese Arten der Evidenz können unter Anwendung des Arguments Based Approach gesammelt werden (vgl. Kapitel 5.1.2). Auch stimmt die Beschreibung der Validität in den Standards als einheitliches Konzept und argumentativer Prozess größtenteils überein (vgl. Kapitel 1.1.3). Ein Unterschied zwischen den beiden Validitätsansätzen ist, dass in den Standards das Konstrukt als Basis für die Validierung definiert ist, wohingegen das IUA im Argument Based Approach die Basis für die Validierung darstellt.

Auch das "Evidence Centered Assessment Design" von Mislevy et al. (2003) und Mislevy (2007) hat viele Gemeinsamkeiten mit dem Argument Based Approach von Kane. Die Basis für die Entwicklung eines Tests ist in dem Modell von Mislevy et al. (2003) und Mislevy (2007) die Entwicklung des Assessment Designs und Use Arguments. Anders als im Argument Based Approach wird die Validierung nicht nach den Interpretationen, die getätigt werden sollen, strukturiert, sondern nach dem Prozess der Testentwicklung. Dadurch ist der Aufbau der Assessment und Use Argumente durch Mislevy nur grob vorstrukturiert und wird erst durch den Entwicklungsprozess des Tests und die beabsichtigten Testwertinterpretationen konkreter. Die Schlussfolgerungen des Argument Based Approach können prinzipiell in das Assessment Design und Use Argument integriert werden. Deutlich werden jedoch nur die Schlussfolgerungen *Bewertung*, *Konstruktbezug* und *Extrapolation* in den Argumenten aufgezeigt.

Das Assessment Use Argument von Bachman (2005) ist dem Argument Based Approach ebenfalls in vielen Punkten ähnlich. So lassen sich die Schlussfolgerung *Konstruktbezug* im dem Qualitätsaspekt Konstruktvalidität, die Schlussfolgerung *Generalisierung* in den Qualitätsaspekten Konstruktvalidität und Reliabilität, die Schlussfolgerungen *Domänenbeschreibung* und *Bewertung* in dem Aspekt Interaktivität, die Schlussfolgerung *Extrapolation* in dem Aspekt Authentizität und die Schlussfolgerung *Extrapolation*

in dem Aspekt Einfluss wiederfinden. Das Assessment Use Argument ist ebenfalls auf Basis der beabsichtigten Interpretationen strukturiert und das Validitätsargument ist vom Utilization Argument getrennt. Jedoch werden die Qualitätsaspekte nicht in Form von Argumenten formuliert und das Validitätsargument ist nicht genau ausgearbeitet. Das Assessment Use Argument geht vor allem auf die Konsequenzen beziehungsweise Entscheidungen bezüglich der Testnutzung ein, für die das Utilization Argument genau ausdifferenziert ist. Mit vier konkreten Typen von Konsequenzen setzt das Assessment Use Argument mehr Evidenzen für die Entscheidung voraus als der Argument Based Approach.

Aus diesem Vergleich wird ersichtlich, dass der Argument Based Approach wichtige Aspekte der anderen Validitäsansätze beinhaltet. Es wird auch deutlich, dass der Argument Based Approach eine deutlichere und transparentere Struktur als die übrigen hier vorgestellten Ansätze hat. Der Argument Based Approach basiert auf dem IUA, also auf den beabsichtigten Testwertinterpretationen und den dazugehörigen Argumenten. Dies bietet gegenüber den Ansätzen, die auf dem Testkonstrukt beruhen, viele Vorteile. Zum einen ist für die Validierung kein präzise ausdifferenziertes Konstrukt notwendig. Zum anderen ist es lediglich erforderlich, die Schlussfolgerungen und Argumente zu evaluieren, die für die beabsichtigte Testwertinterpretation relevant sind. Wenn dem Test kein Konstrukt unterliegt und ein solches dementsprechend nicht in die Testwertinterpretation eingebunden ist, so braucht dieses auch nicht evaluiert zu werden. Die Aufgliederungen aller Interpretationen in aufeinanderfolgende Schlussfolgerungen macht den Validierungsprozess logisch nachvollziehbar. Auch wird für Außenstehende deutlich, welche Interpretationen in den Validierungsprozess eingeschlossen sind und welche nicht, auf welcher Stützung die Interpretationen basieren und eventuell auch welche Ausnahmen gelten.

Trotz der vielen Vorteile des Argument Based Approachs können auch einige kritische Aspekte des Ansatzes benannt werden, die während des Validierungsprozesses in dieser Arbeit sowie in anderen Studien sichtbar wurden. Ein Mangel des Ansatzes von Kane ist beispielsweise, dass gewisse Bestandteile beziehungsweise Merkmale nicht einfach zu verstehen und anzuwenden sind. So ist die Art der Unterscheidung zwischen dem IUA und dem Validitätsargument nicht ganz deutlich (Newton, 2013). Im Argument Based Approach wird zuerst ein IUA gebildet. Anschließend wird dieses IUA ausgewertet. Diese Auswertung des IUA wird als Validity Argument bezeichnet. Zusätzlich werden die Entwicklungsphase und die Beurteilungsphase unterschieden. In der Entwicklungsphase

wird ein Test entwickelt, welcher eine bestimme Testinterpretation erlauben soll. Wenn sich während der Testentwicklung herausstellt, dass bestimmte Annahmen nicht durch den Test abgedeckt werden können, wird entweder die Testinterpretation oder der Test selbst modifiziert. Das Ergebnis der Testentwicklungsphase ist der entwickelte Test sowie eine klare Aussage über die angestrebten Testwertinterpretationen in Form eines ausformulierten IUA. In der Begutachtungsphase folgt eine kritische und unabhängige Evaluation der beabsichtigten Testwertinterpretationen. Das entwickelte IUA sollte in dieser Phase in Frage gestellt werden. Dabei werden versteckte Annahmen oder alternative Interpretationen der Testergebnisse aufgedeckt und die Annahmen des IUA kritisch geprüft. Diese Phase beinhaltet in den meisten Fällen empirische Auswertungen der bedenklichsten Annahmen (siehe auch Kapitel 5.1.2). Obwohl schon in der Entwicklungsphase Evidenz für das IUA gesammelt wird, spricht Kane hier noch nicht von einem Validity Argument. Außerdem schreibt er, dass die beiden Argumente in der Praxis miteinander verwoben sind (Kane, 2013). Undeutlich ist auch, ob das Validity Argument die konkrete Sammlung der Evidenzen beinhaltet oder das auf Evidenzen basierende, angepasste IUA umfasst. Diese Unklarheiten erschweren die Kommunikation über die Validierung und die Umsetzung der Validierung. Ein weiteres ungünstiges Merkmal ergibt sich aus der Flexibilität des Ansatzes. Die Flexibilität an sich ist zwar ein positives Merkmal und ermöglicht die Nutzung des Ansatzes für viele verschiedene Testwertinterpretationen ohne viele Voraussetzungen. Jedoch gibt es aufgrund dieser Flexibilität keine Checkliste oder klare Regeln, welche Schritte bei der Validierung von Testwertinterpretationen notwendig sind. In dem Ansatz von Kane (2013) werden lediglich vier Schlussfolgerungen beschrieben, die in vielen Testungen getätigt werden und einige Vorschläge für weitere möglich Ergänzungen gemacht. Dies führt dazu, dass die Validierung unterschiedlicher Testwertinterpretationen unterschiedlich strukturierte IUAs produziert. So unterscheiden sich beispielsweise die Argumentationsketten aus dieser Arbeit, aus der Studie von Chapelle et al. (2010) und der Studie von Shaw und Crisp (2012). Dabei sind nicht nur die verwendeten Schlussfolgerungen und deren Reihenfolge verschieden, sondern auch die dazugehörigen Argumente und Annahmen. Die Flexibilität des Ansatzes bringt also eine gewisse Unsicherheit mit sich, in welcher die Testentwicklerinnen und Testentwickler beziehungsweise Testnutzerinnen und Testnutzer selbst das Rahmenkonzept ihrer Validierung bestimmen müssen. Eine weitere Unsicherheit ist das Formulieren einer abschließenden Beurteilung der Validität der Testwertinterpretation. Zum einen wird im Ansatz nicht genau beschrieben, wann genügend Evidenz für die abschließende Beurteilung vorhanden ist und wie eine allumfassende Beurteilung auf Basis aller

