Studien zum Physik- und Chemielernen

M. Hopf, H. Niedderer, M. Ropohl, E. Sumfleth [Hrsg.]

316

Volker Brüggemann

Entwicklung und Pilotierung eines adaptiven Multistage-Tests zur Kompetenzerfassung im Bereich naturwissenschaftlichen Denkens

λογος

Studien zum Physik- und Chemielernen

Herausgegeben von Martin Hopf, Hans Niedderer, Mathias Ropohl und Elke Sumfleth

Diese Reihe im Logos Verlag Berlin lädt Forscherinnen und Forscher ein, ihre neuen wissenschaftlichen Studien zum Physik- und Chemielernen im Kontext einer Vielzahl von bereits erschienenen Arbeiten zu quantitativen und qualitativen empirischen Untersuchungen sowie evaluativ begleiteten Konzeptionsentwicklungen zu veröffentlichen. Die in den bisherigen Studien erfassten Themen und Inhalte spiegeln das breite Spektrum der Einflussfaktoren wider, die in den Lehr- und Lernprozessen in Schule und Hochschule wirksam sind.
Die Herausgeber hoffen, mit der Förderung von Publikationen, die sich mit dem Physik- und Chemielernen befassen, einen Beitrag zur weiteren Stabilisierung der physik- und chemiedidaktischen Forschung und zur Verbesserung eines an den Ergebnissen fachdidaktischer Forschung orientierten Unterrichts in den beiden Fächern zu leisten.

Martin Hopf, Hans Niedderer, Mathias Ropohl und Elke Sumfleth

Studien zum Physik- und Chemielernen

Band 316

Volker Brüggemann

Entwicklung und Pilotierung eines adaptiven Multistage-Tests zur Kompetenzerfassung im Bereich naturwissenschaftlichen Denkens

Logos Verlag Berlin

λογος

Studien zum Physik- und Chemielernen
Martin Hopf, Hans Niedderer, Mathias Ropohl und Elke Sumfleth [Hrsg.]

Bibliografische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

ISBN 978-3-8325-5331-9
ISSN 1614-8967

Logos Verlag Berlin GmbH
Georg-Knorr-Str. 4, Geb. 10
D-12681 Berlin

Tel.: +49 (0)30 / 42 85 10 90
Fax: +49 (0)30 / 42 85 10 92
https://www.logos-verlag.de

Entwicklung und Pilotierung eines adaptiven Multistage-Tests zur Kompetenzerfassung im Bereich naturwissenschaftlichen Denkens

Dissertation

zur Erlangung des Grades eines
Doktors der Naturwissenschaften

am Fachbereich Physik
der Freien Universität Berlin

vorgelegt von

Herr Volker Brüggemann, M.Ed.

aus Köln

Berlin, 2020

Erstgutachter: Prof. Dr. Volkhard Nordmeier

Zweitgutachter: Prof. Dr. Thomas Trefzger

Tag der Disputation: 31.03.2021

Kurzfassung

Im Zuge der Bologna-Hochschulreformen wurden in den letzten Jahren verstärkt Kompetenzmodelle und Messinstrumente für ihre Erfassung entwickelt. Da die untersuchten Kompetenzen häufig sehr komplexe Konstrukte mit mehreren Dimensionen sind, gestalten sich die Testinstrumente ebenfalls komplex und damit zeitaufwändig. Zudem werden zur quantitativen Kompetenzmessung vorrangig Leistungstests verwendet. Das Resultat sind häufige und teils langwierige Befragungen von Studierenden. Diese hohe Testbelastung der Proband*innen bringt Nachteile mit sich, wie beispielsweise eine sinkende Teilnahmebereitschaft. Auffallend ist im Zusammenhang damit der noch sehr seltene Einsatz adaptiver Testformate. Diese weisen eine deutlich höhere Effizienz (also Messgenauigkeit über Testlänge) auf als klassische, lineare Formate, womit sie die Belastung eingrenzen und die Teilnahmebereitschaft von Studierenden erhöhen könnten. Daher wird in dieser Arbeit die Anpassung eines bereits bestehenden Leistungstests von einem linearen in ein adaptives Format beschrieben.

Grundlage dieses Vorhabens ist der sogenannte *Ko-WADiS-Test*, der in den Projekten *Ko-WADiS* und *ValiDiS* entwickelt wurde. Dieser Multiple-Choice-Test wird eingesetzt, um die Kompetenz naturwissenschaftlichen Denkens bei Studierenden der Fächer Biologie, Chemie und Physik zu messen.

Um der Arbeit den notwendigen theoretischen Rahmen zu geben, werden dieser Test sowie die Grundkonzepte adaptiver Testformate und der für sie erforderlichen Item Response Theory vorgestellt und erläutert.

Es folgt die Beschreibung der Konstruktion des neuen Tests. Hierzu wurde auf Grundlage der Daten aus den beiden Projekten zunächst der vorhandene Itempool überprüft und für die Testkonstruktion normiert. Danach wurden verschiedene Testformate in Simulationsstudien getestet und im Hinblick auf Messgenauigkeit und Effizienz verglichen. Ein Multistage-Test mit Aufspaltung in drei Teststufen und zwei Fähigkeitsniveaus erwies sich als die beste der betrachteten Testvarianten.

Als Abschluss der Testentwicklung wurde der neue adaptive Test pilotiert und mit dem linearen Ko-WADiS-Test verglichen. Durch das neue Testformat konnte eine signifikante Steigerung der Effizienz um 53% erreicht werden.

Die Ergebnisse werden zum Abschluss der Arbeit zusammengefasst und in ihrer Bedeutung für zukünftige Kompetenzmessungen in der didaktischen Forschung diskutiert.

Abstract

In the course of the Bologna university reforms in recent years, an increasing number of competence models and measuring instruments have been developed. Since the competences studied are often extraordinarily complex constructs with several dimensions, the testing instruments are also complex and therefore time-consuming. In addition, performance tests are primarily used for quantitative competence measurement. The result is a high level of testing load on the test subjects. In this context, the rare use of adaptive tests is striking. These tests show a significantly higher efficiency (i.e. measurement accuracy over test length) than classic, linear formats, thus limiting the burden and increasing the willingness of students to participate. It is suspected that the effort required to develop an adaptive test is not considered justified by many researchers. Therefore, this thesis describes the adaptation of an already existing performance test from a linear to an adaptive format.

The basis of this project was the *Ko-WADiS*-test, which was developed in the research projects *Ko-WADiS* and *ValiDiS*. This multiple-choice test is used to measure the competence of scientific thinking in students of biology, chemistry and physics.

In the theory section of the thesis, this test as well as the basic concepts of adaptive test formats and item response theory (which is a prerequisite for adaptive testing) are presented and explained.

This is followed by the description of the construction of the new test. Based on the data from both projects, the existing pool of items was first checked and standardized. Afterwards, different test formats were tested in simulation studies and compared in terms of measurement accuracy and efficiency. A multistage test with three test stages and two skill levels proved to be the best of the test variants considered.

As a conclusion of test development, the new adaptive test was piloted in a group with known proficiency levels and compared with the linear Ko-WADiS test. The new test format achieved a significant increase in efficiency of 53%.

The results will be discussed at the end of the thesis, and their significance for future competence measurements in didactic research will be assessed.

Inhalt

1 Einleitung

Seit dem Beginn der 2000er Jahre ist die Diskussion um Kompetenzerwerb und -messung immer mehr in den Fokus deutscher Didaktik und Bildungsforschung gerückt. Auch wenn es sicherlich nicht der alleinige Ursprung war, so ist diese Entwicklung spätestens seit dem sogenannten *PISA-Schock* in der gesamten Forschungslandschaft ebenso wie in der Öffentlichkeit und Politik präsent. Das schlechte Abschneiden deutscher Schüler*innen in internationalen Bildungsvergleichen führte zu einer anhaltenden Diskussion über die deutsche Schulbildung.

Eines der Resultate aus diesem Diskurs war der Beschluss neuer Bildungsstandards für ganz Deutschland (Kultusministerkonferenz, 2005). Mit diesen wurde offiziell die Abkehr von der früheren Input-Orientierung beschlossen, Lehrpläne sollten sich also nicht mehr auf zu erlernendes Faktenwissen konzentrieren. Stattdessen wurde der Fokus im Rahmen der neuen Output-Orientierung auf den Kompetenzerwerb gelegt. Neben inhaltlichen Konsequenzen dieser Neuausrichtung war das Ziel vor allem die Qualitätssicherung und Vergleichbarkeit von Bildungssystemen der verschiedenen Bundesländer (Kultusministerkonferenz, 2005).

Parallel zu diesen Veränderungen fand an den deutschen Hochschulen eine Reform der bestehenden Studiengänge im *Bologna-Prozess* statt (Europäische Kommission, 2019). Hierbei wurden europaweit Studienstrukturen verändert, am offensichtlichsten ist der Wechsel der meisten Studiengänge hin zum Bachelor-Master-System sowie zu einer Einstufung von Studienleistungen anhand der European Credit Transfer Systems (ECTS) (Bundesministerium für Bildung und Forschung [BMBF], 2015, 2018). Mit den gemeinsamen Reformen sollten die Vergleichbarkeit von Abschlüssen in verschiedenen Ländern, die bessere Anschlussfähigkeit der Absolvent*innen an den Arbeitsmarkt sowie eine neue Stufe der Qualitätssicherung an Hochschulen erreicht werden (Bosbach, 2007).

Beide Prozesse, auch wenn durch verschiedene Ereignisse primär angestoßen, hatten also im Kern sehr ähnliche Ziele. Ebenso ist beiden

Entwicklungen gemein, dass eine der großen Umstellungen der Wechsel zur Kompetenzorientierung war.

Die neue Definition der Bildungsziele an Hochschulen sowie die gesteigerten Anforderungen an Qualitätssicherung und Monitoring der tatsächlichen Erfolge führten in den anschließenden Jahren zur Entwicklung einer Vielzahl neuer Kompetenzmodelle und Messinstrumente. In Deutschland wurde im Zuge dieser Entwicklungen unter anderem die BMBF[1]-Förderlinie *Kompetenzmodellierung und Kompetenzerfassung im Hochschulsektor* (KoKoHs) eingerichtet. Im Rahmen von KoKoHs wurden im Zeitraum von 2011 bis 2019 16 Forschungsprojekte durchgeführt und über 40 neue Messinstrumente entwickelt (Zlatkin-Troitschanskaia et al., 2020).

Die bei diesen Vorhaben eingesetzten Kompetenzmodelle zeichnen sich häufig durch eine hohe Komplexität aus, sie umfassen motivationale Aspekte und verschiedene Wissensbereiche (vgl. Hartig & Klieme, 2006). Die Grundlagenforschung und Testentwicklung werden dadurch erschwert, weil neben der Erfassung der Kompetenzen selbst auch die Untersuchung und Abgrenzung von nahestehenden Konstrukten, wie beispielsweise Intelligenz, relevant ist (Schweizer, 2006). Das kann zu ganzen „Batterien" von Befragungen führen, die für die Beantwortung von einzelnen Forschungsfragen notwendig sind, womit der Testaufwand im Bereich der Kompetenzmessung relativ hoch ist (Ehmke, 2006). Unter diesem Gesichtspunkt scheint es logisch, möglichst effiziente und zeitsparende Messinstrumente zu verwenden.

Im internationalen Raum werden zu diesem Zweck immer öfter *adaptive Testformate* (Kapitel 4) eingesetzt, bei denen sich die verwendeten Messinstrumente selbstständig an Fähigkeiten der Befragten Personen anpassen (van der Linden & Glas, 2010). Sie haben eine grundsätzlich höhere Effizienz als klassische lineare Tests, wie beispielsweise in Papierform angewandte Fragebögen (siehe Abschnitt 4.1 und 4.4). Im deutschsprachigen Raum werden solche Instrumente vor allem in den Bereichen der Psychologie und Medizin eingesetzt; in der Didaktik bisher nur selten (wobei es

[1] Bundesministerium für Bildung und Forschung

durchaus Initiativen zur adaptiven Kompetenzmessung gibt, vgl. Klieme, (2010)).

Ein Grund für die langsame Etablierung solcher Testverfahren könnte im damit verbundenen Aufwand vermutet werden: Mit der Konstruktion eines adaptiven Tests sind tendenziell höhere Anforderungen verbunden als mit den bisher verbreiteten linearen Testformaten (Segall, 2005). Möglicherweise wird dieser Aufwand aus Sicht der aktuell Forschenden als nicht angemessen beurteilt.

In dieser Arbeit soll als Beispiel für den praktikablen Einsatz von adaptiven Testformaten ein bereits etablierter Kompetenztest aus der bestehenden linearen Form zu einer adaptiven Version weiterentwickelt werden. Als Anwendungsfall für dieses Vorhaben wird der *Ko-WADiS-Test* genutzt, der im gleichnamigen Projekt Ko-WADiS[2] entwickelt wurde. Die Arbeit ist im Rahmen des Projekts ValiDiS[3] entstanden, welches sich mit Validierungsstudien zum Einsatz dieses Instruments befasst. Die beiden Projekte und das Instrument selbst werden in Kapitel 2 vorgestellt.

Die Kapitel 3 und 4 befassen sich danach mit theoretischen Grundlagen adaptiver Testformate. Im dritten Kapitel wird zunächst die Item Response Theory (IRT) betrachtet. Es handelt sich dabei um die mathematische Grundlage, die adaptives Testen erst ermöglicht: Die Anpassung der Befragungen, also die Adaption, folgt bestimmten Regeln und Gleichungen der IRT. Daher ist es unumgänglich, vor der Erläuterung des Testformats selbst einige Grundkonzepte und Begriffe der IRT zu kennen.

Nach der Klärung der mathematischen Grundlagen folgt in Kapitel 4 die Betrachtung aller anderen Aspekte von adaptiven Testverfahren. Das umfasst vor allem die praktische Gestaltung und Umsetzung solcher Tests, die verschiedenen möglichen Variationen und den Vergleich von linearen und

[2] Kompetenzmodellierung und -erfassung zum Wissenschaftsverständnis über naturwissenschaftliche Denk- und Arbeitsweisen bei Studierenden (Lehramt) in den drei Naturwissenschaften Bio-logie, Chemie und Physik

[3] Kompetenzmodellierung und -erfassung: Validierungsstudie zum wissenschaftlichen Denken im naturwissenschaftlichen Studium.

adaptiven Formaten in Bezug auf den Testeinsatz sowie mögliche Vor- und Nachteile.

Kapitel 5 eröffnet den praktischen Teil der Arbeit. Es dient der genaueren Eingrenzung und Definition der Problemstellung, die in den durchgeführten Studien bearbeitet wird. Dementsprechend werden an dieser Stelle die forschungsleitenden Fragestellungen der Arbeit (im Folgenden *Forschungsfragen*) ausformuliert und ein Plan für ihre Beantwortung erstellt.

Diese Fragen werden in jeweils einzelnen aufeinanderfolgenden Schritten in den Kapiteln 6 bis 8 beantwortet. Hierbei handelt es sich jeweils um die Erarbeitung der Datenlage für die Testentwicklung, die Entwicklung selbst sowie die anschließende Erprobung in einer Pilotstudie. In jedem dieser Kapitel werden Vorgehen, Methodik, Ergebnisse sowie Diskussion der notwenigen Studien dargestellt.

In Kapitel 9 werden die Ergebnisse und Schlussfolgerungen aus den vorangegangenen Studien zusammengefasst und diskutiert. An dieser Stelle findet auch die Diskussion darüber statt, ob der Aufwand der Testentwicklung als gerechtfertigt eingestuft wird. Hierbei muss zum einen der Nutzen für das Projekt selbst in Betracht gezogen werden, zum anderen muss aber auch abstrahiert und eine mögliche Verbreitung adaptiver Verfahren in didaktischen Forschungsprojekten diskutiert werden. Da hierfür nur Erfahrungen aus einer Testanpassung vorhanden sind, wird es bei einer qualitativen Betrachtung des Themas bleiben. Somit greift das letzte Kapitel alle vorigen noch einmal auf und bildet den Abschluss dieser Arbeit.

2 Der Ko-WADiS-Test

In diesem Kapitel soll der Ko-WADiS-Test als Ausgangspunkt der späteren Studien vorgestellt werden.

Als erstes wird in den Abschnitten 2.1 und 2.2 das Kompetenzkonstrukt betrachtet, für dessen Messung der Test entwickelt wurde. Das beinhaltet das projektintern verwendete Modell zu naturwissenschaftlichem Denken ebenso wie die bereits vorher veröffentlichten Kompetenzmodelle, auf denen der Ko-WADiS-Test basiert.

In welcher Weise diese Kompetenz in den Aufgaben des Tests operationalisiert wurde, wird getrennt in Abschnitt 2.3 beschrieben.