Ergebnisse zusammengefasst wird. Zum anderen ist unklar, wie viel Evidenz für eine gewisse Testwertinterpretation gesammelt werden muss, bevor diese angenommen werden kann. Dass der Argument Based Approach schon seit den neunziger Jahren im Gespräch ist und bisher nur sehr selten eingesetzt wurde (Newton & Shaw, 2014) sowie die hier aufgeführten Kritikpunkte sprechen für eine gewisse Komplexität des Ansatzes.

5.2.2 Implikationen für die Validierungspraxis

Mit der Anwendung des Argument Based Aproach für die Validierung des NEPS-K9-Mathematiktests sowie der kritischen Betrachtung des Ansatzes und der Durchführung wurde ein reflektiertes Praxisbeispiel für die Validierung auf Basis eine argumentationsbasierten Ansatzes der Validierung geschaffen. Des Weiteren können aus der Anwendung und Diskussion auch Implikationen für die Validierungspraxis abgeleitet werden. Im Folgenden sollen daher Richtlinien für die Validierung formuliert werden.

Bei der Validierung sollte immer von den beabsichtigten Testwertinterpretationen ausgegangen werden. Aussagen über die Validität sollten sich immer auf bestimmte Testwertinterpretationen bei einer spezifischen Testnutzung beziehen.

> „It is incorrect to use the unqualified phrase "the validity of the test". “ (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 2014, S.1)

Außerdem ist es vorteilhaft, schon bei der Entwicklung des Tests auf die Validierung der beabsichtigten Testwertinterpretation einzugehen (Kane, 2013; Mislevy et al., 2003). Dies ermöglicht eine genaue Abstimmung des Tests mit der beabsichtigten Testwertinterpretation sowie eine Optimierung im Entwicklungsprozess.

Validierung ist ein unendlicher Prozess, da immer neue Informationen gesammelt werden können, welche helfen, den Test und die Schlussfolgerungen besser zu verstehen. Validierung sollte daher ein Prozess des Bildens und Evaluierens von Argumenten für und gegen die beabsichtigte Testwertinterpretation und Testnutzung sein (Kane, 2013; Mislevy et al., 2003). Die Argumentation im Validierungsprozess ist in der Regel präsumtiv, da Annahmen zugunsten der Schlussfolgerungen aufgestellt werden, diese Annahmen aber

nicht definitiv bewiesen werden. Aus diesem Grund eignet sich das Argumentationsmodell von Toulmin (1958, 2003) besonders für die Bildung der Argumente (Kane, 2013; Mislevy, 2007; Bachman, 2005)

Für die Verbindung der Beobachtungen mit der beabsichtigten Testwertinterpretation eignet sich der Argument Based Approach von Kane besonders (siehe Kapitel 5.2.1 für eine Besprechung der Vorteile dieses Ansatzes). Während Kane (2013) keine genauen Angaben zum Testentwicklungsprozess machte, entwickelte Mislevy (2003) das Modell des Assessment Design hierzu (vgl. 1.1.3). Eine Orientierung an den Stufen des Assessment Design bietet nicht nur die Möglichkeit, den Prozess transparenter zu gestalten, sondern auch eine Orientierung, an welcher Stelle Schlussfolgerungen, Argumente und Annahmen modelliert werden können. Um dies zu verdeutlichen, wird im folgenden ein Durchlaufen der Stufen des Assessment Designs nach Mislevy, bei Anwendung des Argument Based Approach nach Kane (2013), kurz skizziert. Nach einer Phase der Informationssammlung und Beschreibung der Domäne, bei der die beabsichtigte Testinterpretation eine Schlüsselrolle spielt, kann mit der Modellierung der Domäne begonnen werden. An dieser Stelle werden die systematischen Strukturen, welche die Leistung beschreiben, die Arten der Beobachtungen, welche auf die Leistung schließen lassen, und die Arten der Aufgaben beziehungsweise Situationen, in denen diese Leistungen von den Testpersonen gefordert werden, basierend auf den zuvor gesammelten Informationen entwickelt. Diese Phase bietet also die Möglichkeit, alle Interpretationsschritte, die zur Verbindung der Beobachtungen im Test mit der beabsichtigten Testwertinterpretation getätigt werden, offen zu legen und nach Kane (2013) als IUA modellieren. Dabei bekommen die Interpretationsschritte die Funktion der Schlussfolgerung im IUA. Die Argumente und Annahmen werden aus den gesammelten Informationen in der ersten Phase abgeleitet. Beispielsweise kann aus der Informationssammlung deutlich werden, welche Teilbereiche beziehungsweise Inhalte durch die Testaufgaben abgedeckt werden müssen. Auf Basis dieser Definitionen können bereits vor der Entwicklung der Aufgaben Argumente und Annahmen entwickelt werden (beispielsweise über die Gewichtung der Teilbereiche, die Art der Aufgaben etc.), die für Validierung der Schlussfolgerung bestätigt werden müssen. In der nächsten Phase wird ein umfassendes Konzept für den Einsatz des Tests, basierend auf dem IUA, entwickelt. Dies bietet die Möglichkeit, die Aufgabeninhalte und -eigenschaften sowie alle Aspekte der Durchführung auf das IUA abzustimmen. Des Weiteren ist es möglich, das Sammeln bestimmter Evidenzen für die Validierung frühzeitig einzuplanen, wie zum Beispiel eine stichprobenartige Qualitätskontrolle der

Testdurchführung. Auch ein Schritt zurück in die Phase der Domänenmodellierung ist möglich, sodass die beabsichtigten Interpretationen frühzeitig angepasst werden können. Für die letzte Phase der Entwicklung, die tatsächliche Durchführung beziehungsweise Operationalisierung, wird durch die vorangegangene Phase sichergestellt, dass alle für die Validierung wichtigen Informationen gesammelt werden. Vor der tatsächlichen Testdurchführung sind auch kleinere Zwischenstudien wie eine Beurteilung der Aufgaben durch Expertinnen und Experten denkbar. Im Anschluss an die Sammlung von Evidenz durch die Testdurchführung, Expertenbeurteilungen, Pilotierungsstudien, etc. kann die Validität der Testwertinterpretation evaluiert und das Validitätsargument gebildet werden. Die Auswertungen können auf die Notwendigkeit hinweisen, in eine andere Phase zurückzukehren und beispielsweise neue Informationen zu sammeln, Interpretationen anzupassen oder Aufgaben und Durchführungsbedingungen zu ändern. Die Phasen können solange durchlaufen werden, bis ein Validitätsargument für den Test gebildet ist.

Es besteht ein professionelles Einvernehmen darüber, dass die Evaluation von Konsequenzen der Testnutzung ein wichtiger Bestandteil der Validierung ist (Shaw & Crisp, 2012). Welche Konsequenzen in die Validierung einbezogen werden, unterscheidet sich zwischen Validierungsansätzen. An dieser Stelle wird daher empfohlen, mindestens die Konsequenzen bezüglich der Entscheidungsregeln zu evaluieren, wie es im Argument Based Approach vorgesehen ist. Sollte eine ausführlichere Analyse der Konsequenzen angestrebt werden, so ist es möglich, sich an anderen Ansätzen wie dem Assessment Use Argument von Bachman (2005) oder der Konstruktvalidität nach Messick (1989a, 1994, 1995) zu orientieren.

Anhang

Vergleich der Validierungsstichprobe und der Stichprobe der Haupterhebung

Im Folgenden soll die Vergleichbarkeit der Validierungsstichprobe mit der Stichprobe der Haupterhebung dargestellt werden. Für den Vergleich wurde die Stichprobe hinsichtlich der Verteilung der mathematischen Fähigkeitswerte sowie einiger soziodemografischer Variablen untersucht. Für die Berechnungen wurden jeweils die Schülerinnen und Schüler ausgewählt, die den NEPS-K9-Mathematiktest bearbeitet hatten, da diese Stichprobe die Grundlage für die unterschiedlichen Berechnungen darstellte. Die Unterschiede in der mathematischen Fähigkeit der beiden Stichproben wurde in einem ersten Schritt durch die Gegenüberstellung der Verteilungsmerkmale analysiert.

Abbildung 38 und 39 zeigen die Verteilungen der mathematischen Fähigkeiten in der Validierungsstichprobe und in der Haupterhebung. Die jeweils gestrichelte Kurve zeigt eine Normalverteilung. Tabelle 22 zeigt die Verteilungsmerkmale der beiden Stichproben. Die Unterschiede wurden analysiert, indem das 95%-Konfidenzintervall um den Mittelwert der Haupterhebungsstichprobe berechnet wurde. Enthält dieses Konfidenzintervall den Mittelwert der Validierungsstichprobe, so kann geschlussfolgert werden, dass diese aus einer Population mit einem solchen Mittelwert stammt. Das 95%-Konfidenzintervall um den Mittelwert der Haupterhebungsstichprobe liegt bei KI[0.00;0.04]. Die Berechnungen zeigen, dass die Validierungsstichprobe im Mittel eine höhere mathematische Fähigkeit aufweist als die Stichprobe der Haupterhebung. Beide Verteilungen zeigen eine Rechtsschiefe und ebenso eine steilere Form im Vergleich zur Normalverteilung. Die Validierungsstichprobe weist außerdem einen höheren Standardfehler auf. Obwohl beide Verteilungen die gleiche Richtung in ihrer Abweichung von der Normalverteilung zeigen, unterscheiden sich die Abweichungen in ihrer Ausprägung. So fällt die Verteilungskurve der Validierungsstichprobe steiler und weniger rechtsschief aus als die Verteilung der Haupterhebung. Ein Grund für diese Unterschiede kann unter anderem die Zusammensetzung der Stichprobe sein. Im Folgenden wird eine Analyse der Verteilungen hinsichtlich mehrerer soziodemografischer Variablen dargelegt.

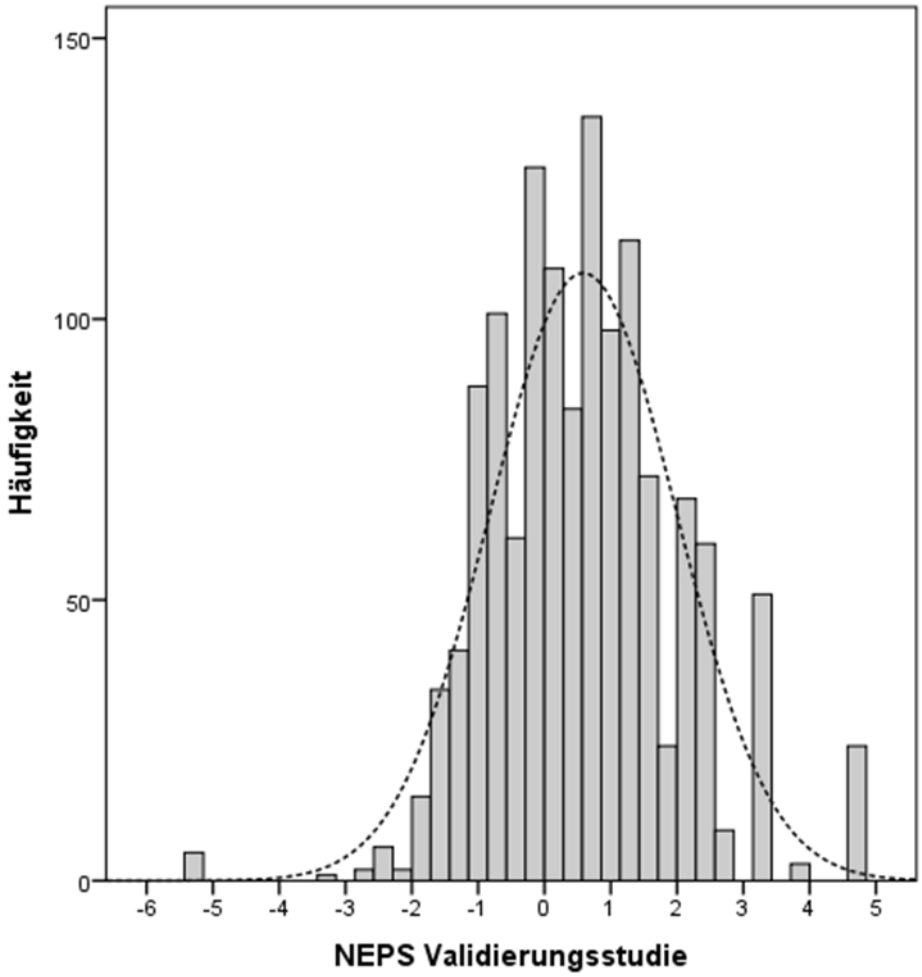


Abbildung 38: Verteilung der mathematischen Fähigkeiten in der Validierungsstudie im Vergleich zu einer Normalverteilung