Im letzten Abschnitt (2.4) folgt dann die Darstellung der beiden Forschungsprojekte Ko-WADiS und ValiDiS sowie der Studien, in deren Rahmen das Instrument entwickelt und erprobt wurde. Insbesondere kann hier geklärt werden, welche Ziele mit der Testerstellung verfolgt und in welchem Umfeld der Test bereits eingesetzt und empirisch abgesichert wurden.

2.1 Kompetenzbegriff

Mit dem Ko-WADiS-Test soll die Kompetenz naturwissenschaftlichen Denkens erfasst werden. Für die Betrachtung des Testinstruments und die Anpassung in ein adaptives Format im späteren Verlauf der Arbeit (Kapitel 5 bis 8) ist es wichtig, die Struktur dieser Kompetenz genau zu verstehen. Die vermutete Kompetenzstruktur hatte nicht nur Folgen bei der schon abgeschlossenen Aufgabenkonstruktion, sondern spielt auch eine elementare Rolle bei der mathematischen Beschreibung der Aufgaben und Messwerte in der Auswertung von Befragungen (vgl. Kapitel 3).

Als erstes ist festzuhalten, dass der Begriff *Kompetenz* im Rahmen von Ko-WADiS der Definition als „kontextspezifische kognitive Leistungsdispositionen“ (Hartig & Klieme, 2006, S. 128) folgt. Danach handelt es sich bei dem gemessenen Konstrukt also um ein kognitives Merkmal. Dieses wird durch mehrere Aspekte gegen allgemeinere kognitive Fähigkeiten, wie beispielsweise Intelligenz, abgegrenzt.

Kompetenzen gelten im Gegensatz zu Intelligenz als gezielt und mittelfristig veränderbar. Sie werden nicht vererbt, sondern durch die Bewältigung von Anforderungen und Problemen mit der Zeit erlernt und trainiert. Das kann nicht nur passiv geschehen, sondern auch aktiv und gezielt gefördert werden (Hartig & Klieme, 2006).

Weiterhin werden Kompetenzen als bereichsspezifische Fähigkeiten betrachtet (Weinert, 2002). Damit ist gemeint, dass eine Kompetenz immer im Rahmen von bestimmten, miteinander verwandten Situationen benötigt und erworben wird. Im vorliegenden Beispiel wird die Kompetenz naturwissenschaftlichen Denkens im Kontext von wissenschaftlichen Untersuchungen erworben und kann auch nur in begrenztem Maße darüber hinaus Anwendung finden. Für das Halten eines Vortrages wäre ein gänzlich anderer Satz an Kompetenzen notwendig, da die dabei gestellten Anforderungen sich zu stark unterscheiden.

Neben dieser Unterscheidung zu anderen kognitiven Merkmalen grenzt sich die Kompetenzauffassung auch gegen andere Personenmerkmale ab, die Handlungen beeinflussen und steuern können, wie emotionale, affektive und motivationale Dispositionen oder Zustände (Hartig & Klieme, 2006).

2.2 Naturwissenschaftliches Denken

Nachdem in Abschnitt 2.1 der verwendete Kompetenzbegriff allgemein definiert wurde, kann jetzt die konkrete Beschreibung naturwissenschaftlichen Denkens folgen.

Tabelle 1 zeigt das Strukturmodell naturwissenschaftlichen Denkens, das im Projekt Ko-WADiS entwickelt wurde und bei der Konstruktion des Testinstruments Verwendung fand (Krüger, Upmeier zu Belzen & Hartmann, 2020). Die Kompetenz wird in zwei grundlegende Kompetenzbereiche, *Naturwissenschaftliche Untersuchungen* und *Modelle Nutzen*, sowie in sieben untergeordnete Facetten aufgeteilt. Die beiden übergeordneten Kompetenzbereiche entstammen jeweils verschiedenen Modellen zur Erkenntnisgewinnung, die kurz vorgestellt werden.

Tabelle 1: Strukturmodell der Kompetenz naturwissenschaftlichen Denkens.

Naturwissenschaftliche Untersuchungen	Fragestellungen	Hypothesen	Planung und Durchführung	Auswertung und Interpretation
Modelle Nutzen	Zweck von Modellen	Testen von Modellen	Ändern von Modellen	

Beim ersten Kompetenzmodell handelt es sich um die allgemeine Beschreibung naturwissenschaftlicher Erkenntnisgewinnung durch Mayer (2007). Hiernach handelt es sich bei naturwissenschaftlichem Denken um einen der drei Kompetenzbereiche, die Erkenntnisgewinnung als Ganzes ausmachen (Abbildung 1). *Naturwissenschaftliches Denken* bezeichnet damit nicht mehr und nicht weniger als diejenigen Denkmuster und Problemlösefähigkeiten, die die Logik und den Prozess einer wissenschaftlichen Untersuchung steuern. Ausgeschlossen werden aus dem Konstrukt alle metakognitiven Ansichten zum Wissenschaftsbegriff (also Einstellungen zu aktuell vorherrschenden Paradigmen, gesellschaftliche und politische Folgen der Untersuchungen und Ähnliches), ebenso wie alle manuellen und praktischen Fähigkeiten, die für die tatsächliche Durchführung einer geplanten Untersuchung notwendig sind.

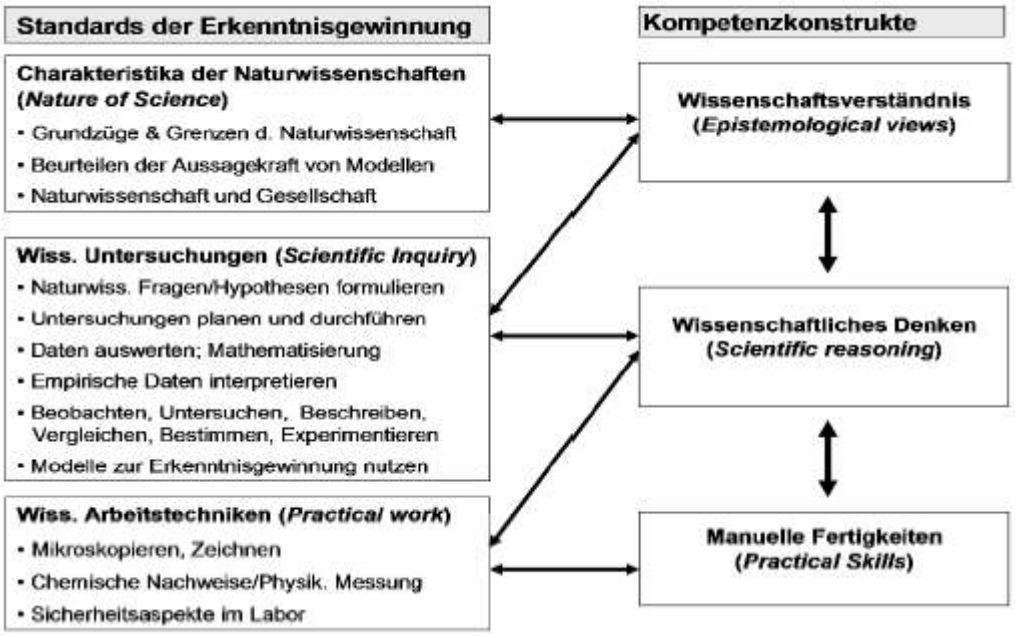

Abbildung 1: Kompetenzen naturwissenschaftlicher Erkenntnisgewinnung nach Mayer (2007).

Daneben grenzt Mayer (2007) die prozesssteuernden Fähigkeiten noch extra gegen allgemeine kognitive Fähigkeiten und gegen deklaratives Wissen ab (Abbildung 2). Die Unterscheidung zwischen Kompetenzen und Intelligenz deckt sich dabei mit der im vorigen Abschnitt (2.1) erläuterten allgemeinen Definition von Kompetenzen nach Hartig und Klieme (2006).

Der Ausschluss von deklarativem Wissen über spezifische wissenschaftliche Methoden ist weniger offensichtlich, grenzt die zu messende Kompetenz aber noch einmal in einem wichtigen Bereich ein. Zur Kompetenz naturwissenschaftlichen Denkens gehört damit nicht das grundlegende Wissen über einen vorhandenen Satz an Methoden und Werkzeugen der Erkenntnisgewinnung, sondern allein die problembezogene Auswahl, Anpassung oder kreative Neugestaltung eines dieser Werkzeuge und eines zugehörigen Untersuchungsplans (Mayer, 2007).

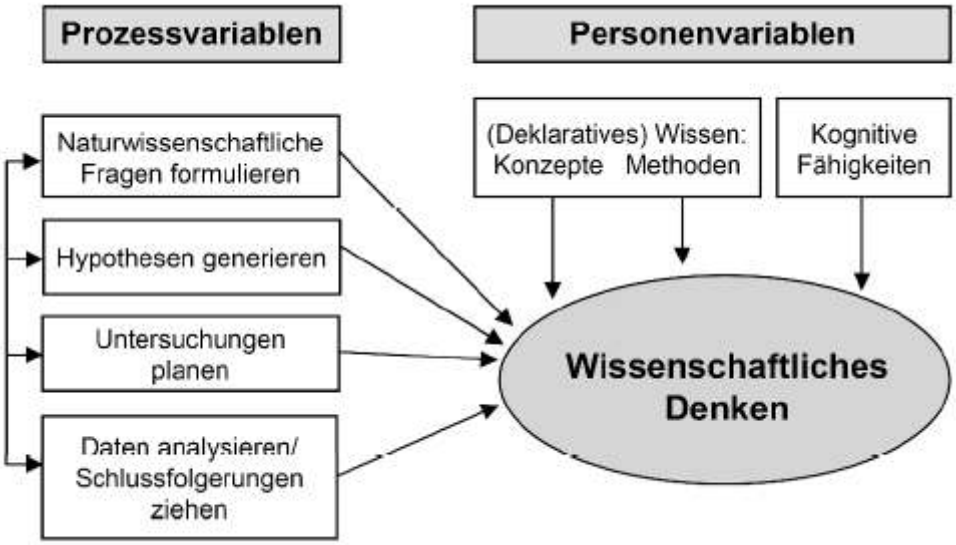

Abbildung 2: Abgrenzung der Kompetenz naturwissenschaftlichen Denkens gegen deklaratives Methodenwissen und Intelligenz nach Mayer (2007).

Im Modell von Mayer (2007) werden verschiedene Handlungsschritte aufgelistet, die bei der Erkenntnisgewinnung durch eine praktische Untersuchung, beispielsweise ein Experiment, eine Rolle spielen. Was dabei nicht berücksichtigt wird, ist die Erkenntnisgewinnung durch die Arbeit mit Modellen, die allerdings eine durchaus wichtige Rolle spielt.

Aus diesem Grund wurde das Kompetenzmodell im Projekt Ko-WADiS erweitert. Dabei wurde auf die Beschreibung von Modellkompetenz durch Upmeier zu Belzen und Krüger (2010) zurückgegriffen (Abbildung 3). Darin wird unterschieden zwischen dem Kompetenzbereich *Kenntnisse über Modelle*, der analog zum Bereich *Nature of Science* in Mayers Modell (Abbildung 1) das Wissenschaftsverständnis repräsentiert, und dem Kompetenzbereich *Modellbildung*, der den Prozess wissenschaftlichen Denkens sowie praktische Fähigkeiten umfasst. Zusammen umfassen beide Bereiche den gesamten Prozess der Erkenntnisgewinnung anhand von wissenschaftlichen Modellen[4].

[4] Der Begriff Modell bezeichnet in diesem Zusammenhang abstrahierte und gezielt konstruierte Repräsentationen von Objekten oder beobachtbaren Phänomenen Upmeier zu Belzen und Krüger (2010).

science education				Quellen
learning science	learning about science	doing science		Hodson (1992)
learn scientific models	nature of models	act of modelling		Henze et al. (2007)
Modellwissen	Modellverständnis	Modellarbeit		Meisert (2008)
	Erkenntnisgewinnung			
	Wissenschafts-verständnis	Wissenschaftliches Denken	Manuelle Fertigkeiten	Mayer (2007)
	Kenntnisse über Modelle	Modellbildung		Upmeier zu Belzen & Krüger

Abbildung 3: Verortung von Erkenntnisgewinnung und Modellkompetenz nach Upmeier zu Belzen und Krüger (2010).

Ebenso wie aus dem ersten Modell wurde auch hier nur der Kompetenzbereich des Denkprozesses selbst berücksichtigt, weshalb lediglich die drei dort dargestellten Facetten der Modellbildung in das Ko-WADiS-Modell aufgenommen wurden. Durch das Zusammenführen der relevanten Bestandteile aus beiden Modellen entstand das in Tabelle 1 dargestellte Strukturmodell.

Da sich die konkrete Durchführung der verschiedenen Handlungsfacetten des Modells innerhalb der Fächer Biologie, Chemie und Physik nicht wesentlich unterscheidet (relevant ist hier nur der Untersuchungsgegenstand, vgl. Straube, 2016), wird dieses Kompetenzmodell als für die drei Naturwissenschaften fachübergreifend angesehen.

Fasst man Abschnitt 2.1 und 2.2 zusammen, kann die Kompetenz naturwissenschaftlichen Denkens wie folgt beschrieben werden. Es handelt sich dabei um eine

a) kognitive,
b) für die drei Naturwissenschaften übergreifende aber
c) bereichsspezifische sowie
d) gezielt erlernbare

Fähigkeit.

Sie umfasst weder

e) affektive, emotionale oder motivationale Aspekte, noch
f) deklaratives Wissen jedweder Art, noch
g) manuelle Experimentierfähigkeiten oder
h) individuelle Ansichten und Einstellungen zur Nature of Science.

In einer konkreten Problemstellung zeigt sich naturwissenschaftliches Denken in der erfolgreichen Durchführung der in Tabelle 1 dargestellten Facetten.

2.3 Vorstellung des Testformats

Bei der Testentwicklung wurde beschlossen, die modellierte Kompetenz in einem linearen Papiertest sowie durch ein geschlossenes Multiple-Choice-Aufgabenformat zu messen. Grund für diese Auswahl war einerseits die hohe Objektivität bei der Auswertung dieser Art von Aufgaben. Andererseits handelte es sich dabei um das in Betracht gezogene Format mit dem geringsten Zeitaufwand pro Aufgabe innerhalb einer Messung, womit die Abdeckung aller Facetten (Tabelle 1) und verschiedener Fachkontexte in einer Befragung ermöglicht wurde (Hartmann, Mathesius et al., 2015).

Für das Testinstrument wurden pro Facette und Fachbereich (Biologie, Chemie, Physik) mehrere Aufgaben entwickelt und pilotiert. Bei der Aufgabenentwicklung wurden mehrere Pilotierungsphasen durchlaufen, in denen unter Einbezug von Expert*inneninterviews aus zunächst offenen Aufgaben die später geschlossenen Formate inklusive der Antwortmöglichkeiten entstanden. Dieses Vorgehen wurde gewählt, um die Validität des Testeinsatzes aus inhaltlichen Gesichtspunkten zu gewährleisten (für eine detaillierte Beschreibung der gesamten Testentwicklung siehe Straube, 2016). Insgesamt entstand so ein Pool aus 123 Aufgaben. Diese zeichnen sich durch ein einheitliches Format aus (siehe Abbildung 4).

Zunächst wird die Aufgabe im Itemstamm in einen Sachkontext eingebettet, der neben der eigentlichen Problemstellung auch alle notwendigen Informationen präsentiert. Dieser Itemstamm wurde – soweit möglich – ohne die Verwendung von spezifischen Fachbegriffen erstellt, damit die Aufgaben ohne Vorwissen im jeweiligen Kontext bearbeitet werden können.

Gegebenenfalls notwendige Konzepte werden ansonsten im Itemstamm selbst erläutert.

Auf den Itemstamm folgt ein vereinheitlichter Impuls. Dieser ist für alle Aufgaben aus einer der sieben Facetten jeweils gleich und unabhängig von der fachlichen Ausrichtung des Itemstamms formuliert.

Auf den Impuls folgen dann wiederum vier Antwortmöglichkeiten. Drei davon sind falsche Antworten und stellen Distraktoren dar, die vierte Antwort ist korrekt. Die Information, dass pro Aufgabe immer nur exakt eine richtige Lösungsmöglichkeit besteht, wird zum Anfang des Tests gegeben und ist den Proband*innen somit bekannt.

Für die Untersuchung der Kompetenz wurden aus dem Pool von Aufgaben mehrere Testhefte erstellt, da nicht alle 123 Aufgaben in einer den Versuchspersonen zumutbaren Zeit bearbeitet werden können. In jedem dieser Hefte ist jeweils eine Aufgabe pro Facette und Fachbereich enthalten, insgesamt also 21 Aufgaben pro Heft. Die Aufgaben überschneiden sich zwischen Testheften teilweise in Ankerblöcken, sodass eine gemeinsame Auswertung ermöglicht wird. Den Proband*innen wurde bei jeder Messung eines dieser Hefte zufällig vorgelegt.

Durch diese Form der Befragung konnten alle 123 Aufgaben angewendet werden ohne eine Überforderung der Proband*innen zu riskieren. Daneben vermindert das Design die Gefahr von Erinnerungseffekten bei mehrfacher Befragung derselben Proband*innen.

Meeresspiegel

Durch die globale Erwärmung ist weltweit ein Rückgang permanenter Eisvorkommen an Land und im Wasser feststellbar.

Mit einem Modellversuch können die Auswirkungen des schmelzenden Eises auf den Meeresspiegel modelliert werden (siehe Abbildung). Dazu wird ein Zylinder in ein Becherglas gestellt. Auf den Zylinder wird ein Eiswürfel gelegt. In ein anderes Becherglas wird nur ein Eiswürfel gelegt. Beide Bechergläser werden mit der gleichen Menge Wasser befüllt. Im rechten Becher ragt der Zylinder noch aus dem Wasser heraus. Schmelzen nun beide Eiswürfel, steigt der Wasserstand in dem Becherglas mit dem Zylinder an, in dem anderen bleibt er gleich.

Abbildung. Modellversuch zum Schmelzen von Eisvorkommen.

(Foto: Helmuth Grötzebauch, FU-Berlin)

Wie kann man die Gültigkeit des Modellversuchs überprüfen? Kreuzen Sie an.

- ☐ Man überprüft, welcher der beiden Eiswürfel im Modellversuch schneller schmilzt.
- ☐ Man leitet Prognosen über den Wasserspiegel aus dem Modellversuch ab und prüft diese in der Natur.
- ☐ Man ersetzt das Süßwasser durch Salzwasser und wiederholt den Modellversuch.
- ☐ Man prüft, ob sich die Wasserspiegel in Modellversuch und Natur ändern.

Abbildung 4: Beispielaufgabe "PT_Meeresspiegel_021" aus den Bereichen "Physik" und "Testen von Modellen".

2.4 Ausgewählte Ergebnisse im Fach Physik

In diesem Abschnitt werden die Erkenntnisse zusammengefasst, die aus bisherigen Studien mit dem Testinstruments gewonnen wurden. Bisher wurde der Ko-WADiS-Test vor allem in den Projekten Ko-WADiS und ValiDiS eingesetzt.

Ko-WADiS fand im Zeitraum von 2011 bis 2015 im Rahmen der BMBF-Förderinitiative KoKoHs statt. Zu den zentralen Zielen des Projekts Ko-WADiS gehörte der Vergleich von Kompetenzständen im Bereich der Erkenntnisgewinnung von Studierenden mit und ohne Lehramtsoption sowie die langfristige Beobachtung von Kompetenzentwicklungen dieser Gruppen (Hartmann, Upmeier zu Belzen & Krüger, 2015).

Zu diesem Zweck wurde das in Abschnitt 2.2 dargestellte Kompetenzmodell entwickelt sowie das gesamte Testinstrument von Grund auf konstruiert und pilotiert. Danach wurden Langzeitbeobachtungen mehrerer Kohorten von Studierenden begonnen. In diesen Kohorten waren die Fächergruppen Biologie, Chemie und Physik vertreten, innerhalb jeder Fachgruppe wurden sowohl Monostudierende als auch Studierende des Lehramts im Erst- oder Zweitfach befragt (Hartmann, Mathesius et al., 2015; Straube, 2016).

Das Projekt ValiDiS baute im Zeitraum von 2016 bis 2019 auf den Ergebnissen von Ko-WADiS auf. Das Hauptanliegen dieses Projekts war die Absicherung der validen Auslegung der Testergebnisse als Maß für die Kompetenz naturwissenschaftlichen Denkens. Zudem wurden die in Ko-WADiS begonnenen Langzeitstudien fortgesetzt, bis die beobachteten Studierendenkohorten ihre Regelstudienzeit absolviert hatten (Krüger, Upmeier zu Belzen & Hartmann, 2016).

Im Folgenden werden die wichtigsten Befunde aus beiden Projekten zusammengefasst, die zum aktuellen Zeitpunkt veröffentlicht wurden.

In einem Vergleich verschiedener mathematischer Modelle zur Datenbeschreibung wurde die Dimensionalität des Kompetenzkonstrukts untersucht. Es bestand die Hypothese, dass sich in der Praxis eine Aufspaltung

der Kompetenz in mehrere Teilbereiche zeigen würde. Dazu wurden Aufspaltungen nach

a) Fach, also Aufgabenkontexten aus Biologie, Chemie und Physik,
b) übergeordnetem Kompetenzbereich, also Untersuchungen und Modellen,
c) sowie den sieben einzelnen Handlungsfacetten

überprüft. Die mathematische Modellierung dieser drei Aufspaltungen wurde durch Modelltests (AIC, BIC und X^2-Test, vgl. Abschnitt 6.3.1) mit einer eindimensionalen Beschreibung der Kompetenz verglichen, in der alle Facetten und Fächer zu einer einzigen, nicht trennbaren Dimension gehören. Die verwendete Stichprobe umfasste N = 2646 Studierende; dabei hatten von den Studierenden 695 Physik, 1663 Biologie und 552 Chemie als mindestens eines ihrer Fächer.

Als Resultat der Vergleiche wurden alle mehrdimensionalen Modelle aufgrund schlechterer Datenpassung (verglichen zum eindimensionalen Modell) verworfen. Es wird also davon ausgegangen, dass es sich bei der beschriebenen Kompetenz um ein eindimensionales und fachübergreifendes Konstrukt handelt (Hartmann, Mathesius et al., 2015). Dieses äußert sich in Untersuchungsprozessen in verschiedenen Handlungsmustern, weshalb die getrennte Beschreibung der sieben Facetten nach wie vor als sinnvoll betrachtet wird.

Anhand der gleichen Stichprobe wurden weitere Untersuchungen zum Kompetenzstand der Physikstudierenden im Vergleich von Teilstichproben vorgenommen. Hierbei zeigten sich folgende Effekte (Straube, 2016):

a) Lehramtsstudierende mit Physik als einem ihrer Wahlfächer wiesen einen signifikant niedrigeren Kompetenzstand auf als Mono-Fachstudierende der Physik (ANOVA: $F(1, 675) = 28.126$, $p < .001$, $\eta_p^2 = .040$, $r = .20$)[5].

[5] Zur Erläuterung des Testberichts sowie der enthaltenen Kennwerte siehe Field (2011).

b) Der mittlere Kompetenzstand der Studierenden war im Master signifikant höher als im Bachelor (ANOVA: $F(2, 675) = 46.630$, $p < .001$, $\eta_P^2 = .121$, $r = .35$).
c) Die Leistungsunterschiede zwischen Lehramts- und Fachstudierenden blieben über die verschiedenen Studienphasen hinweg erhalten (ANOVA: $F(2, 675) = 2.691$, $p > .05$, $\eta_P^2 = .008$).
d) Männliche Studierende wiesen im Mittel einen höheren Kompetenzstand auf (ANOVA: $F(1, 675) = 16.741$, $p < .001$, $\eta_P^2 = .024$, $r = .15$).
e) Lehramtsstudierende mit mehreren naturwissenschaftlichen Fächern zeigten keinen höheren Kompetenzstand als Studierende mit nur einem solchen Fach.

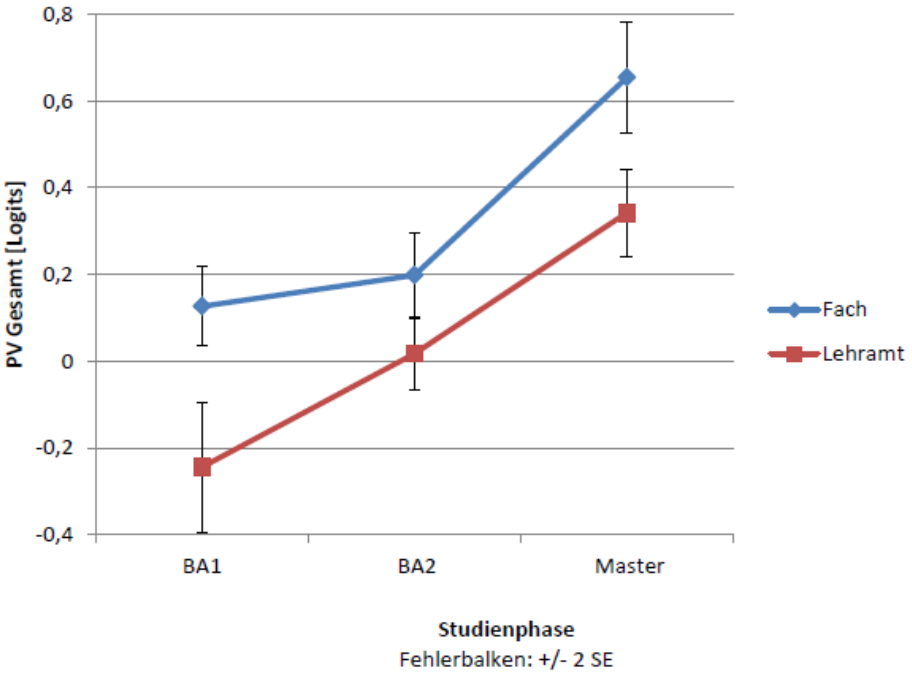

Abbildung 5: Darstellung des Kompetenzstands in verschiedenen Studienphasen und Fachgruppen (Straube, 2016).

Insgesamt liegen aus den durchgeführten Studien etwa 7000 ausgefüllte Testhefte vor (Hartmann, Upmeier zu Belzen & Krüger, 2015). Die theoriekonformen Beobachtungen aus Längsschnittstudien werden im Zusammenspiel mit der ausführlichen Aufgabenkonstruktion als ausreichende Evidenz angesehen, um die Interpretation der Testwerte als Maß für die Kompetenz naturwissenschaftlichen Denkens zu akzeptieren.

3 Item Response Theory

In diesem Kapitel erfolgt die theoretische Rahmung der Item Response Theory (IRT). Sie stellt eine zwingend notwendige Voraussetzung für den Umgang mit adaptiven Testformaten (siehe Kapitel 4) dar. Das bezieht sich nicht nur auf die Erstellung solcher Tests, sondern auch schon auf die Beschreibung ihrer Konstruktion, ihrer Vor- und Nachteile. Es müssen daher zunächst einige Begriffe aus der IRT erläutert werden, bevor im nächsten Kapitel adaptive Tests besprochen werden können.

Andererseits wird mit der Darstellung der IRT auch der Anfang aller methodischen Ausführungen in dieser Arbeit gelegt, da die mathematischen Auswertungen hiermit durchgeführt werden. Somit hat das Kapitel sowohl inhaltlichen als auch methodischen Charakter.

Die ersten beiden Abschnitte dieses Kapitels (3.1 und 3.2) befassen sich mit den Grundannahmen der IRT sowie der dahinterstehenden Denkweise. Danach werden die wichtigsten Modelle der IRT beschrieben und erläutert, wie diese berechnet werden können (Abschnitt 3.3). Den Abschluss des Kapitels bildet Abschnitt 3.4 zur Einschätzung von Personenmerkmalen, also zur Bestimmung der finalen Messwerte aus einer Erhebung.

Alle weiteren Themen, wie etwa die Überprüfung der Güte/Messgenauigkeit einer konkreten Messung, werden in den Praxisteil und die Studienauswertungen (Kapitel 6 bis 8) verlagert. Sie sind nicht zwingend notwendig, um adaptive Verfahren zu beschreiben und würden dieses Kapitel unnötig in die Länge ziehen.

3.1 Testtheorien

Testtheorien stellen einen Teilbereich der psychologischen Messtheorie dar. Sie befassen sich mit den Zusammenhängen zwischen Personenmerkmalen, die gemessen werden sollen, und den aus ihnen resultierenden empirischen Messdaten. Testtheorien bestehen aus einer Reihe von Axiomen über Probandenverhalten und mathematischen Modellen, die den Zusammenhang zwischen Personenmerkmalen und Messdaten beschreiben sollen. Damit wird die notwendige Grundlage geschaffen, um a) psychometrische Messinstrumente zu entwickeln und b) mit diesen gewonnene Daten auswerten zu können (Bühner, 2011; Moosbrugger & Kelava, 2012; Rost, 2004).

Um die Gültigkeit von Schlussfolgerungen aus den gewonnenen Daten beurteilen zu können, liefern Testtheorien zudem notwendige Gütekriterien (Objektivität, Reliabilität und Validität, Döring & Bortz, 2016) und Prüfverfahren.

Die beiden aktuell am weitesten verbreiteten Testtheorien innerhalb der psychologischen, erziehungswissenschaftlichen und didaktischen Forschung stellen die klassische Testtheorie (KTT) und Item Response Theory (IRT) dar.

Die KTT besteht schon deutlich länger als die IRT und ist (unter anderem) deshalb besser und weiter etabliert (Moosbrugger & Kelava, 2012). Das gilt auch in Forschungszweigen wie den Didaktiken, da hier nicht aktiv an Methoden und Testkonstruktion geforscht wird – sie stellen hier lediglich ein Werkzeug dar. Daher soll die (den meisten Lesern vertrautere) KTT als Ausgangspunkt der Betrachtung dienen.

In der KTT wird davon ausgegangen, dass zwischen Merkmalen einer Person und den dadurch gesteuerten Handlungen ein deterministischer Zusammenhang besteht (Bortz & Döring, 2006). Wenn also eine Person grundsätzlich fähig genug ist, eine bestimmte Aufgabe zu lösen (zum Beispiel in einem Mathematiktest), dann ist die Annahme, dass sie diese Aufgabe grundsätzlich immer richtig lösen wird. Als Ursache für ein

eventuelles Scheitern kommen in der KTT Störvariablen in Frage, in unserem Beispiel also etwa Unterbrechungen, ablenkender Lärm oder Ähnliches.

Mathematisch betrachtet lässt sich daher in der KTT jede Messung als die Summe aus einem wahren, unverzerrten Messwert und einem zufälligen Messfehler, bedingt durch nicht gemessene und unkontrollierte Störfaktoren, darstellen (Bortz & Döring, 2006; Moosbrugger & Kelava, 2012). Dieser Messfehler ist also möglichst zu minimieren. Auf der einen Seite bedeutet dies, bei der Testkonstruktion neben dem angesteuerten Merkmal keine weiteren wichtigen Faktoren (wie z. B. Textverständnis) in die Items einzubauen. Auf der anderen Seite bedeutet dies, Störfaktoren zu minimieren oder wenigstens bei der Messung zu dokumentieren. Ist der Fehler im Sinne einer konkreten Messung klein genug, so wird der Messwert häufig als Ausprägung des Personenmerkmals interpretiert (Brennan, 2010).

In der IRT wird eine andere Denkweise verfolgt. Die Annahme ist hier, dass zwischen dem Personenmerkmal und Handlungen beziehungsweise Antwortmustern in einem Test kein deterministischer, sondern ein probabilistischer[6] Zusammenhang besteht (Bortz & Döring, 2006; Moosbrugger & Kelava, 2012). Je besser also eine Person rechnen kann, desto höher ist die Chance für sie, Rechenaufgaben korrekt zu lösen (Bortz & Döring, 2006). Es wird aber nur von Wahrscheinlichkeiten gesprochen: Eine sehr fähige Person kann, nach der IRT betrachtet, selbst in Idealsituationen an einer leichten Aufgabe scheitern. Die Wahrscheinlichkeit ist jedoch sehr gering.

Mathematische Modelle der IRT bestehen dementsprechend immer aus einem Wahrscheinlichkeitsmaß, das einen Zusammenhang zwischen Merkmalen, Aufgaben und Lösungswahrscheinlichkeit herstellt, siehe dazu Abschnitt 3.3. Im Gegensatz zur KTT ist kein Fehlermaß in den

[6] Es findet sich daher in der Literatur auch die Bezeichnung *probabilistische Testtheorie* (PTT) (Moosbrugger & Kelava, 2012) als Alternative zur IRT. Daneben existieren noch weitere Bezeichnungen wie *latent trait theory* (Weiss, 1983). Im weiteren Verlauf wird als Bezeichnung IRT verwendet, da diese am häufigsten zu finden ist. Mitunter wird allgemein von einer „Raschanalyse“ gesprochen, siehe dazu auch Abschnitt 3.3.4.

Modellgleichungen enthalten. Stattdessen wird für die Beurteilung einer Messung der Fokus auf die Passung von Modell und Daten sowie auf das Konzept der Testinformation gelegt, beides wird in späteren Abschnitten (siehe 6.2 und 6.3) betrachtet.