Tabelle 22: Vergleich der Verteilungsmerkmale der Validierungsstichprobe und der Haupterhebung

	N	Min	Max	MW	SD	Schiefe	Kurtosis
Haupterhebung	14523	-4.37	4.62	0.02	1.21	0.68	0.77
Validierungsstudie	1335	-5.35	4.62	0.58	1.41	0.18	1.01

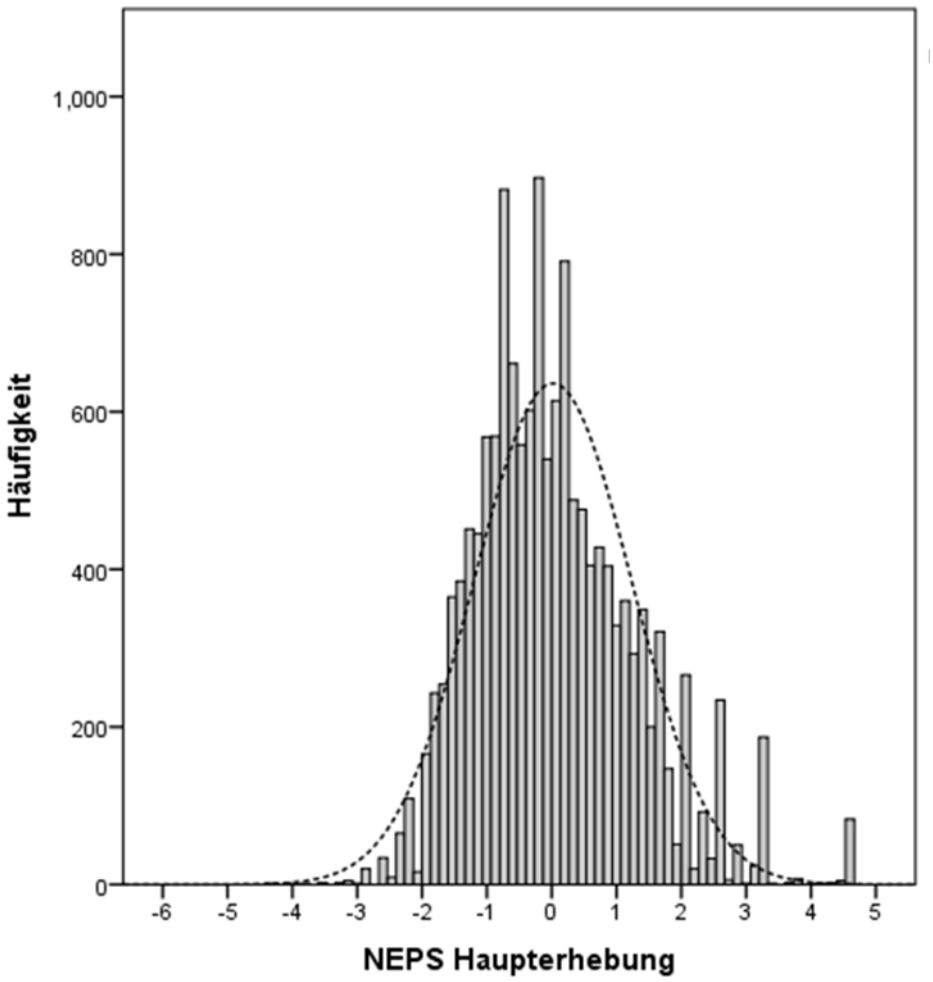


Abbildung 39: Verteilung der mathematischen Fähigkeiten in der Haupterhebung im Vergleich zu einer Normalverteilung

Unterschiede in der Verteilung der Stichproben in den Variablen Geschlecht, Schulform und Migrationshintergrund wurden mit Hilfe eines χ^2-Tests analysiert.

Tabelle 23: Vergleich der Stichproben der Haupterhebung und der Validierungsstudie hinsichtlich der Variablen Geschlecht, Schulform und Migrationshintergrund

			Haupt-erhebung	Validierungs-studie
Geschlecht	Männlich	Anzahl	7297	677
		Prozent	50.3	50.7
	Weiblich	Anzahl	7224	658
		Prozent	49.7	49.3
Schulform	Hauptschule	Anzahl	3563	35
		Prozent	24.5	2.6
	Schule mit mehreren Bildungsgängen	Anzahl	1125	145
		Prozent	7.7	10.9
	Realschule	Anzahl	3106	297
		Prozent	21.4	22.2
	Gymnasium	Anzahl	5115	628
		Prozent	35.2	47
Migrations-hintergrund	Kein Migrationshin-tergrund	Anzahl	10254	1125
		Prozent	74	88.9
	Ein Elternteil im Ausland geboren	Anzahl	1164	59
		Prozent	8.4	4.7

Fortsetzung auf der nächsten Seite

			Haupt-erhebung	Validierungs-studie
	beide Elternteile im Ausland geboren	Anzahl	1537	69
		Prozent	11.1	5.5
	Kind im Ausland geboren	Anzahl	902	12
		Prozent	6.2	0.9
Noten	Deutsch	MW	2.9	3.9
		SD	0.8	0.8
		Gültige Anzahl	13985	1329
	Mathematik	MW	3.0	3.1
		SD	1.1	1.0
		Gültige Anzahl	13953	1329
ISEI		MW	48.8	41.7
		SD	16.2	15.7
		Gültige Anzahl	11353	1186

Die Stichproben der Validierungsstudie und der Hauperhebung unterscheiden sich nicht signifikant bezüglich der Variable Geschlecht ($\chi^2 = 0.10$; $df = 1$, *ns.*). Beide Stichproben beinhalten ebenso viele Jungen wie Mädchen. Hinsichtlich der besuchten Schulformen unterscheiden sich die Verteilungen signifikant zwischen den Stichproben ($\chi^2 = 360$), $df = 4$). Deutlich weniger Schülerinnen und Schüler aus der Validierungsstudie (2.6%) als aus der Stichprobe der Haupterhebung (24.5%) besuchten eine Hauptschule. Dafür besuchten mehr Schülerinnen und Schüler der Validerungsstudie ein Gymnasium (47%) im Gegensatz zur Haupterhebungsstichprobe (35.2%). Auch die Verteilung der beiden Stichproben unterscheidet sich signifikant bezüglich des Migrationshintergrundes ($\chi^2 = 148$, $df = 3$). In der Validierungsstichprobe haben mit 88.9 % deutlich weniger Schülerinnen und Schüler einen Migrationshintergrund als in der Stichprobe der Haupterhebung (74 %).