In Teilen der in diesem Kapitel zitierten Literatur werden die beiden Theorien verglichen oder als alternative Herangehensweisen an die grundsätzliche Problematik der Testkonstruktion und Testauswertung betrachtet. Es wäre passender, die IRT als eine Weiterentwicklung der KTT zu betrachten (Moosbrugger & Kelava, 2012). Sie liefert mathematische Werkzeuge, um einige der Schwächen der KTT zu beheben. So ist es innerhalb der KTT beispielsweise nicht möglich, die axiomatischen Annahmen auf ihre Gültigkeit zu überprüfen (Bortz & Döring, 2006; Moosbrugger & Kelava, 2012). Zudem ist die Beurteilung von Aufgaben an die hierfür verwendete Stichprobe gebunden (Moosbrugger & Kelava, 2012). Durch Anwendung der IRT können diese Probleme umgangen werden (Amelang & Zielinski, 1997; Moosbrugger & Kelava, 2012). Weiterhin erlaubt die IRT eine getrennte Betrachtung von Eigenschaften der Testitems und der Probanden, was für die Entwicklung von Instrumenten interessant sein kann und den Einsatz adaptiver Testverfahren (siehe Kapitel 4) erst ermöglicht. Gleichzeitig stellt die IRT jedoch höhere Anforderungen an die Testentwicklung (Moosbrugger & Kelava, 2012). Bei der Wahl der geeigneten Testtheorie muss also zwischen Anforderungen an die Testinstrumente und dem dafür gerechtfertigten Aufwand abgewogen werden.

3.2 Grundannahmen der IRT

Die Item Response Theory baut im Allgemeinen auf zwei Grundannahmen auf, die unabhängig vom gewählten Modell[7] erfüllt sein müssen.

Die erste besagt, dass das Antwortverhalten einer Person durch *latente Merkmale* (latent trait) gesteuert wird, die nicht direkt beobachtet werden können (Bortz & Döring, 2006). Auf der einen Seite handelt es sich dabei um verschiedene Merkmale der Person selbst, wie zum Beispiel eine Fähigkeit, Kompetenz oder Einstellung. Auf der anderen Seite haben auch die bearbeiteten Testitems latente Merkmale, wie zum Beispiel die Schwierigkeit oder die Chance, eine Aufgabe durch Raten korrekt zu lösen. Die latenten Merkmale sind nicht direkt beobachtbar und beeinflussen lediglich die Wahrscheinlichkeit einer bestimmten Antwort (Bortz & Döring, 2006). Die Annahme ist nur indirekt prüfbar, indem die Zusammenhänge zwischen dem manifesten Antwortmuster und den durch das Modell geschätzten latenten Variablen untersucht werden.

Als zweites besteht die Annahme der *lokalen stochastischen Unabhängigkeit* (Liu & Maydeu-Olivares, 2013) für alle Testitems: Solange die Ausprägungen der latenten Probandenmerkmale konstant bleiben, dürfen die Antworten auf die verschiedenen Items des Tests nicht korrelieren. Umgekehrt ausgedrückt sollen die Items also nur die latenten Personenmerkmale und keine weiteren Faktoren messen. Ist die Bedingung der lokalen stochastischen Unabhängigkeit erfüllt, spricht man auch von *Itemhomogenität* (Christensen, Mesbah & Kreiner, 2013). Diese Annahme kann direkt durch das sogenannte *Multiplikationstheorem für unabhängige Ereignisse* (Behnke & Behnke, 2006) überprüft werden.

Weitergehende Beschreibungen der internen Struktur der latenten Variablen können durch das jeweilige mathematische Modell vorgenommen und als zusätzliche Annahmen festgelegt werden (vgl. Reckase, 2009).

[7] Im weiteren Verlauf der Arbeit ist, sofern nicht anders spezifiziert, mit dem Begriff *Modell* stets ein mathematisches Modell aus der IRT gemeint.

3.3 Modelle der IRT

Mit der steigenden Beliebtheit und Verbreitung der IRT wurden auch immer neue Modelle spezifiziert und eingesetzt. Aus der ursprünglich geringen Zahl an Modellen wuchs so eine auf den ersten Blick schwer überschaubare Menge an einzelnen Modellen und ganzen Familien von Funktionen (vgl. Wright & Masters, 1982). Es erscheint deshalb nötig, eine Klassifikation vorzunehmen und einen Überblick zu ermöglichen.

Es bestehen verschiedene Möglichkeiten, IRT-Modelle in Gruppen einzuteilen. Eine beliebte Trennung ist die nach der angenommenen Art der latenten Personenvariablen. Sogenannte *Latent-Class-Modelle* gehen davon aus, dass sich die Personenmerkmale in unterschiedlichen Klassen oder Stufen ausprägen. Sie sind damit für die Analyse von qualitativen Daten geeignet. *Latent-Trait-Modelle* stehen dem gegenüber und setzen eine kontinuierlich veränderbare Ausprägung der Personenmerkmale voraus. Diese Einteilungsmöglichkeit wird beispielsweise von Rost (2004) vorgenommen.

Weitere Einteilungsmöglichkeiten sind die Unterscheidung nach der Menge an möglichen Antworten (dichotome und polytome Modelle) oder nach der angenommenen Dimensionalität der latenten Personenmerkmale. Dabei werden Modelle, in denen mehrdimensionale Merkmale untersucht werden, durch die Bezeichnung *multidimensional item response theory* (MIRT oder M-IRT, vgl. Reckase 2009) gegenüber eindimensionalen Modellen abgegrenzt.

Daneben können innerhalb dieser Gruppierungen noch verschiedene Modellfamilien anhand des funktionalen Zusammenhangs unterschieden werden. Auch wenn logistische Funktionen am häufigsten zu finden sind, können IRT-Modelle durchaus anhand von linearen Funktionen (McDonald, 1999) oder Ogiven (Lord, Novick & Birnbaum, 2008) definiert werden.

Im weiteren Verlauf der Arbeit werden ausschließlich Latent-Class-Modelle aus der Familie der eindimensionalen, dichotomen, logistischen Modelle (kurz als logistische Modelle bezeichnet, da mehrdimensionale oder

polytome Varianten meist gesonderte Bezeichnungen bekommen) verwendet und besprochen. Die Gründe hierfür sind:

1. Im konkreten Anwendungskontext wird von einer eindimensionalen Struktur des Personenmerkmals ausgegangen.
2. Es liegen im verwendeten Testinstrument nur Items mit dichotomem Antwortformat vor.
3. Die Familie der logistischen Modelle wird bereits sehr lange und häufig verwendet, weshalb sie wohl am besten dokumentiert ist.
4. Aus mathematischer Sicht sind diese Modelle vergleichsweise einfach und eignen sich damit gut für die Darstellung von Grundlagen der IRT.

Innerhalb dieser Familie werden die einzelnen Modelle anhand der Menge der berücksichtigten Itemmerkmale oder auch *Parameter* unterschieden. Daraus folgt die Bezeichnung der verschiedenen Modelle als einparametrisch (1pl), zweiparametrisch (2pl) und so weiter.

3.3.1 Das einparametrische logistische Modell

Das 1pl-Modell berücksichtigt neben der Merkmalsausprägung[8] der Proband*innen nur einen Parameter, und zwar die *Itemschwierigkeit* beziehungsweise den *Schwierigkeitsparameter* des Items. Die zugehörige Modellgleichung kann wie folgt geschrieben werden:

$$P(\theta_i, b_j) = \frac{e^{(\Theta_i - b_j)}}{1 + e^{(\Theta_i - b_j)}} = \frac{1}{e^{(b_j - \Theta_i)} - 1}$$

$P(\Theta_i, b_j)$ stellt die Wahrscheinlichkeit der richtigen Lösung $x_{ij} = 1$ der Aufgabe j durch Person i dar. Θ_i steht für die Fähigkeit der Person i, b_j ist der Schwierigkeitsparameter der Aufgabe j.

[8] Das in dieser Arbeit behandelte Testinstrument misst eine Fähigkeit, weswegen die Begriffe Fähigkeit und Merkmal im Folgenden synonym verwendet werden. Gleiches gilt für die Begriffe Item und Aufgabe, da es sich bei allen Items im Testinstrument um Aufgaben handelt.

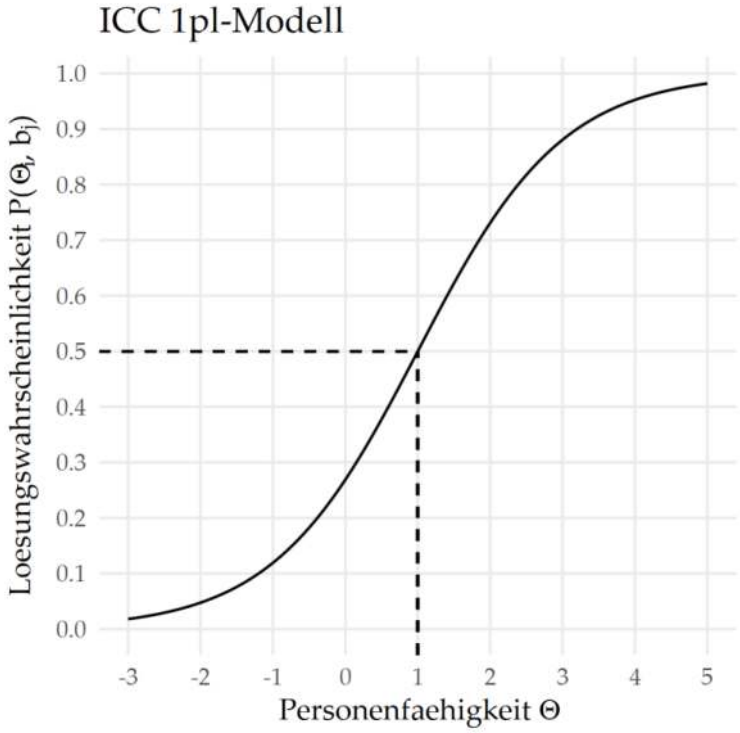

Abbildung 6: Itemcharakteristische Kurve (ICC) für ein Item im 1pl-Modell. Dargestellt ist die itemcharakteristische Funktion einer Aufgabe mit Schwierigkeitsparameter b = 1.

Setzt man den Schwierigkeitsparameter für eine gegebene Aufgabe fest in die Gleichung ein, so erhält man die *Itemcharakteristische Funktion* (IC-Funktion). Sie stellt die Lösungswahrscheinlichkeit für das gegebene Item allein in Abhängigkeit der Fähigkeit der bearbeitenden Probanden dar.

$$P(\theta) = \frac{1}{e^{(b-\theta)} - 1}$$

Der Graph der IC-Funktion wird als *Itemcharakteristische Kurve* (ICC, von *Item characteristic curve*) bezeichnet. Sie ist in Abbildung 6 dargestellt. Anhand der Darstellung lassen sich mehrere interessante Eigenschaften der Funktion beobachten. Zunächst ist festzustellen, dass die Funktion für alle Werte der Parameter Θ und b lösbar ist. Somit können auch extreme Ausprägungen der latenten Variablen im Modell berücksichtigt werden, ohne auf mathematische Komplikationen zu stoßen. Die Funktionswerte liegen für alle diese Ausprägungen im Bereich von 0 bis 1, womit diese ohne weitere Umrechnung direkt als Wahrscheinlichkeit interpretiert werden können.

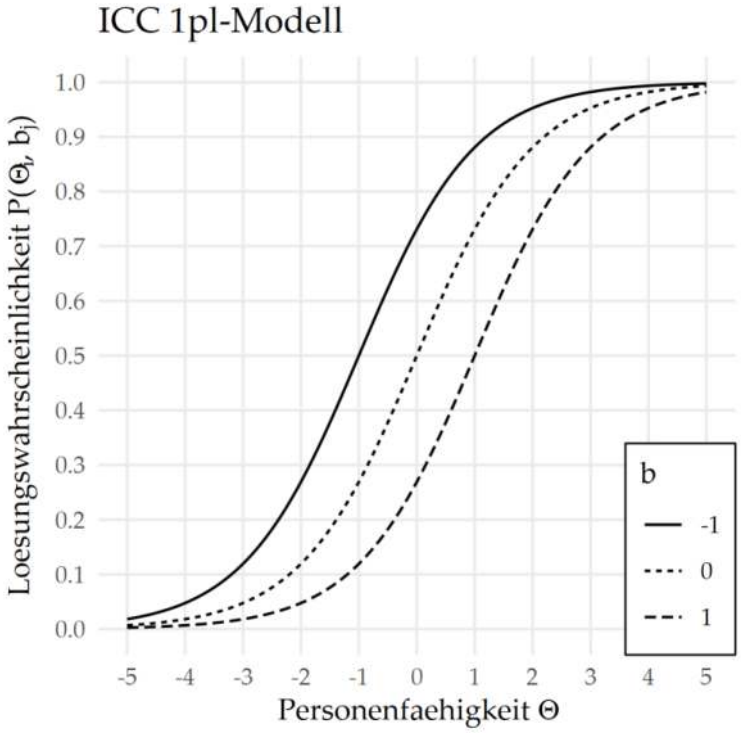

Abbildung 7: ICCs dreier 1pl-konformer Items mit unterschiedlichen Schwierigkeitsparametern b (siehe Legende)

Weiterhin ist zu erkennen, dass Schwierigkeitsparameter und Personenfähigkeit auf derselben Skala abgetragen werden, siehe hierzu auch Abbildung 7. Die gemeinsame Einheit der Werte ist der *Logit* (die logarithmierte Chance). Sind Personenfähigkeit und Itemschwierigkeit identisch, so liegt die Wahrscheinlichkeit einer korrekten Lösung bei exakt 50%.

Zuletzt sind zwei Anmerkungen zur Steigung der IC-Funktionen zu machen. Erstens ist Abbildung 7 zu entnehmen, dass die Steigung von IC-Funktionen von Aufgaben verschiedener Schwierigkeiten identisch ist, die Funktionen verlaufen parallel. Zweitens ist die Steigung einer einzelnen IC-Funktion nicht konstant. Liegt die Fähigkeit einer Person weit unter der Schwierigkeit einer gegebenen Aufgabe, ist die Lösungswahrscheinlichkeit nur knapp über 0%. Kleine Steigerungen der Fähigkeit haben nur geringe Änderungen der Lösungswahrscheinlichkeit zu Folge. Dieser Fakt kann so interpretiert werden, dass die Aufgabe nach wie vor viel zu schwierig ist. Je näher sich beide Parameter kommen, desto größer wird die Steigung. Ist die Aufgabe nur ein klein wenig zu schwer, so würde eine Fähigkeitssteigerung (oder ein Lernzuwachs, um auf das frühere Beispiel

der Rechenaufgaben zurückzugreifen) schnell starke Änderungen in der Lösungswahrscheinlichkeit hervorrufen. Bei einer Chance von 50%, also identischen Parametern, liegt der Wendepunkt der Funktion – ab hier sinkt die Steigung für wachsende Fähigkeitsparameter. Je leichter die Aufgabe für die Person wird, desto weniger stark beeinflussen weitere Lernfortschritte die Lösungswahrscheinlichkeit. Ab einer gewissen Fähigkeit liegt die Lösungschance bei über 99%, die Aufgabe ist quasi trivial geworden.

Im Zusammenhang mit der letzten Bemerkung kann folgende Feststellung gemacht werden: Für den Vergleich zweier Personen mit unterschiedlicher Fähigkeit sind am besten solche Aufgaben geeignet, deren Schwierigkeit in der Nähe der beiden Personenfähigkeiten ist.

Legen wir beiden Personen eine Reihe von Aufgaben vor, die für beide sehr leicht/sehr schwer sind, besteht bei jeder Aufgabe nur ein kleiner Unterschied in der Lösungswahrscheinlichkeit. Dementsprechend müssen viele Aufgaben vorgelegt werden, bis sich der Unterschied eindeutig aus den gewonnenen Daten ergibt und eine Unterscheidung zwischen beiden Personen sicher vorgenommen werden kann. Der Informationsgewinn jeder einzelnen Aufgabe ist also in diesem Fall klein.

Befinden sich die Aufgaben jedoch im Bereich der Fähigkeiten beider Personen, zeigt sich ein anderes Bild. Die Steigung der ICC ist hier deutlich höher, womit auch der Unterschied zwischen den Lösungswahrscheinlichkeiten unserer Testpersonen weitaus größer ist. Es müssen deshalb weniger Aufgaben bearbeitet werden, bis sich ein eindeutiger Unterschied zeigt. Der Informationsgewinn durch jede einzelne Aufgabe ist höher.

Es gibt in der Statistik verschiedene Möglichkeiten, um die durch Zufallsereignisse gewonnene Informationsmenge über eine unabhängige Variable abzuschätzen. Durchgesetzt hat sich in der IRT die Verwendung der Fisher-Information, benannt nach dem britischen Statistiker Sir Ronald Aylmer Fisher (Baker & Kim, 2017).

Für die Information gilt:

$$I = \frac{1}{\sigma^2}$$

wobei I die Information und σ^2 die Varianz des geschätzten Parameters ist.

Für das 1pl-Modell kann der Informationsgewinn einer Aufgabe mit Schwierigkeit b in Abhängigkeit der Personenfähigkeit θ wie folgt berechnet werden (Baker & Kim, 2017):

$$\begin{aligned} I(\theta) &= PQ \\ &= P(1-P) \\ &= \frac{e^{(\theta-b)}}{(1+e^{(\theta-b)})^2} \end{aligned}$$

wobei $Q = 1 - P$ die Gegenwahrscheinlichkeit beziehungsweise die Wahrscheinlichkeit einer falschen Lösung bezeichnet. Durch die graphische Darstellung des Zusammenhangs erhält man die *Informationskurve* der Aufgabe (vgl. Abbildung 8). Anschaulich kann die Fisher-Information als eine Art diagnostischer Empfindlichkeit der Aufgabe betrachtet werden. Sie gibt also an, wie sensibel die Aufgabe an verschiedenen Stellen auf Unterschiede oder Änderungen der Personenfähigkeit reagiert.

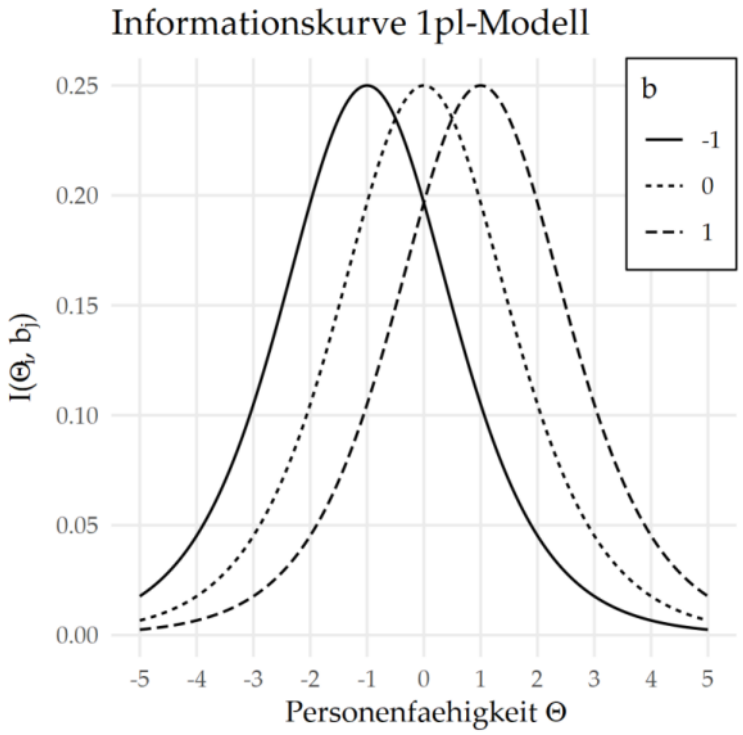

Abbildung 8: Informationskurven dreier 1pl-konformer Aufgaben mit unterschiedlichen Schwierigkeitsparametern (siehe Legende).

Wie leicht zu erkennen ist, erreicht der Informationsgewinn durch eine Aufgabe sein Maximum von 0.25, wenn Fähigkeit der Proband*innen und Schwierigkeit der Aufgabe exakt übereinstimmen. Das ist gleichzeitig die Stelle, an der Lösungswahrscheinlichkeit und Gegenwahrscheinlichkeit jeweils 50% betragen. Die Information, wie oben zu sehen das Produkt beider Wahrscheinlichkeiten, erreicht daher im Maximum einen Wert von 0,25. Bedeutung kommen der Testinformation im adaptiven Testen (Kapitel 4) sowie in der Beurteilung der Messgenauigkeit (Abschnitt 6.3.2) zu.

3.3.2 Das zweiparametrische logistische Modell

Das 2pl-Modell, nach dem amerikanischen Mathematiker Alan Birnbaum auch als Birnbaum-Modell benannt (Birnbaum, 2008; Moosbrugger, 2012), ergänzt das 1pl-Modell um den sogenannten *Diskriminationsparameter*. Es wird mathematisch durch folgende Gleichung ausgedrückt:

$$P(\Theta_i, b_j, a_j) = \frac{e^{a_j\,(\Theta_i - b_j)}}{e^{a_j\,(\Theta_i - b_j)} + 1}$$

Wobei a_j der Diskriminationsparameter des Items j ist, die restlichen Bezeichnungen sind identisch zum 1pl-Modell. Die grundlegenden Eigenschaften der IC-Funktion sind identisch zu der im 1pl-Modell, mit einer Ausnahme: Der Diskriminationsparameter stellt mathematisch betrachtet die Steigung der Kurve im Wendepunkt dar, siehe Abbildung 9.

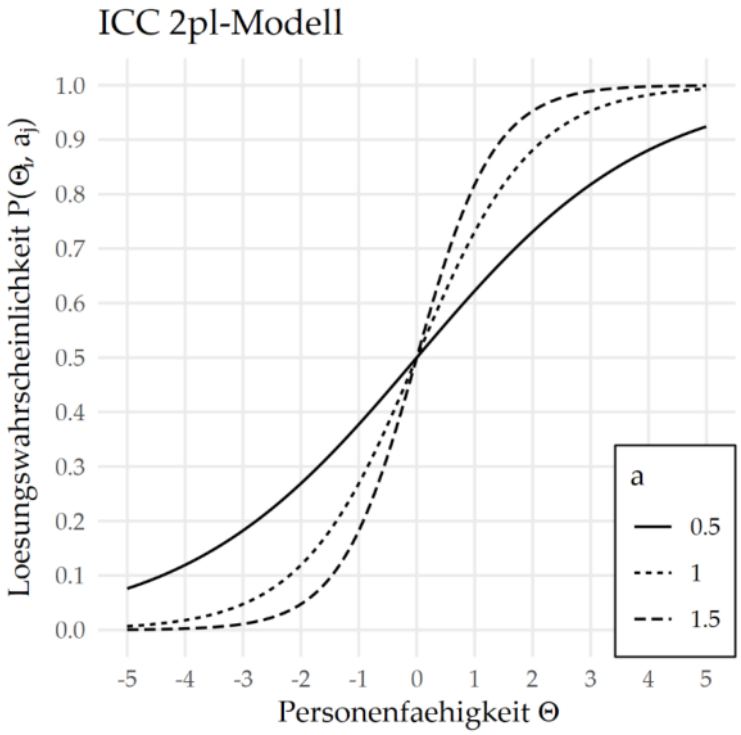

Abbildung 9: ICCs dreier 2pl-konformer Items mit unterschiedlichen Diskriminationsparametern (siehe Legende). Alle Items haben den Schwierigkeitsparameter b = 0.

Inhaltlich interpretiert kann durch die Variation des Diskriminationsparameters modelliert werden, wie stark die Lösungswahrscheinlichkeit einer Aufgabe durch Änderungen der Personenfähigkeit verändert wird. Ist die Diskrimination sehr hoch, führen kleine Änderungen der Fähigkeit im Bereich um die Aufgabenschwierigkeit zu starken Veränderungen der Lösungschance. Im Gegenzug wird bei steigendem Abstand zur Aufgabenschwierigkeit schnell ein Punkt erreicht, ab dem Änderungen der Fähigkeit kaum noch messbare Unterschiede ausmachen. Ist die Diskrimination hingegen sehr klein, flacht die gesamte Kurve ab. Auch bei großen Abständen zwischen Fähigkeit und Schwierigkeit können noch Änderungen in der Lösungschance festgestellt werden. Bei flachen Kurven ist diese Änderung

aber kaum noch abhängig von der Fähigkeit der Person, was eine Einschätzung deutlich erschwert. Der Diskriminationsparameter bestimmt also, wie gut unterschiedlich fähige Personen anhand der Aufgaben unterschieden, also diskriminiert werden können. Er kann analog zur Trennschärfe von Aufgaben innerhalb der KTT betrachtet werden.

Sowohl anhand der inhaltlichen Interpretation als auch der mathematischen Bedeutung als Steigung der ICC erscheint es nur sinnvoll, dass der Diskriminationsparameter Einfluss auf die Informationskurve von Aufgaben hat. Diese ist für das 2pl-Modell wie folgt definiert (Baker & Kim, 2017):

$$I(\theta) = a^2 PQ$$
$$= a^2 \frac{e^{a(\theta-b)}}{(1+e^{a(\theta-b)})^2}$$

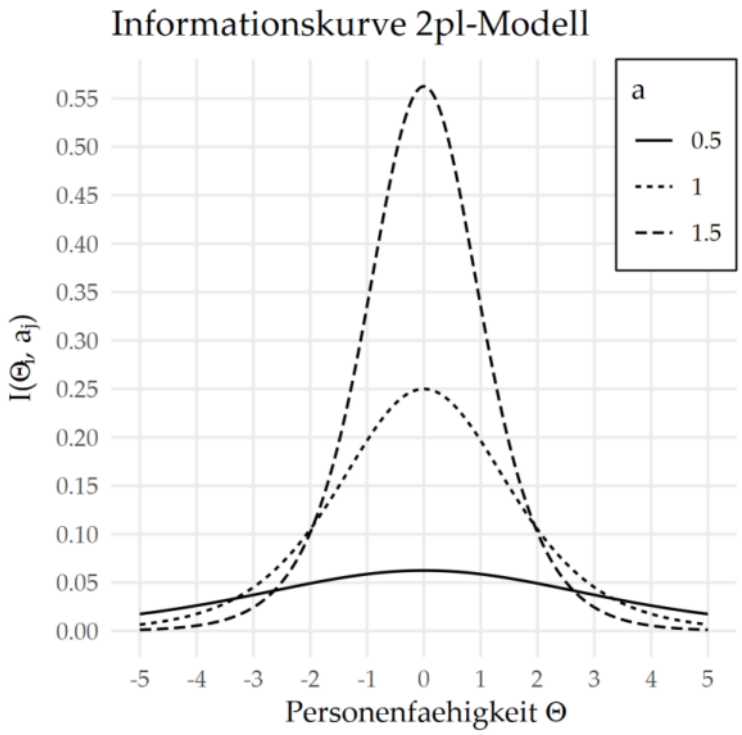

Abbildung 10: Informationskurven dreier 2pl-konformer Items mit unterschiedlichen Diskriminationsparametern (siehe Legende). Alle Items haben den Schwierigkeitsparameter b = 0.

Genau wie beim 1pl-Modell bestimmt der Schwierigkeitsparameter einer Aufgabe die Position des maximal möglichen Informationsgewinns. Der Einfluss des Diskriminationsparameters zeigt sich in einer Streckung oder

Stauchung der Informationskurve. Er verändert damit ihre Steigung und den Wert des maximalen Informationsgewinn. Die anschauliche Betrachtung wäre hier: Ist die ICC einer Aufgabe flach, ihre Steigung in verschiedenen Fähigkeitsbereichen also kaum unterschiedlich, ist der mögliche Informationsgewinn auch kaum verschieden. Zudem ist er vergleichsweise niedrig, da die ICC eben nirgends eine hohe Steigung aufweist und geringe Fähigkeitsunterschiede auch nur zu marginal unterschiedlichen Lösungschancen führen. Bei einem hohen Diskriminationsparameter hingegen ist die Steigung im direkten Umfeld der Aufgabenschwierigkeit sehr groß. Unterschiede zwischen Personenfähigkeiten zeigen sich hier schon nach wenigen Aufgabenbearbeitungen deutlich in den Antwortmustern, die gewonnene Information ist daher sehr hoch. Der mögliche Informationsgewinn fällt dafür aber sehr schnell ab, wenn dieser Bereich verlassen wird und die Fähigkeiten von Proband*innen stark von der Aufgabe abweichen.

3.3.3 Das dreiparametrische logistische Modell

Auch das 3pl-Modell kann auf Alan Birnbaum zurückgeführt werden (Lord et al., 2008). Es ergänzt das 2pl-Modell um den *Rateparameter* und wird aus diesen Gründen auch das Rate-Modell von Birnbaum genannt. Um Verwechslungen zwischen den Modellen zu vermeiden, werden im weiteren Verlauf die Bezeichnungen 1pl, 2pl und 3pl verwendet. Mathematisch wird das 3pl-Modell durch folgende Funktion ausgedrückt:

$$P(\Theta_i, a_j, b_j, c_j) = c_j + \frac{(1 - c_j)}{e^{a_j(b_j - \Theta_i)} - 1}$$

Hierbei steht c_j für den Rateparameter des Items j, alle weiteren Bezeichnungen sind identisch zum 2pl- beziehungsweise 1pl-Modell. Das Verhalten der Funktion ist weitgehend identisch zum 2pl-Modell.

Der Rateparameter entspricht inhaltlich betrachtet der Wahrscheinlichkeit, ein Item allein durch Raten zu lösen. Auch für geringe Fähigkeiten kann die Lösungschance im 3pl-Modell nicht unter diesen Wert fallen, er stellt mathematisch also den unteren Grenzwert der Funktionswerte dar (siehe Abbildung 11). Hierdurch wird der gesamte Funktionsverlauf zwischen dem Rateparameter als Minimum und einer hundertprozentigen

Lösungschance als Maximum gestaucht. Dadurch wird auch die Lösungswahrscheinlichkeit im Wendepunkt verändert. Sie beträgt nicht mehr 50%, sondern $50\% + \frac{c}{2}$.

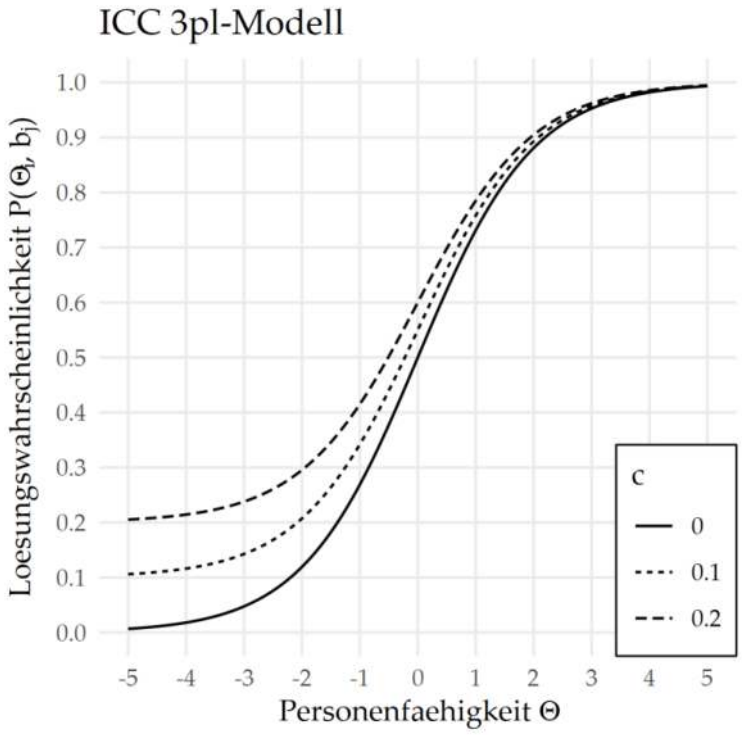

Abbildung 11: ICCs dreier 3pl-konformer Items mit unterschiedlichen Rateparametern (siehe Legende). Alle Items haben eine Schwierigkeit von b = 0 und eine Diskrimination von a = 1.

Die Informationskurve für das 3pl-Modell ist definiert als (Baker & Kim, 2017):

$$I(\theta) = a^2 PQ \left(\frac{P - c}{1 - c}\right)^2$$

Anhand von Abbildung 12 ist eine Eigenschaft der Informationskurve zu erkennen, die spezifisch für das 3pl-Modell und durch den Rateparameter bedingt ist. Größer werdende Rateparameter reduzieren den maximal möglichen sowie wie den insgesamt erreichbaren Informationsgewinn (die Fläche unter der Informationskurve) (Baker & Kim, 2017). Das macht insofern Sinn, als dass bei jeder beobachteten Antwort eine gewisse Unsicherheit durch den Rateparameter einbezogen werden muss, was die Güte der

gewonnenen Information beeinflusst. Wäre die Ratewahrscheinlichkeit im Extremfall 100%, würde keine Antwort an egal welcher Stelle des Fähigkeitsspektrums verwertbare Information liefern.

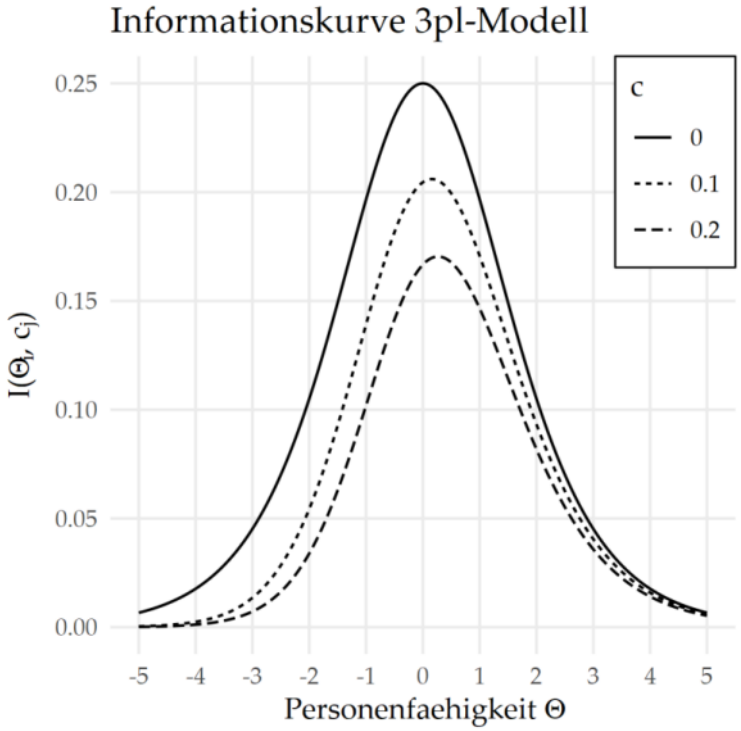

Abbildung 12: Informationskurven dreier 3pl-konformer Items mit unterschiedlichen Rateparametern (siehe Legende). Alle Items haben eine Schwierigkeit von b = 0 und einen Diskriminationsparameter von a = 1.