Betreffend der Deutsch- und Mathematiknoten unterscheiden sich die Schülerinnen und Schüler der beiden Stichproben kaum. Aufgrund der höheren mathematischen Fähigkeiten der Testpersonen der Validierungsstichprobe entspricht dies nicht den Erwartungen. Eine Erklärung dieses Phänomens könnte sein, dass die Schülerinnen und Schüler an Gymnasien im Vergleich zu den übrigen Schulformen strenger benotet wurden beziehungsweise dass Hauptschülerinnen und -schüler im Vergleich zu Schülerinnen und Schülern anderer Schulformen weniger streng benotet wurden. Aufgrund des hohen Anteils an Gymnasien und des niedrigen Anteils an Hauptschulen in der Validierungsstichprobe ist es denkbar, dass die Schülerinnen und Schüler der Valideirungsstichprobe insgesamt strenger benotet wurden.

Dass die Validierungsstichprobe einen höheren Mittelwert bezüglich des International Socio-Economic Index of Occupational Status (ISEI) aufweist, spiegelt unter anderem die gefundenen Unterschiede bezüglich der Schulform und des Migrationshintergrundes wider.

Die Unterschiede in den soziodemografischen Variablen können die Unterschiede in den mathematischen Fähigkeitsverteilungen mit bedingen. Die Validerungsstichprobe besteht aufgrund des Untersuchungsdesigns nur aus Schulen, die an dem Schulentwicklungsprogramm SINUS beziehungsweise SINUS Transfer teilgenommen haben. Dies könnte ebenfalls einen Einfluss auf die Unterschiede in den mathematischen Kompetenzen haben.

Insgesamt muss bei der Interpretation der Analysen mit Daten der Validierungsstudie beachtet werden, dass sich die Stichprobe von der Stichprobe der Haupterhebung unterscheidet. Bei der Validierungsstichprobe handelt es sich um eine positive Selektion. Die Stichprobe zeigt eine positivere mathematische Fähigkeitsverteilung, höhere ISEI-Werte, weniger Schülerinnen und Schüler mit Migrationshintergrund und mehr Schülerinnen und Schüler an höheren Schulformen. Für einige Analysen, für welche die Daten beider Studien vorhanden waren (wie zum Beispiel der Zusammenhang der mathematischen Fähigkeit mit Noten), wurden die Berechnungen daher mit beiden Stichproben durchgeführt. Insgesamt zeigen diese doppelt durchgeführten Auswertungen, dass sich die Werte zwar leicht unterscheiden, die Interpretation jedoch die gleiche bleibt (vgl. Kapitel 2.6).

Literaturverzeichnis

Adams, R. J., Wilson, M. & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21* (1), 1–23. doi: 10.1177/0146621697211001

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

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

Anastasi, A. (1986). Evolving concepts of test validation. *Annual Review of Psychology*, *37* (1), 1–16. doi: 10.1146/annurev.ps.37.020186.000245

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly*, *2* (1), 1–34. doi: 10.1207/s15434311laq0201\textunderscore1

Bachman, L. F. & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford, New York: Oxford University Press.

Baumert, J., Trautwein, U. & Artelt, C. (2003). Schulumwelten - institutionelle Bedindungen des Lehrens und Lernens. In Deutsches PISA-Konsortium (Hrsg.), *Pisa 2000* (S. 261–332). Opladen: Leske + Budrich.

Bertschy, K., Cattaneo, M. A. & Wolter, S. C. (2009). PISA and the transition into the labour market. *Labour*, *23*, 111–137. doi: 10.1111/j.1467-9914.2008.00432.x

Blömeke, S., Gustafsson, J.-E. & Shavelson, R. J. (2015). Beyond dichotomies: Competence viewed as a continuum. *Zeitschrift für Psychologie*, *223* (1), 3–13. doi: 10.1027/2151-2604/a000194

Blossfeld, H.-P. & von Maurice, J. (2011). Education as a lifelong process. *Zeitschrift für Erziehungswissenschaft*, *14* (S2), 19–34. Zugriff am 28.06.2012 auf http://www

.springerlink.com/content/m0820p1801p42p84/fulltext.pdf doi: 10.1007/s11618-011-0179-2

Blossfeld, H.-P., von Maurice, J. & Schneider, T. (2011). The national educational panel study: need, main features, and research potential. *Zeitschrift für Erziehungswissenschaft*, *14* (S2), 5–17. Zugriff am 28.06.2012 auf http://www.springerlink.com/content/186px2140141211p/fulltext.pdf doi: 10.1007/s11618-011-0178-3

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37* (1), 29–51.

Borsboom, D., Mellenbergh, G. J. & van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111* (4), 1061–1071. doi: 10.1037/0033-295X.111.4.1061

Bortz, J. & Döring, N. (2006). *Forschungsmethoden und Evaluation* (4., überarbeitete Aufl.). Berlin, Heidelberg: Springer Berlin Heidelberg. doi: 10.1007/978-3-540-33306-7

Bos, W. & Schwippert, K. (2002). TIMSS, PISA, IGLU & Co: vom Sinn und Unsinn internationaler Schulleistungsuntersuchungen. *Bildung und Erziehung*, *55* (1), 5–23.

Brennan, R. L. (2013). Commentary on "validating the interpretations and uses of test scores". *Journal of Educational Measurement*, *50* (1), 74–83. doi: 10.1111/jedm.12001

Brunner, M. (2006). *Mathematische Schülerleistung : Struktur, Schulformunterschiede und Validität* (Dissertation, Humboldt-Universität, Berlin). Zugriff am 01.10.2012 auf http://library.mpib-berlin.mpg.de/diss/Brunner_Dissertation.pdf

Burstein, L., Aschbacher, P., Chen, Z. & Lin, L. (1990). *Establishing the content validity of tests designed to serve multiple purposes: Bridging secondary-postsecondary mathematics* (Nr. 313). Los Angeles, CA.

Carstensen, C. H. (2013). Linking PISA competencies over three cycles – Results from Germany. In M. Prenzel, M. Kobarg, K. Schöps & S. Rönnebeck (Hrsg.), *Research outcomes of the pisa research conference 2009* (S. 199–214). New York: Springer.

Chapelle, C. A. (2012). Validity argument for language assessment: The framework is simple. *Language Testing*, *29* (1), 19–27. doi: 10.1177/0265532211417211

Chapelle, C. A., Enright, M. K. & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice*, *29* (1), 3–13. doi: 10.1111/j.1745-3992.2009.00165.x

Chapelle, C. A., Enright, M. K. & Jamieson, J. M. (2009). *Building a validity argument*

for the test of English as a foreign language tm. New York and NY: Routledge.

Cizek, G. J., Bowen, D. & Church, K. (2010). Sources of validity evidence for educational and psychological tests: A follow-up study. *Educational and Psychological Measurement*, *70* (5), 732–743. doi: 10.1177/0013164410379323

Cizek, G. J., Rosenberg, S. L. & Koons, H. H. (2008). Sources of validity evidence for educational and psychological tests. *Educational and Psychological Measurement*, *68* (3), 397–412. doi: 10.1177/0013164407310130

Cohen, A. S. & Wollack, J. A. (2006). Test administration, security, scoring, and reporting. In R. L. Brennan (Hrsg.), *Educational measurement* (S. 355–386). Westport, CT: Praeger Publishers.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16* (3), 267–334.