3.3.4 Exkurs: Raschmodell und 1pl-Modell

Nachdem die notwendigen Begriffe eingeführt wurden, folgt ein kurzer Exkurs zur Verwendung von Begrifflichkeiten im Zusammenhang mit IRT-Analysen.

Das erste IRT-Modell, welches eine logistische Funktion verwendete, wurde von dem dänischen Mathematiker Georg Rasch publiziert (Rasch, 1980):

$$P(x_{ij}) = \frac{e^{x_{ij}(\theta_i - b_j)}}{1 + e^{x_{ij}(\theta_i - b_j)}}$$

Als Name etablierte sich schnell „Rasch-Modell“. Da dieses Modell sowie die gesamten Arbeiten von Rasch wegweisend in der IRT waren, hat sich dieser Name erhalten und eingeprägt. Dies geht sogar so weit, dass des Öfteren "Rasch-Analysen" synonym für die Arbeit mit IRT-Modellen verwendet wird (Boone, Staver & Yale, 2014).

Nach der Publikation von Rasch folgte die Vorstellung eines erweiterten Modells durch Frederic M. Lord. Er schlug vor, das Modell um eine für alle Items konstante Steigungskorrektur von 1,7 zu erweitern (Lord et al., 2008):

$$P(x_{ij}) = \frac{e^{1{,}7 * x_{ij}(\theta_i - b_j)}}{1 + e^{x_{ij}(\theta_i - b_j)}}$$

Hintergrund hierfür war der Versuch, den Funktionsgraphen der logistischen Funktion demjenigen der Ogiven-Verteilung anzugleichen, welche bis dahin das verbreitetste Modell innerhalb der aufkommenden IRT war. Da die Diskrimination für alle Items konstant sein sollte, stellte sie keinen weiteren Parameter dar. Das Modell, für welches Lord den Namen *one parameter logistic model*, also einparametrisches logistisches Modell (1pl) vorschlug, unterscheidet sich also deutlich vom 2pl-Modell, obwohl die Ähnlichkeit sehr groß ist.

Festzuhalten ist, dass die inzwischen oft übliche Bezeichnung des Rasch-Modells als 1pl-Modell technisch nicht ganz korrekt ist. Da das ursprüngliche 1pl-Modell de facto nicht mehr genutzt wird, stellt dies aber kein praktisches Problem dar. Auch in dieser Arbeit wird, um die

Modellbezeichnungen einheitlich zu halten, der Begriff 1pl-Modell verwendet, obwohl eine Diskrimination von eins festgelegt wird und somit eigentlich das Rasch-Modell vorliegt (für eine kompakte Gegenüberstellung beider Modelle siehe Linacre, 2005).

Die bewusste Wortunterscheidung zwischen IRT- und Rasch-Analysen stellt bei vielen Methodiker*innen eher die Positionierung in ein bestimmtes Forschungsparadigma dar: Verfechter*innen der Rasch-Analyse stellen, sofern das Modell nicht auf vorliegende Daten passt, die Frage nach möglichen Fehlern bei der Datenerhebung und Aufgabenkonstruktion. Verfechter*innen der IRT gehen in diesem Fall zu einem anderen Modell über, damit sie die vorhandenen Daten gut beschreiben können (Boone et al., 2014).

3.4 Parameterschätzung

In den vorigen Abschnitten wurde erläutert, in welcher Form IRT-Modelle mathematisch dargestellt werden können, welche theoretischen Annahmen für ihre Gültigkeit wichtig sind und welche Eigenschaften von Testinstrumenten und Proband*innen durch sie untersucht werden können. Daraus ergibt sich allerdings noch keine Antwort auf die Frage, wie die konkreten Werte für die untersuchten Parameter in einer Messung bestimmt, also tatsächlich die Messung und Auswertung durchgeführt werden. Das soll in diesem Abschnitt erläutert werden.

Bei der Auswertung von empirischen Daten anhand von IRT-Modellen ergibt sich die Schwierigkeit, dass keiner der besprochenen Parameter direkt beobachtbar ist. Auf Aufgaben- und Personenseite sind sowohl die einzelnen Werte der Parameter als auch die Eigenschaften der Parameterverteilungen vollständig unbekannt – auch wenn sich teils begründete Vermutungen anstellen lassen.

Für sich gesehen wäre das im Vergleich zu klassischen Auswertungen weder ungewöhnlich noch problematisch. Jedoch lassen sich die Parameter in der IRT auch nicht direkt aus den empirischen Daten berechnen: Jede Antwort entspricht lediglich einer einzelnen Zufallsmessung für die jeweilige Kombination aus Parametern und Modellgleichung. Es kann beobachtet

werden, ob Person x Aufgabe y richtig oder falsch beantwortet hat. Damit ist aber noch keinerlei Information über die Wahrscheinlichkeit dieser Antwort gewonnen. Würde Person x Aufgabe y mehrfach beantworten, könnte man daraus bei ausreichenden Bearbeitungen eine Verteilung konstruieren. Das mehrfache Vorlegen der gleichen Aufgaben innerhalb einer Messung ist jedoch leider nicht praktikabel. Die Wahrscheinlichkeit der Ereignisse ist aus den Messdaten allein so nicht bestimmbar, womit die Modellgleichung unbestimmt bleibt.

Aus diesem Grund ist für alle nicht-deterministischen Modelle die Verwendung numerischer Verfahren notwendig. Praktisch finden verschiedene Verfahren der Likelihood-Maximierung sowie Methoden aus der bayesschen Statistik Anwendung (Embretson & Reise, 2000; Johnson, 2007). Die Unterschiede zwischen ihnen bestehen vor allem in der Art, wie die Personenparameter geschätzt und welche Annahmen über diese getroffen werden.

3.4.1 Maximum-Likelihood-Verfahren

Die Grundlage aller *Maximum-Likelihood-Schätzverfahren* (Maximum Likelihood Estimation, MLE) ist die *Likelihoodfunktion* L. In ihr wird für ein angenommenes Modell die Gesamtwahrscheinlichkeit des Auftretens der empirischen Daten ausgedrückt. So kann für verschiedene Sätze an Parametern geprüft werden, wie wahrscheinlich diese die beobachteten Antworten erzeugt hätten. Durch systematische Veränderung der Parameterwerte wird mittels numerischer Algorithmen derjenige Satz an Werten bestimmt, für den die Likelihoodfunktion ihr Maximum erreicht – also die Werte, die die empirischen Beobachtungen am wahrscheinlichsten erzeugt haben könnten.

Am einfachsten darzustellen ist die Definition der *Joint Maximum Likelihood* (JML). Da eine der Annahmen für die Verwendung eines IRT-Modells die lokale stochastische Unabhängigkeit der einzelnen Antworten ist, kann die Wahrscheinlichkeit der gesamten Messung als Produkt der Wahrscheinlichkeiten aller einzelnen Antworten ausgedrückt werden:

$$L = \prod_i \prod_j P(x_{ij})$$

Um die Likelihood zu berechnen, wird $P(x_{ij})$ je nach verwendetem IRT-Modell für jede Antwort einzeln berechnet und über alle Beobachtungen multipliziert. Die Verwendung der JML kann genutzt werden, um alle Personen- und Aufgabenparameter des Modells zu schätzen. Hier liegt jedoch auch das Problem des Verfahrens, die JML-Methode führt insbesondere bei kleinen Probandenzahlen oder höherdimensionalen Modellen zu inkonsistenten und verzerrten Parameterschätzungen (Drasgow, 1989).

Eine Alternative ist die Verwendung der *Marginal Maximum Likelihood* (MML). Die Wahrscheinlichkeit des Antwortmusters einer einzelnen Person kann ausgedrückt werden als ein Produkt aus den bedingten Wahrscheinlichkeiten

a) ihres Rohwertes (Anzahl richtiger Antworten) unter Kenntnis von Aufgaben- und Personenparametern und
b) des Antwortmusters unter Kenntnis von Rohwert, Aufgaben- und Personenparametern (Andersen, 1970):

$$\begin{aligned} P(\overrightarrow{x_i}) &= P(\overrightarrow{x_i}|r_i.\,\theta_i, \beta) * P(r_i|\theta_i, \beta) \\ &= \int P(\overrightarrow{x_i}|\theta_i, \beta)\, dG(\theta) \end{aligned}$$

β steht hierbei für die Menge aller Aufgabenparameter und $\overrightarrow{x_i}$ für das Antwortmuster, r_i für den Rohwert und θ_i für die Fähigkeit von Person i. $G(\theta)$ ist die angenommene Verteilungsfunktion für die Personenparameter innerhalb der Stichprobe, üblicherweise wird die Standardnormalverteilung $N(0,1)$ verwendet.

Die Likelihoodfunktion berechnet sich dann als Produkt über die Wahrscheinlichkeiten aller beobachteten Antwortmuster:

$$L = \prod_i \prod P(\vec{x_i})$$

Durch die Integration über eine angenommene Verteilungsfunktion der Personenparameter fallen die einzelnen Fähigkeiten aus der Berechnung der Likelihoodfunktion weg. Somit können durch das MML-Verfahren die Aufgabenparameter unabhängig von den Personenparametern geschätzt werden, was die numerische Suche nach den optimalen Parametern erleichtert. Weiterhin handelt es sich bei der MML um ein Verfahren, das die Schätzung höherparametrischer Modelle wie 2pl und 3pl ermöglicht (Drasgow, 1989). Der potentielle Nachteil des Vorgehens liegt in der Wahl von $G(\theta)$; weichen die theoretisch begründeten Annahmen über die Verteilung der Personenparameter weit von der Realität ab, so sind die anhand der MML geschätzten Parameter stark verzerrt (Hatzinger, 2010). Für das 1pl-Modell besteht mit der C*onditional Maximum Likelihood* ein weiteres Verfahren, mit dem die Aufgabenparameter einzeln geschätzt werden können. Hierbei sind, im Gegensatz zur MML, keine Annahmen über die Verteilung der Personenparameter nötig. Jedoch kann das Verfahren nicht für höherparametrische Modelle verwendet werden (Hatzinger, 2010).

Sofern Aufgabenparameter mit MML- oder CML-Verfahren bestimmt wurden, kann die Schätzung der Personenparameter hinterher auch anhand der JML erfolgen. Die Aufgabenparameter werden dabei als bekannte Größen in die Likelihood-Funktion der JML eingesetzt. Die bereits angesprochenen Inkonsistenzen und unsystematische Verzerrungen treten dann aufgrund der reduzierten Freiheitsgrade in der Schätzung nicht mehr auf. Trotzdem kann, egal welches Verfahren angewandt wurde, für geschätzte Personenparameter aus MLE-Verfahren eine systematische Verzerrung (oder Bias) gezeigt werden (Lord, 2012). Üblicherweise wird daher aus den zunächst gefundenen MLE-Werten ein um dieses Bias

korrigierter Wert berechnet. Dieser wird als *Weighted Likelihood Estimate* (WLE)[9] bezeichnet.

Für die vorliegende Familie an IRT-Modellen wird eine Korrektur nach Warm angewandt (Linacre, 2009; Warm, 1989).

3.4.2 Bayessche Schätzverfahren

Eine weitere Möglichkeit der Bestimmung von Personenparametern sind Verfahren aus der Bayesschen Statistik. Sie setzen voraus, dass die Aufgabenparameter bekannt sind. Dementsprechend ist die Anwendung von MLE-Verfahren in einem vorigen Schritt unabdingbar.

Da die folgenden Ausführungen auf dem Satz von Bayes beruhen, soll dieser kurz erläutert werden. Mathematisch ausgedrückt lautet er wie folgt:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Dabei sind $P(A)$ und $P(B)$ die Anfangswahrscheinlichkeiten (a-priori-Wahrscheinlichkeiten) der Ereignisse A und B; $P(A|B)$ ist die bedingte Wahrscheinlichkeit (A-Posteriori-Wahrscheinlichkeit) von A für den Fall, dass B bereits eingetreten ist und umgekehrt. Der Satz ermöglicht also Aussagen über die Wahrscheinlichkeit eines Ereignisses, falls ein anderes damit zusammenhängendes Ereignis bereits eingetreten ist. Er besagt: Die Wahrscheinlichkeit von A ist, sofern B bereits eingetreten ist, gleich der Wahrscheinlichkeit eines gemeinsamen Auftretens von A und B geteilt durch die Wahrscheinlichkeit von B (Behnke & Behnke, 2006):

In der IRT kann der Satz von Bayes genutzt werden, um die A-Posteriori-Wahrscheinlichkeit von Personenparametern beziehungsweise eine Verteilungsfunktion für diese zu bestimmen.

Hierfür ist es notwendig, Annahmen über die Verteilung der Personenfähigkeiten θ innerhalb der Stichprobe aufzustellen. $G(\theta)$ soll diese Verteilung der Fähigkeiten bezeichnen. Dann ist die Wahrscheinlichkeit für die Fähigkeit θ_n innerhalb der Stichprobe gleich $P(\theta_n) = G(\theta_n)$. Mittels des Satzes von Bayes kann nun unter Kenntnis des beobachteten

[9] Teils auch WMLE für W*eighted Maximum Likelihood Estimate*

Antwortverhaltens $\vec{x}$ einer Person die bedingte Wahrscheinlichkeit dafür bestimmt werden, dass diese Person die Fähigkeitsausprägung θ_n besitzt:

$$P(\theta_n|\vec{x}) = \frac{P(\vec{x}|\theta_n)P(\theta_n)}{P(\vec{x})} = \frac{P(\vec{x}|\theta_n)G(\theta_n)}{P(\vec{x})} = \frac{P(\vec{x}|\theta_n)G(\theta_n)}{\sum_m P(\vec{x}|\theta_m)G(\theta_m)}$$

Wobei θ_m für alle einzelnen möglichen Ausprägungen von θ steht. Die Gleichung gilt daher leider nur für diskrete Fähigkeitsverteilungen, dazu im nächsten Absatz mehr. Zunächst ist festzuhalten: Sofern die Aufgabenparameter aus den vorigen MLE-Verfahren bekannt sind, können alle Wahrscheinlichkeiten auf der rechten Seite der Gleichung analytisch berechnet werden. Setzt man eine vermutete Fähigkeitsverteilung als gegeben voraus, kann somit die Gleichung gelöst und für jedes beobachtete Antwortmuster die Wahrscheinlichkeit unterschiedlicher Fähigkeitsausprägungen bestimmt werden.

Meist wird aber davon ausgegangen, dass die zu messenden Fähigkeiten sich nicht nur diskret in verschiedenen Niveaus bewegen, sondern eine stetige Variable sind. In diesem Fall kann, analog zum vorigen Absatz, die Wahrscheinlichkeitsdichte von θ_n bestimmt werden:

$$h(\theta_n|\vec{x}) = \frac{P(\vec{x}|\theta_n)G(\theta_n)}{\int P(\vec{x}|\theta_m)G(\theta_m)d\theta}$$

Die A-Posteriori-Verteilung von θ_n wird durch Integration über die Fähigkeit erlangt:

$$f(\theta_n|\vec{x}) = \int h(\theta_n|\vec{x})\, d\theta$$

Ihr Erwartungswert berechnet sich wie folgt:

$$E\big(f(\theta_n|\vec{x})\big) = \int \theta * h(\theta_n|\vec{x})\, d\theta$$

Dieser Erwartungswert wird als *expected a posteriori* (EAP) (Linacre, 2002) bezeichnet. Er stellt den Fähigkeitswert dar, der (bei einer gegebenen und bekannten Fähigkeitsverteilung in der Stichprobe) für ein gegebenes

Antwortmuster am wahrscheinlichsten auftritt. Aus diesem Grund wird er als Alternative zu MLE-Werten als Schätzung für Personenfähigkeiten genutzt.

Schließlich gibt es die Möglichkeit, sogenannte *plausible values* (PV) (Wu, 2004) aus der A-Posteriori-Verteilung zu ziehen. Hierbei werden für jede Person aus der individuellen Verteilung zufällige Werte gezogen. Die individuelle Diagnose einzelner Personen ist damit nicht genau möglich. Zwar sind die Werte, wie der Name verrät, plausibel. Schließlich werden sie aus der Wahrscheinlichkeitsverteilung gezogen. Jedoch handelt es sich aufgrund der zufälligen Ziehung selten um die Werte, die als die wahrscheinlichsten anzusehen sind (also die EAPs). Wieso PVs trotzdem sinnvolle Schätzmaße sind, wird im nächsten Abschnitt erläutert.

3.4.3 Schätzverfahren im Vergleich

In den vorigen Abschnitten wurden drei verschiedene Schätzwerte für Personenfähigkeiten vorgestellt: WLE, EAP und PV. Alle drei setzen voraus, dass die Aufgabenparameter schon vor der Messung bekannt sind oder durch Anwendung eines MLE-Verfahrens bestimmt wurden. Ebenfalls ist es für alle drei Schätzer notwendig, Annahmen über die Verteilung der Fähigkeiten innerhalb der Stichprobe anzustellen und diese als gegeben in die Berechnung einzufügen. Es scheint also sinnvoll, sich damit zu befassen, worin der Unterschied zwischen den drei Schätzwerten besteht.

a) Wie bereits unter Abschnitt 3.4.2 erwähnt, handelt es sich bei WLEs und EAPs um Erwartungswerte für individuelle Personenparameter, bei PVs um zufällig gezogene Parameter aus einer individuellen Verteilung wahrscheinlicher Fähigkeiten. Für die Diagnose einzelner Probanden ist daher die Genauigkeit von PVs geringer als von den anderen beiden Werten (Lüdtke & Robitzsch, 2017; Monseur & Adams, 2009).
b) In Abschnitt 3.4.1 wurde der Bias von MLEs angesprochen. Hierfür gibt es Korrekturen, anhand derer WLEs berechnet werden. Sie sind also weniger verzerrt. Dennoch kann beobachtet werden, dass je nach Länge des angewandten Testinstruments die Varianz der

Fähigkeitsverteilung verzerrt ist, also ein gewisser Bias verbleibt. EAPs zeigen ein ähnliches Verhalten. Einzig PVs stellen die Personenverteilung unverzerrt dar.

Durch die simulierte Befragung einer Proband*innengruppe kann der zweite Aspekt anschaulich gemacht werden: Die folgenden Daten (siehe Abbildung 13) basieren auf Testdurchläufen von 1000 virtuell generierten Proband*innen mit standardnormalverteilten Fähigkeiten. Die Testaufgaben sind mit einer Varianz von $\sigma^2 = 2$ normalverteilt, zur Antwortgenerierung und Auswertung wurde jeweils ein 1pl-Modell verwendet.

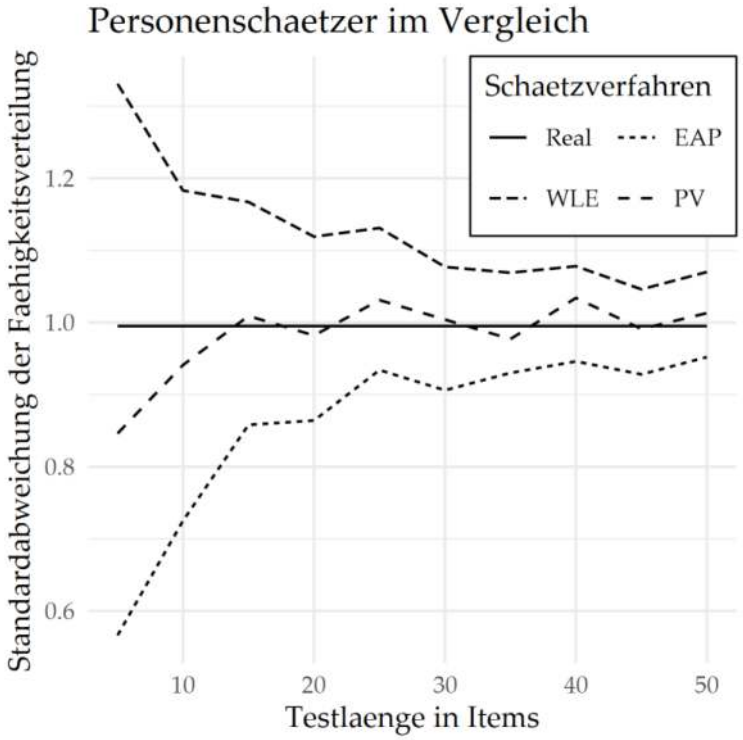

*Abbildung 13: Vergleich unterschiedlicher Schätzmaße für Personenfähigkeiten durch simulierte Befragungen. In der Abbildung dargestellt ist die Standardabweichung der Fähigkeiten einer hypothetischen Proband*innengruppe. Die durchgezogene Linie stellt den realen Wert der Standardabweichung dar. Die drei unterbrochenen Linien zeigen die Standardabweichungen, die sich aus der Auswertung und Fähigkeitsschätzung durch verschiedene Verfahren ergeben. Sie sind abhängig von der Länge der simulierten Tests.*

In Abbildung 13 ist der Unterschied zwischen den drei vorgestellten Personenschätzern klar zu erkennen. Werden bei der Auswertung der Testdaten für jede Person PVs gezogen, kann die Fähigkeitsverteilung in der gesamten Stichprobe recht genau reproduziert werden. Die anhand der

Schätzwerte berechnete Standardabweichung schwankt nur in einem kleinen Bereich um den wahren Wert.

Für die beiden anderen Maße, WLE und EAP, können systematische Fehler beobachtet werden. Beide Verfahren führen bei geringer Testlänge zu Abweichungen zwischen der beobachteten und realen Verteilung der Fähigkeiten in der Stichprobe. Dabei wird die Streuung der Fähigkeiten durch EAPs unterschätzt und durch WLEs überschätzt. Die Abweichung ist abhängig von der Menge der bearbeiteten Aufgaben.

Während WLE und EAP also genauere Verfahren für die Feststellung individueller Fähigkeiten sind, eignen sich PVs besser, um Merkmalsverteilungen in ganzen Stichproben wiederzugeben. Wu (2005) zeigte diese Unterschiede ebenfalls in einer von ihr durchgeführten Simulationsstudie.

4 Adaptive Testverfahren

Die Kernidee adaptiver Verfahren ist die individuelle Anpassung von Tests an das Antwortverhalten verschiedener Personen (Frey, 2012). Es sollen dabei Aufgaben und Fragen so ausgewählt werden, dass sie möglichst gut für die jeweiligen Proband*innen geeignet sind. Das kann zum Beispiel bedeuten, keine schweren Aufgaben an leistungsschwache Proband*innen zu vergeben – das Ergebnis einer falschen Lösung ist dabei zu vorhersehbar und liefert keine sinnvollen, neuen Erkenntnisse.

Mit der Idee eines optimalen Tests für jede Person müssen solche Situationen also ausgeschlossen werden. Eine direkte Konsequenz davon ist die Notwendigkeit, allen Proband*innen unterschiedliche Items vorlegen zu können. Die Auswahl der geeigneten (beziehungsweise der Ausschluss der ungeeigneten) Items erfolgt als direkte Reaktion auf das Verhalten der Person im bisherigen Verlauf des Tests.

Traditionell werden persönliche Merkmale und Fähigkeiten mit Tests in Papierform erfasst (Magis, Yan & Davier, 2017; Moosbrugger & Kelava, 2012). Diese werden zuvor mit einer festen Zusammenstellung von Items konstruiert und auch in genau dieser Zusammenstellung an alle befragten Personen verteilt. Dementsprechend sind sie starr und nicht adaptiv. Im englischsprachigen Raum finden sich hierfür unter anderem die Bezeichnungen als *Fixed-Item-Tests* (FIT) (Ling, Attali, Finn & Stone, 2017) oder *Linear Tests* (Magis et al., 2017). Um Verwechslungen mit der ebenfalls üblichen Bezeichnung der mathematischen Modellpassung als Fit zu vermeiden, wird im weiteren Verlauf der Arbeit die zweite Option, *lineare Tests*, verwendet.

Dass lineare Tests in ihrer Form starr sind, hat direkte Folgen für die Menge an benötigten Items. Für jeden zu messenden Merkmals- oder Fähigkeitsbereich müssen schließlich ausreichend Items im Test enthalten sein, um diesen ausreichend genau zu erfassen. Gerade bei der Erfassung von Kompetenzen, für die oft schon theoretisch mehrere Niveaustufen erwartet werden, sind die nötigen Messinstrumente entsprechend umfangreich.

Diese Menge an vorgelegten Fragen kann negative Auswirkungen auf die Konzentrationsfähigkeit und Motivation der Befragten haben. Mögliche Folgen sind unter anderem eine geringe Teilnahmebereitschaft und damit Populationsverluste sowie auffälliges Rateverhalten und allgemein verminderte Leistungen während der Befragung (Bühner, 2011; Moosbrugger & Kelava, 2012; Rost, 2004).

Diese Problematiken sind, wenn auch teils nur anekdotisch belegt, schon seit langer Zeit bekannt. Es gab deswegen schon sehr früh Bemühungen, bei psychometrischen Tests Anpassungen an die einzelnen Testteilnehmer*innen vorzunehmen. Bereits seit 1905 ist im Binet-Simon-Intelligenztests (Boake, 2002) ein Verfahren dokumentiert, bei dem die Items während der Messung durch geschultes Personal ausgewählt und angepasst wurden. Die Aufgaben des Tests waren nach dem mentalen Alter eingeteilt, das sie repräsentieren sollten. Die Testleitung sollte anhand der bearbeiteten Items das mentale Alter der Versuchspersonen abschätzen und das nächste Item danach auswählen, bis ein finaler Messwert festgestellt war. Solche Testformate sind die Vorläufer dessen, was heutzutage als adaptiver Test verstanden wird.

An der Beschreibung wird leicht ersichtlich, wie aufwändig frühe adaptive Messungen waren. Für regelmäßige oder großflächige Messungen wäre ein solches Vorgehen kaum praktikabel. In den letzten 50 Jahren konnten jedoch enorme Fortschritte in mathematischen, messtheoretischen und technischen Bereichen gemacht werden. Hierzu gehören die Entwicklung der Item-Response Theory (Kapitel 3), der breitflächige Einsatz von Computern sowie die Weiterentwicklung von Ideen und Verfahren zur Itemselektion. Einen weiteren Einblick in diesen Prozess geben van der Linden und Glas (2010).

Das erste Instrument, das nach heutigem Verständnis erfolgreich als adaptiver Test umgesetzt wurde, ist die „Armed Services Vocational Aptitude Battery“ (ASVAB, in adaptiver Form CAT-ASVAB) (van der Linden & Glas, 2010). Dabei handelt es sich einen Multiple-Choice-Leistungstest aus mehreren Bestandteilen. Er umfasst unter anderem mathematische, technische und sprachliche Grundfähigkeiten und wird vom US-Militär

verwendet, um die Eignung zum Militärdienst zu beurteilen (Army Marketing and Research Group & The United States Army, 2020). Die Testergebnisse korrelieren stark mit kristalliner Intelligenz (Roberts et al., 2000; Schmidt-Atzert, Deter & Jaeckel, 2004).

Moderne adaptive Tests zeichnen sich im Vergleich zu ihren Vorgängern durch zwei grundlegende Merkmale aus: Erstens werden sie immer an Computern durchgeführt (Frey, 2012), weshalb sich der Begriff *computeradaptiver Tests* (CAT) etabliert hat und im Folgenden verwendet wird. Zweitens findet die Itemauswahl inzwischen fast ausschließlich auf Basis der IRT statt (H.-H. Chang, 2015).

In diesem Kapitel werden grundlegende Merkmale und Bestandteile adaptiver Tests vorgestellt. Zudem werden die verschiedenen Testformate auf der Basis vorliegender Studien untereinander verglichen. Die Auswahl beschränkt sich dabei auf elementare beziehungsweise im Kontext dieser Arbeit relevante Inhalte: Das breite Feld der Testkonstruktion im Bereich adaptiver Tests voll abzudecken, wäre im Rahmen dieser Arbeit auch nicht zielführend.

4.1 Vergleich papier- und computerbasierter Tests

Für den späteren Vergleich von linearen Tests und CATs (theoretischer Vergleich in Abschnitt 4.4 und praktischer Vergleich in Bezug auf das KoWADiS-Instrument in Abschnitt 8.3) lohnt es sich, an dieser Stelle kurz die Unterschiede zwischen papierbasierten und am Computer administrierten Tests aufzuzählen. Traditionell werden lineare Tests noch häufig in Papierform verwendet. Zu den Vor- und Nachteilen gegenüber CATs könnten daher schnell auch solche gezählt werden, die allein durch das Medium verursacht sind. Da sich aber selbstverständlich auch lineare Tests in digitaler Form durchführen lassen, sollte das vermieden werden.

Zu möglichen Vorteilen einer computerbasierten Messung zählen:

1. Höhere Flexibilität bei der Vereinbarung und Organisation von Erhebungszeitpunkten. Testinstrumente, die nicht zwingend unter Aufsicht durchgeführt werden müssen, können von Probanden zeitlich und räumlich ungebunden bearbeitet werden. Reisezeiten und -kosten der Testleitung können ebenfalls minimiert werden.
2. Erleichterte und schnellere Dateneingabe und Auswertung, da keine Übertragung der Daten zwischen mehreren Medien nötig ist.
3. Damit im Zusammenhang steht eine prinzipiell höhere Objektivität der Durchführung, da weniger Interpretationsfehler durch manuelles Auslesen der Antworten entstehen.
4. Einfache Randomisierung und Testvergabe bei mehreren Testgruppen.
5. Größere Gestaltungsfreiheit bei der Erstellung neuer Items.

Als Nachteile gelten dagegen:

1. Unsicherheit bezüglich der Echtheit/Unverfälschtheit der Daten, falls Tests unbeaufsichtigt durchgeführt werden.
2. Verringerung menschlicher Kommunikation, wodurch zum Beispiel weniger Möglichkeit zum Nachfragen besteht.
3. Durch den größeren Einfluss automatisierter Prozesse kann es dazu kommen, dass Ergebnisse unhinterfragt übernommen werden. Das gilt besonders, wenn ein Test von Dritten ausgeführt wird.

Daneben gibt es anhaltende Diskussionen über den Einfluss verschiedener Medien auf die Motivation, Teilnahmebereitschaft sowie die tatsächlich erbrachte Leistung von Proband*innen – sowohl im Allgemeinen als auch in Abhängigkeit von vorherigen Erfahrungen mit den jeweiligen Medien. Studien, die beide Testmodi in Bezug auf Leistungstests und nicht etwa Selbsteinschätzungsskalen oder Ähnliches vergleichen, sind bis heute selten. In den letzten Jahren sind sie vor allem im Bereich der Sprachwissenschaften, speziell in Bezug auf Englisch als Fremdsprache zu finden. Dort kommen die bestehenden Veröffentlichungen allerdings zu widersprüchlichen Ergebnissen.

Al-Amri (2007) konnte in einer Studie zur Lesefähigkeit mit 167 Medizinstudent*innen keinen Einfluss des Testmodus feststellen. Ebenso konnte kein signifikanter Einfluss durch frühere Erfahrung und Geübtheit im Umgang mit Computern sowie Vorlieben bezüglich des Testmodus auf die Performanz im computerbasierten Test gefunden werden.

In einer weiteren Studie zu Leseverständnis, diesmal in der Primarstufe (N = 216), fanden Higgins, Russell und Hoffmann (2005) ebenfalls keinen Unterschied zwischen den Testmodi und auch keinen signifikanten Einfluss von Vorerfahrungen auf Performanz am Computer. Bei der Befragung bevorzugte ein Großteil der Proband*innen die Computerversion. Identische Resultate fanden sich bei Khoshsima, Hashemi Toroujeni, Thompson und Reza Ebrahimi (2019) in Vergleichen mit 58 Englischstudent*innen und bei Öz und Özturan (2018) mit 97 angehenden Englischlehrkräften bezüglich ihrer allgemeinen Sprachkenntnisse.

Dagegen berichtet Paek (2005) in einer Metaanalyse insbesondere in Bezug auf Leistungen im Leseverständnis eine Benachteiligung von Proband*innen durch scheinbar schwerere Tests am Computer, während Laborda, Santiago, Juan und Álvarez (2014) in ihrer Studie eine höhere Performanz bei computerbasierten Tests vorfanden.

Insgesamt kann anhand der veröffentlichten Studien keine eindeutige Schlussfolgerung gezogen werden. Es ist aber festzustellen, dass die Mehrzahl der Studien keinen Unterschied in Bezug auf Performanz nachweist. In den Fällen, in denen signifikante Unterschiede aufgezeigt werden, sind die Effekte schwach. Zudem widersprechen sich die angeführten Quellen. Die Testinstrumente aus den zitierten Studien bewegen sich hierbei alle in einem anderen Inhaltsgebiet als das Instrument, das Fokus dieser Arbeit ist. Konkrete Vergleiche des Testmodus in Bezug auf die Erfassung von naturwissenschaftlichen Denk- und Arbeitsweisen konnten nicht gefunden werden.