Cronbach, L. J. (1989). Construct validation after thirty years. In R. L. Linn (Hrsg.), *Intelligence* (S. 147–171). Urbana: University of Illinois Press.

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

Cronbach, L. J. & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological bulletin*, *52* (4), 281–302.

Deutsches PISA-Konsortium (Hrsg.). (2003). *PISA 2000: Ein differenzierter Blick auf die Länder der Bundesrepublik Deutschland.* Opladen: Leske + Budrich.

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

Duchhardt, C. & Gerdes, A. (2013). *NEPS technical report for mathematics – scaling results of starting cohort 4 in ninth grade: (neps working paper no. 22).* Bamberg.

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 & M. Grüßing (Hrsg.), *Mathematiklernen vom Kindergarten bis zum Studium. Kontinuität und Kohärenz als Herausforderung für den Mathematikunterricht* (S. 313–327). Münster: Waxmann.

Ehmke, T., van den Ham, A.-K., Sälzer, C. & Heine, J.-H. (in Vorbereitung). Measuring mathematics competence in international and national large scale assessments: Linking PISA with the National Educational Panel Study in Germany.

Fan, X. & Sun, S. (2013). Item response theory. In T. Teo (Hrsg.), *Handbook of quan-*

titative methods for educational research (S. 45–67). Rotterdam: SensePublishers. doi: 10.1007/978-94-6209-404-8\textunderscore3

Geisinger, K. F. (1992). The metamorphosis to test validation. *Educational Psychologist*, *27* (2), 197–222.

Guion, R. M. (1978). Content validity in moderation. *Personnel Psychology*, *31* (2), 205–213. doi: 10.1111/j.1744-6570.1978.tb00440.x

Gulliksen, H. (1950). *Theory of mental tests.* New York: Wiley.

Haberkorn, K., Pohl, S., Carstensen, C. H. & Wiegand, E. (2013). Incorporating different response formats of neps competence tests in an irt-model: Manuscript submitted for publication.

Hansen, H. V. & Pinto, R. C. (1995). *Fallacies: Classical and contemporary readings.* University Park and Pa: Pennsylvania State University Press.

Hansen, J. (2010). *How does academic ability affect educational and labour market pathways in canada* (Nr. 30).

Hartig, J. & Frey, A. (2012). Konstruktvalidierung und Skalenbeschreibung in der Kompetenzdiagnostik durch die Vorhersage von Aufgabenschwierigkeiten. *Psychologische Rundschau*, *63* (1), 43–49. Zugriff am 07.08.2012 auf http://www.psycontent.com/content/8w4607412m163861/fulltext.pdf doi: 10.1026/0033-3042/a000109

Hartig, J. & Höhler, J. (2010). Modellierung von Kompetenzen mit mehrdimensionalen IRT-modellen: Projekt MIRT. *Zeitschrift für Pädagogik, Beiheft*, *56* (56), 189–198.

Hartig, J. & Klieme, E. (2006). Kompetenz und Kompetenzdiagnostik. In K. Schweizer (Hrsg.), *Leistung und Leistungsdiagnostik* (S. 127–143). Heidelberg: Springer Medizin Verlag.

Hogan, T. P. & Agnello, J. (2004). An empirical study of reporting practices concerning measurement validity. *Educational and Psychological Measurement*, *64* (5), 802–812.

Huyhn, H. M. & Ferrara, S. (1995). *A comparison of three statistical procedures to identify clusters of items with local dependency.* San Francisco, CA.

IEA Data Processing and Research Center. (2013). *Methodenbericht NEPS Startkohorte 4: Haupterhebung – Herbst/Winter 2010; A46, A67, A83.*

Johansone, I. (2000). Quality control observations of the TIMSS 2011 data collection. In M. O. Martin & I. V. Mullis (Hrsg.), *Methods and procedures in TIMSS and PIRLS 2011* (S. 1–14). Chestnut Hill, MA: TIMSS & PIRLS International Study

Center, Lynch School of Education, Boston College.

Jonkisz, E., Moosbrugger, H. & Brand, H. (2012). Planung und Entwicklung von Tests und Fragebögen. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 27–74). Berlin, Heidelberg: Springer-Verlag Berlin Heidelberg.

Kane, M. T. (1990). *An argument-based approach to validation.* Iowa City: American College Testing Program.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, *38* (4), 319–342. doi: 10.1111/j.1745-3984.2001.tb01130.x

Kane, M. T. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice*, *21* (1), 31–41. doi: 10.1111/j.1745-3992.2002.tb00083.x

Kane, M. T. (2004). Certification testing as an illustration of argument-based validation. *Measurement*, *2* (3), 135–170.

Kane, M. T. (2006). In praise of pluralism. A comment on borsboom. *Psychometrika*, *71* (3), 441–445. doi: 10.1007/s11336-006-1491-2

Kane, M. T. (2008). Terminology, emphasis, and utility in validation. *Educational Researcher*, *37* (2), 76–82. doi: 10.3102/0013189X08315390

Kane, M. T. (2012). Validating score interpretations and uses. *Language Testing*, *29* (1), 3–17. doi: 10.1177/0265532211417210

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50* (1), 1–73. doi: 10.1111/jedm.12000

Kelava, A. & Moosbrugger, H. (2012). Deskriptivstatistische Evaluation von Items (Itemanalyse) und Testwertverteilungen. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 75–102). Berlin, Heidelberg: Springer-Verlag Berlin Heidelberg.

Kelley, T. L. (1927). *Interpretation of educational measurements.* Yonkers-on-Hudson, NY: World Book Co.

Klieme, E. et al. (Hrsg.). (2010). *PISA 2009: Bilanz nach einem Jahrzehnt.* Münster: Waxmann.

Knigge, J. (2010). *Modellbasierte Entwicklung und Analyse von Testaufgaben zur Erfassung der Kompetenz „Musik wahrnehmen und kontextualisieren“* (Unveröffentlichte Dissertation). Universität Bremen, Bremen.

Köller, O. (2010). Bildungsstandards. In R. Tippelt & B. Schmidt (Hrsg.), *Handbuch Bildungsforschung* (S. 529–548). Wiesbaden: VS Verlag für Sozialwissenschaften / GWV Fachverlage, Wiesbaden.

Köller, O., Eßel-Ullmann, G. & Paasch, D. (2012). Validierung eines Instruments zur

Erfassung Standard-basierter mathematischer Kompetenzen in der Grundschule. *Psychologie in Erziehung und Unterricht*, *59* (3), 177–190. doi: 10.2378/peu2012.art14d

Lane, S. & Stone, C. A. (2006). Performance assessments. In R. L. Brennan (Hrsg.), *Educational measurement* (S. 387–432). Westport, CT: Praeger Publishers.

Lennon, R. T. (1956). Assumptions underlying the use of content validity. *Educational and Psychological Measurement*, *16* (3), 294–304. doi: 10.1177/001316445601600303

Lingel, K., Neuenhaus, N., Artelt, C. & Schneider, W. (2014). Der Einfluss des metakognitiven Wissens auf die Entwicklung der Mathematikleistung am Beginn der Sekundarstufe I. *Journal für Mathematik-Didaktik*, *35* (1), 49–77. doi: 10.1007/s13138-013-0061-2

Linn, R. L. (1997). Evaluating the validity of assessments: The consequences of use. *Educational Measurement: Issues and Practice*, *16* (2), 14–16. doi: 10.1111/j.1745-3992.1997.tb00587.x

Lissitz, R. W. & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher*, *36* (8), 437–448. doi: 10.3102/0013189X07311286

Lockl, K. (2013). *Assessment of procedural meta cognition: Scientific use file 2013.* Bamberg.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports* (3), 635–694.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, N.J: L. Erlbaum Associates.

Masters, G. N. (1982). A rasch model for partial credit scoring. *Psychometrika*, *47* (2), 149–174. doi: 10.1007/BF02296272

Maul, A. (2012). The validity of the mayer-salovey-caruso emotional intelligence test (MSCEIT) as a measure of emotional intelligence. *Emotion Review*, *4* (4), 394–402. doi: 10.1177/1754073912445811

McClelland, D. C. (1973). Testing for competence rather than for "intelligence". *American Psychologist*, *28* (1), 1–14.

McGhee, D., Peterson, J., Gillmore, J. & Lowell, N. (2008). *General mathematics placement test (mpt-g): Initial test development.*

Mehrens, W. A. (1997). The consequences of consequential validity. *Educational Measurement: Issues and Practice*, *16* (2), 16–18. doi: 10.1111/j.1745-3992.1997.tb00588

.x

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist*, *35* (11), 1012–1027. doi: 10.1037/0003-066X.35.11.1012

Messick, S. (1988). The once and future issues of validity. assessing the meaning and consequences of measurement. In H. Wainer, H. I. Braun & Educational Testing Service (Hrsg.), *Test validity* (Bd. 33-45). L. Erlbaum Associates.

Messick, S. (1989a). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, *18* (2), 5–11. doi: 10.3102/0013189X018002005

Messick, S. (1989b). Validity. In R. L. Linn (Hrsg.), *Educational measurement* (S. 13–103). New York, NY: Macmillan.

Messick, S. (1994). *Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. research report rr-94-45.* Princeton, NJ.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50* (9), 741–749. doi: 10.1037/0003-066X.50.9.741

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, *56* (2), 177–196. doi: 10.1007/BF02294457

Mislevy, R. J. (2007). Validity by design: Comments on lissitz and samuelsen. *Educational Researcher*, *36* (8), 463–469. doi: 10.3102/0013189X07311660

Mislevy, R. J., Steinberg, L. S. & Almond, R. G. (2003). *On the structure of educational assessments: CSE technical report 597.* Los Angeles, CA.

Monseur, C., Baye, A., Lafontaine, D. & Quittre, V. (2011). PISA test format assessment and the local independence assumption. In IERI (Hrsg.), *IERI monograph series, issues and methodologies in large-scale assessments* (Bd. 4, S. 131–156).

Moosbrugger, H. (2012). Item-Response-Theorie (IRT). In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 227–274). Berlin, Heidelberg: Springer-Verlag Berlin Heidelberg.

Moosbrugger, H. & Kelava, A. (2012). Qualitätsanforderungen an einen psychologischen Test (Testgütekriterien). In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 8–26). Berlin, Heidelberg: Springer-Verlag Berlin Heidelberg.

Moss, P. A. (2007). Reconstructing validity. *Educational Researcher*, *36* (8), 470–476. doi: 10.3102/0013189X07311608

Mullis, I. V. S., Martin, M. O., Foy, P. & Arora, A. (2012). *TIMSS 2011 international*

results in mathematics. Chestnut Hill, MA and Amsterdam: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College and International Association for the Evaluation of Educational Achievement (IEA).

Muraki, E. (1993). Information functions of the generalized partial credit model. *Applied Psychological Measurement*, *17* (4), 351–363.

Neumann, I., Duchhardt, C., Grüßing, M., Heinze, A., Knopp, E. & Ehmke, T. (2013). Modeling and assessing mathematical competence over the lifespan. *Journal for Educational Research Online*, *5* (2), 80–110.

Newton, P. E. (2013). Two kinds of argument? *Journal of Educational Measurement*, *50* (1), 105–109. doi: 10.1111/jedm.12004

Newton, P. E. & Shaw, S. (2014). *Validity in educational and psychological assessment*. London: Sage Publications.

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

OECD. (2010). *PISA mathematics framework 2012*. Zugriff am 12.06.2012 auf http://www.oecd.org/dataoecd/11/40/44455820.pdf

OECD. (2012). *PISA 2009 technical report*. OECD Publishing.

OECD. (2013). *PISA 2012 assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy*. Paris: OECD Publishing.

OECD. (2014). *PISA 2012 technical report*. OECD Publishing.

Pant, H. A., Stanat, P., Pöhlmann, C. & Böhme, K. (2013). Die Bildungsstandards im allgemeinbildenden Schulsystem. In H. A. Pant, P. Stanat, U. Schroeders, A. Roppelt, T. Siegle & C. Pöhlmann (Hrsg.), *IQB-Ländervergleich 2012* (S. 13–22). Münster and Berlin [u.a.]: Waxmann.

Pant, H. A., Stanat, P., Schroeders, U., Roppelt, A., Siegle, T. & Pöhlmann, C. (Hrsg.). (2013). *IQB-Ländervergleich 2012: Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I*. Münster and Berlin [u.a.]: Waxmann.

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

Pohl, S. & Carstensen, C. H. (2013). Scaling of competence tests in the National Educational Panel Study - many questions, some answers, and further challenges. *Journal for Educational Research Online*, *5* (2), 186–216.

Pohl, S., Grafe, L. & Rose, N. (2014). Dealing with omitted and not-reached items in competence tests: Evaluating approaches accounting for missing responses in item response theory models. *Educational and Psychological Measurement*, *74* (3), 423–

452. doi: 10.1177/0013164413504926

Popham, W. J. (1997). Consequential validity: Right concern-wrong concept. *Educational Measurement: Issues and Practice*, *16* (2), 9–13. doi: 10.1111/j.1745-3992.1997.tb00586.x

Prenzel, M. et al. (Hrsg.). (2004). *PISA 2003. Der Bildungsstand der Jugendlichen in Deutschland : Ergebnisse des zweiten internationalen Vergleichs*. Münster: Waxmann.

Prenzel, M., Sälzer, C., Klieme, E. & Köller, O. (Hrsg.). (2013). *PISA 2012: Fortschritte und Herausforderungen in Deutschland*. Münster: Waxmann.

Prenzel, M., Walter, O. & Frey, A. (2007). PISA misst Kompetenzen: Eine Replik auf Rindermann (2006): Was messen internationale Schulleistungsstudien? *Psychologische Rundschau*, *58* (2), 128–136. doi: 10.1026/0033-3042.58.2.128

Richter, D., Engelbert, M., Böhme, K., Haag, N., Hannighofer, J., Reimers, H., ... Stanat, P. (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* (S. 85–102). Münster: Waxmann.

Robitzsch, A. (2009). Methodische Herausforderungen bei der Kalibrierung von Leistungstests. In A. Bremerich-Vos, D. Granzer & O. Köller (Hrsg.), *Bildungsstandards Deutsch und Mathematik* (S. 42–106). Weinheim: Beltz.

Roppelt, A., Blum, W. & Pöhlmann, C. (2013). Die im Ländervergleich 2012 untersuchten mathematischen und naturwissenschaftlichen Kompetenzen. In H. A. Pant, P. Stanat, U. Schroeders, A. Roppelt, T. Siegle & C. Pöhlmann (Hrsg.), *IQB-Ländervergleich 2012* (S. 23–37). Münster and Berlin [u.a.]: Waxmann.

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

Rost, J. (2004). *Lehrbuch Testtheorie - Testkonstruktion* (2., vollständig überarb. und erw. Aufl.). Bern: Huber.

Sälzer, C. & Prenzel, M. (2013). PISA 2012 - eine Einführung in die aktuelle Studie. In M. Prenzel, C. Sälzer, E. Klieme & O. Köller (Hrsg.), *PISA 2012* (S. 11–46). Münster: Waxmann.

Schermelleh-Engel, K. & Werner, C. S. (2012). Methoden der Reliabilitätsbestimmung. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 119–142). Berlin, Heidelberg: Springer-Verlag Berlin Heidelberg.

Schneider, W. & Artelt, C. (2010). Metacognition and mathematics education. *ZDM*, *42* (2), 149–161. doi: 10.1007/s11858-010-0240-2

Scriven, M. (1987). Validity in personnel evaluation. *Journal of Personnel Evaluation in Education*, *1* (1), 9–23. doi: 10.1007/BF00143275

Seidl, T. & Prenzel, M. (2008). Assessment in large-scale studies. In J. Hartig, E. Klieme & D. Leutner (Hrsg.), *Assessment of competencies in educational contexts* (S. 279–304). Cambridge and Mass: Hogrefe.

Sekretariat der ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland. (02.06.2006). *Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring: (Beschluss der Kultusministerkonferenz vom 02.06.2006).* Berlin.

Senkbeil, M., Ihme, J. M. & Wittwer, J. (2013). Entwicklung und erste Validierung eines Tests zur Erfassung technologischer und informationsbezogener Literacy (TILT) für Jugendliche am Ende der Sekundarstufe I. *Zeitschrift für Erziehungswissenschaft*, *16*, 671–691.

Shaw, S. & Crisp, V. (2012). An approach to validation: Developing and applying an approach for the validation of general qualifications. *Research Matters: A Cambridge Assessment Publication* (Special Issue 3).

Shepard, L. A. (1993). Evaluating test validity. In L. Darling-Hammond (Hrsg.), *Review of research in education* (S. 405–450). Washington, DC.

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice*, *16* (2), 5–24. doi: 10.1111/j.1745-3992.1997.tb00585.x

Siegle, T., Schroeders, U. & Roppelt, A. (2013). Anlage und Durchführung des Ländervergleichs. In H. A. Pant, P. Stanat, U. Schroeders, A. Roppelt, T. Siegle & C. Pöhlmann (Hrsg.), *IQB-Ländervergleich 2012* (S. 101–122). Münster and Berlin [u.a.]: Waxmann.

Sireci, S. G. (2007). On validity theory and test validation. *Educational Researcher*, *36* (8), 477–481. doi: 10.3102/0013189X07311609

Stalder, B. E., Meyer, T. & Hupka-Brunner, S. (2008). Leistungsschwach – Bildungsarm ? Ergebnisse der TREE-Studie zu den PISA-Kompetenzen als Prädiktoren für Bildungschancen in der Sekundarstufe II. *Die Deutsche Schule*, *100* (4), 436–448.

Standards for educational and psychological tests and manuals. (1966). Washington, DC: American-Psychological-Association.

Standards for educational and psychological tests and manuals. (1974). Washington, DC: American-Psychological-Association.

Strobl, C. (2012). *Das Rasch-Modell: Eine verständliche Einführung für Studium und*

Praxis (2. Aufl., Bd. 2). Mering: Rainer Hampp Verlag.

Thorndike, E. L., Bregman, E. O., Cobb, M. V. & Woodyard, E. (1927). *The measurement of intelligence.* New York.

Toulmin, S. (1958). *The uses of argument.* Cambridge [England]: University Press. Zugriff auf BC177.T61958

Toulmin, S. (2003). *The uses of argument* (erneuerte Aufl.). Cambridge and U.K, New York: Cambridge University Press.

Universität Bamberg, N. (2012). *Startkohorte 4, Haupterhebung 2010/11 (A47) Schüler/innen, Klasse 9 in Regelschulen: Informationen zum Kompetenztest.* Bamberg.

van den Ham, A.-K., Ehmke, T., Roppelt, A. & Stanat, P. (angenommen). Assessments verbinden, Interpretationen erweitern? Lassen sich die mathematischen Kompetenzskalen im Nationalen Bildungspanel und im IQB-Ländervergleich 2012 verbinden? *Zeitschrift für Erziehungswissenschaft.*

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

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.

Walton, D. N. (1989). *Informal logic: A handbook for critical argumentation.* Cambridge: Cambridge University Press.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54* (3), 427–450. doi: 10.1007/BF02294627

Weinert, F. E. (2001). Concept of competence: A conceptual clarification. In D. S. Rychen & L. H. Sagalnik (Hrsg.), *Definition and selection of competencies: theoretical and conceptual foundations* (S. 46–65). Ashland, OH, US: Hogrefe & Huber Publishers.

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

Whitely, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, *93* (1), 179–197. doi: 10.1037/0033-2909.93.1.179

Winkelmann, H., Robitzsch, A., Stanat, P. & Köller, O. (2012). Mathematische Kompetenzen in der Grundschule. *Diagnostica*, *58* (1), 15–30. doi: 10.1026/0012-1924/a000061

Winter, H. (2003). Mathematikunterricht und Allgemeinbildung. In H.-W. Henn & K. Maaß (Hrsg.), *Materialien für einen realitätsbezogenen Mathematikunterricht* (Bd. 8, S. 6–15). Hildesheim: Franzbecker.

Wongwiwatthananukit, S., Popovich, N. G. & Bennett, D. E. (2000). Assessing pharmacy student knowledge on multiple-choice examinations using partial-credit scoring of combined-response multiple-choice items. *American Journal of Pharmaceutical Education*, *64* (1), 1–10.

Wu, M. (2013). Using item response theory as a tool in educational measurement. In M. M. C. Mok (Hrsg.), *Self-directed learning oriented assessments in the asia-pacific* (Bd. 18, S. 157–185). Dordrecht, New York: Springer.

Wu, M., Adams, R. J., Wilson, M. R. & Haldane, S. A. (2007). ACERconquest version 2: Generalised item response modelling software.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, *8* (2), 125–145. doi: 10.1177/014662168400800201

Yen, W. M. & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Hrsg.), *Educational measurement* (S. 111–153). Westport, CT: Praeger Publishers.