4.2 Computeradaptive Tests

Ein grundlegendes Ablaufschema computeradaptiver Tests zeigt Abbildung 14. Nach der Bearbeitung des ersten Items wird durch den vorgegebenen Algorithmus dasjenige Item aus dem Pool gesucht, das am besten für die weitere Messung der Testperson geeignet ist. Die Eignung wird vor allem anhand der Testinformation (Abschnitt 3.3.1) beurteilt. Es wird aus dem Pool der übrigen Items entfernt und der Testperson vorgelegt. Dieser Prozess wird so häufig wiederholt, bis ein vorgegebenes Abschlusskriterium erfüllt wird und die Messung beendet wird.

Daran können verschiedene Elemente ausgemacht werden, die in gewisser Weise als Bausteine eines CATs bezeichnet werden könnten:

a) der Itempool
b) der Startpunkt beziehungsweise die Auswahlregeln, nach denen das erste Item des Tests ausgesucht wird
c) das Abschlusskriterium, bei dessen Erfüllung die Messung beendet wird
d) die Auswahlregeln für Aufgaben innerhalb des laufenden Tests

Die letzten drei Punkte könnten zusammengefasst auch als der Testalgorithmus bezeichnet werden, da sie die logische Struktur und den Verlauf der Messung bestimmen. Da sie bei der Konstruktion eines CATs häufig einzeln bestimmt werden, werden sie auch im Folgenden getrennt betrachtet.

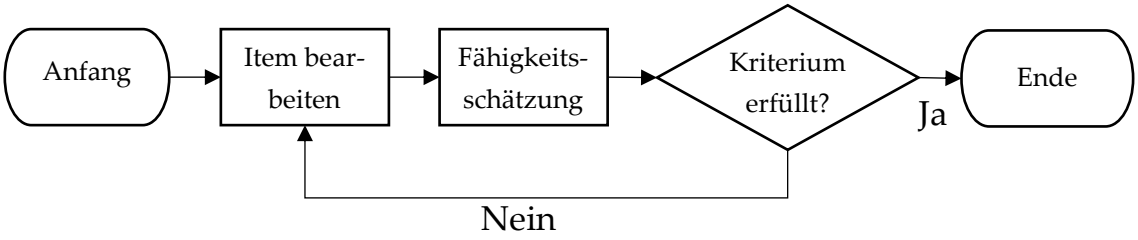

Abbildung 14: Ablaufschema eines adaptiven Tests.

4.2.1 Testalgorithmus: Itemauswahl während des Tests

Innerhalb der Messung erfolgt die Auswahl geeigneter Items anhand von psychometrischen und inhaltlichen Kriterien.

Das übliche psychometrische Kriterium ist die Maximierung des vorausgesagten Informationsgewinns durch das nächste Item (H.-H. Chang & Ying, 1999). Hierzu wird als erstes eine Schätzung der Personenfähigkeit vorgenommen. Das bisher beobachtete Antwortmuster wird zusammen mit den Itemparametern verwendet, um anhand des ausgesuchten IRT-Modells einen Schätzwert für den/die Proband*in zu bestimmen.[10] Danach wird für alle noch im Itempool verfügbaren Items der Informationsgewinn für genau diese Personenfähigkeit bestimmt. Der maximale Wert wird gesucht und das entsprechende Item kann als nächstes vorgelegt werden.

Für die Testinformation liegen mehrere Definitionen vor. Magis et al. (2017) geben einen Überblick über acht verschiedene Möglichkeiten der mathematischen Definition, die alle bereits im Kontext des adaptiven Testens angewendet wurden. Am häufigsten verwendet wird von diesen die Definition nach Fischer, die bereits in Abschnitt 3.3.1 besprochen wurde (Baker & Kim, 2017).

Inhaltliche Kriterien variieren dagegen je nach Test und hängen vor allem von der angenommenen Struktur und Breite des gemessenen Merkmals ab. Als Beispiel kann die Messung von physikalischem Fachwissen dienen. Auch wenn angenommen wird, dass es sich bei Grundlagenwissen zur Mechanik um ein in sich geschlossenes Konstrukt handelt, sollte es nicht nur anhand eines einzelnen Kontextes erfasst werden. Statt Proband*innen eine Reihe von Items zum Kraftmesser zu präsentieren, sollten verschiedene Kontexte behandelt und die gesamte Breite der Thematik abgedeckt werden. Sind im Itempool aber genug Items, um einen Test allein anhand eines einzelnen Kontextes durchzuführen, kann genauso ein Fall auftreten.

[10] Von den in Abschnitt 3.4 genannten Schätzwerten sind dazu sowohl WLE als auch EAP geeignet. PVs kommen aus den im genannten Abschnitt besprochenen Gründen nicht in Frage.

Es muss dann manuell anhand von Auswahlregeln gegengesteuert werden.

Noch deutlicher wird das Beispiel, wenn man mehrere übergeordnete Themenbereiche zusammen messen will, in etwa Wissen im Bereich der Mechanik und Optik. Hier ist es sinnvoll, neben der Variation der einzelnen Kontexte auch für eine ausgewogene Balance zwischen den verschiedenen Themenfeldern zu sorgen. Es wird dann vom Problem der *content balance* gesprochen (Kingsbury & Zara, 1989). Content balance bezieht sich nur auf die Messung von Konstrukten, die trotz ihrer verschiedenen inhaltlichen Aspekte als eine einzelne mathematische Dimension modelliert werden. Andernfalls liegt ein multidimensionales IRT-Modell vor, was hier nicht thematisiert wird.

Ebenfalls getrennt zu betrachten ist die *Itemexposition* (engl. *Item exposure*) (vgl. Stocking & Lewis, 2002). Hiermit wird hauptsächlich eine Gefahr für die Sicherheit des Testinstruments gegenüber absichtlichen Verfälschungen bezeichnet: Es kann Items geben, die besonders häufig durch den Testalgorithmus ausgewählt werden. Ein Grund kann der besonders hohe Informationsgewinn durch dieses Item im Mittel der Stichprobenverteilung sein. Wird der Test regelmäßig eingesetzt, steigt somit insbesondere für diese Items die Chance, bereits im Vorfeld bekannt zu sein. Im Fall von längsschnittlichen Beobachtungen könnten Erinnerungseffekte auftreten. Wird ein Instrument als Eingangstest oder zur Beurteilung in Prüfungen verwendet, bietet sich bei einer hohen Itemexposition die Möglichkeit des gezielten Betrügens.

Für die Definition geeigneter Auswahlregeln gibt es verschiedene Ansätze. Kingsbury und Zara (1989) schlagen im Sinne der content balance vor, die Items den verschiedenen Inhaltsbereichen nach in Subgruppen zu trennen und den Testalgorithmus für alle Gruppen parallel zu durchlaufen. Es gibt dann für jeden Inhaltsbereich jeweils einen nach Informationskriterien ausgesuchten optimalen Kandidaten – davon wird immer das Item ausgewählt, dessen Inhaltsbereich noch am stärksten unterrepräsentiert ist (Leung, Chang & Hau, 2003). Einer Gefährdung durch Itemexposition hingegen kann beispielsweise durch zufällige (Kingsbury & Zara, 1991;

McBride & Martin, 1983) oder systematische (Hetter & Sympson, 1997) Variation innerhalb der informativsten Items entgegengewirkt werden.

S.-W. Chang und Ansley (2003) verglichen anhand von Simulationsstudien verschiedene Alternativen zur Vermeidung von hoher Itemexposition, wobei keine eindeutig beste Methode gefunden werden konnte. Über alle verglichenen Methoden hinweg war festzustellen, dass eine starke Fokussierung auf die inhaltlichen Aspekte zu Einbußen in der Messgenauigkeit führt.

Übergeordnete Vorgehensweisen können beide inhaltlichen Aspekte zusammen mit den Informationskriterien vereinen, benötigen jedoch aufwändigere Testalgorithmen und stellen damit höhere Anforderungen an die Testentwicklung sowie die Infrastruktur zur Testdurchführung. Dazu gehören beispielsweise *shadow tests* oder *Testlet-basierte Tests* (Linden & Glas, 2002; Magis et al., 2017).

4.2.2 Testalgorithmus: Itemauswahl vor dem Test

Das Ziel bei der Auswahl des erstens Items ist prinzipiell das gleiche wie auch bei allen nachfolgenden Items: Die durch Itembearbeitung gewonnene Information soll optimiert werden. Es besteht aber hier der besondere Fall, dass noch kein Antwortmuster von den Proband*innen vorliegt. Da damit aber üblicherweise der mögliche Informationsgewinn geschätzt wird, müssen andere Kriterien verwendet werden.

Üblich ist die Wahl einer Startaufgabe, die eine mittlere Schwierigkeit besitzt (Frey, 2012). Der Grund dafür ist, dass bei noch komplett unbekannten Proband*innen am ehesten von einer mittleren Personenfähigkeit auszugehen ist. Für diesen Fall stellt ein mittelschweres Item den höchsten Informationsgewinn dar. Bei dichotomen Antwortformaten kann außerdem als Alternative eine Startaufgabe ausgesucht werden, deren Schwierigkeit möglichst nah am angenommenen Mittelpunkt der Fähigkeitsverteilung liegt, wie von Urry (1970) vorgeschlagen.

Daneben sind im Vorfeld bekannte Personenmerkmale oder erhobene Kovariaten zu berücksichtigen, die eine Einschätzung der Fähigkeit ermöglichen. Ergebnisse aus ähnlichen Tests oder bekanntermaßen mit der

Fähigkeit korrelierende Größen wie beispielsweise Semesterzahl können genutzt werden, um im Voraus eine Merkmalsschätzung durchzuführen. Damit kann ein individuell passendes Startitem ausgesucht werden (van der Linden, 1999).

Ein Problem bei diesen Auswahlregeln kann auftreten, wenn Personen in den Extrembereichen der Merkmalsverteilung den Test bearbeiten. In diesem Fall ist die Differenz zwischen Itemschwierigkeit und Merkmal sehr groß, was einen minimalen Informationsgewinn zur Folge hätte. Dieses Szenario wird noch häufiger, je flacher die Verteilung des Merkmals ist, da mehr Personen in den Randbereichen liegen. Außerdem könnten (bei der Verwendung von zweiparametrischen oder komplexeren IRT-Modellen) durch die Maximierung der Testinformation Items ausgesucht werden, die eine hohe Trennschärfe aufweisen. Damit ist zwar am angepeilten Wert ein höherer Informationsgewinn möglich, dieser fällt dafür aber schon bei kleinen Abweichungen schnell ab.

Daraus ergeben sich zwei Denkansätze. Zum einen sollte, falls im Itempool viele Items mit hohen Trennschärfen zur Verfügung stehen, eine Begrenzungsregel für die Trennschärfe des ersten Items definiert werden. Zum anderen kann anstelle eines einzelnen Items ein sogenannter *routing test* an den Start des Tests gestellt werden. Es handelt sich dabei um einen vorkonstruierten Block aus mehreren Items. Die Items im routing test werden so gewählt, dass sie eine breitere Fähigkeitsverteilung um den anfangs angenommenen Wert abdecken. Auf diese Weise kann für einen größeren Teil des Fähigkeitsspektrums ein ausreichender Informationsgewinn gewährleistet werden, der für die danach anstehenden Entscheidungen ausreichend ist. Frey (2012) weist allerdings auch darauf hin, dass weder in empirischen noch simulierten Studien ein großer Einfluss des Startitems auf die Performanz festgestellt werden kann.

Unabhängig davon, welche Form die Startauswahl im konkreten Fall annimmt – es gelten dieselben Bedenken bezüglich der Itemexposition wie im Rest des Tests. Dagegen können die gleichen Maßnahmen ergriffen werden wie innerhalb der laufenden Messung, wie zum Beispiel eine Zufallsauswahl aus mehreren geeigneten Items.

4.2.3 Testalgorithmus: Abschlusskriterien

Bei Sichtung der Literatur zur Konstruktion von CATs fällt schnell auf, dass ebenso wie für die Itemauswahl auch eine Vielzahl an Kriterien zum Abbruch einer Messung existieren. Welche dieser Regeln berücksichtigt werden und wie diese im Detail bezeichnet werden, unterscheidet sich je nach Autor*in. Die folgende Auswahl und Bezeichnung der Regeln basiert auf dem Überblick, den Magis et al. (2017) geben. Ihre Übersicht wurde als geeignete Referenz ausgewählt, da sie mehrere nicht triviale Kriterien gegenüberstellen und für diese eine einheitliche Art der Benennung festlegen.

Beim Längenkriterium wird die absolute Länge des gesamten Tests, also die Anzahl der von allen Proband*innen bearbeiteten Items, festgesetzt. Sobald eine Person diese Zahl von Items vorgelegt bekommen hat, wird der Test beendet. Der Vorteil besteht in der präzisen Kontrolle der individuellen Bearbeitungszeit. Steht nur ein begrenztes Zeitfenster zur Verfügung, bietet sich das Längenkriterium an. Als Nachteil steht dem gegenüber, dass CATs bei Proband*innen mit Merkmalsausprägungen im Randbereich der Verteilung mehr Aufgaben benötigen, um sich auf einen genauen Messwert einzupendeln. Dementsprechend schwankt die Messgenauigkeit bei konstanter Testlänge über die Populationsverteilung hinweg und kann bei Ausreißern ungenau werden. Durch die Simulation von Worst-Case-Szenarien könnte die Testlänge bestimmt werden, die auch in Randbereichen noch eine minimal erwünschte Messgenauigkeit ermöglicht. Das hätte aber einen Test zur Folge, der in der Mitte der Verteilung unnötig lang ist.

Das Präzisionskriterium (im Englischen oft auch *SEM stopping rule*, vgl. Lunz, Bergstrom & Gershon, 1994) stellt als Alternative das genaue Gegenteil dar. Hierbei wird der erwünschte Grad der Messgenauigkeit festgelegt, meist über einen Grenzwert des Standardmessfehlers. Alle Proband*innen bekommen in der Messung Items vorgelegt, bis dieser Wert erreicht wird. Unabhängig von der Testlänge wird zu diesem Zeitpunkt die Messung abgebrochen. Dementsprechend kann unabhängig von der Fähigkeitsverteilung und der individuellen Ausprägung eine einheitliche Messgenauigkeit

erreicht werden, was die Auswertung einer Messung vereinfacht. Sollten die Fähigkeiten weit streuen, können dafür große Schwankungen in der Testlänge zustande kommen.

Das Klassifizierungs- oder Konfidenzkriterium wird genutzt, um die Proband*innen in Bezug auf eine festgelegte Fähigkeitsstufe zu prüfen. Das kann zum Beispiel ein Wert sein, der inhaltlich als Übergang in ein höheres Kompetenzniveau interpretiert wird. Ebenso kann es eine Leistungsgrenze sein, die die Grenze zwischen Noten oder Bestehen und Nicht-Bestehen in einer Prüfungssituation darstellt.

Der Abstand zwischen Grenzwert und aktueller Fähigkeitsschätzung wird dazu nach jedem Item durch den aktuellen Standardfehler der Schätzung dividiert. Das resultierende Maß kann mit der Standardnormalverteilung abgeglichen werden, um die Sicherheit der Einordnung über/unter der Grenze zu bestimmen (Lunz et al., 1994).[11]

Lunz et al. (1994) schlagen vor, das Konfidenzkriterium mit einem Längenkriterium zu kombinieren. Liegt die wahre Fähigkeit einer Testperson nah an dem Grenzwert, kann die Zahl der notwendigen Aufgaben für eine Beurteilung (Beispielsweise im 90%-Konfidenzintervall, was für benotete Tests oder Eignungsprüfungen nicht unrealistisch wäre) sonst schnell ansteigen – sie führen Beispiele an, in denen mehrere Hundert Aufgaben bearbeitet werden müssten.

Als letztes sei das Informationskriterium genannt. Hierbei wird der Test beendet, sobald keines der übrigen Items im Pool einen bestimmten Minimalwert an Information liefern kann. So wird vermieden, die Messung für einen marginalen Gewinn an Messgenauigkeit in die Länge zu ziehen. Es erscheint sinnvoll, dieses Kriterium nur in Kombination mit einem der anderen zu verwenden, da es im Gegensatz kein praktisches oder methodisches Ziel der Messung verfolgt.

[11] Dieses Vorgehen ist mathematisch äquivalent zu einem einseitigen t-Test.

4.3 Multistage-Tests

Eine alternative zur adaptiven Messung anhand von diskreten Items stellen *Multistage-Tests* (MST) dar. Sie bestehen aus mehreren *Stufen* (Stages), die von den Proband*innen in einer festen Reihenfolge durchlaufen werden. Die Stufen wiederum enthalten die *Module* des Tests (siehe Abbildung 15).

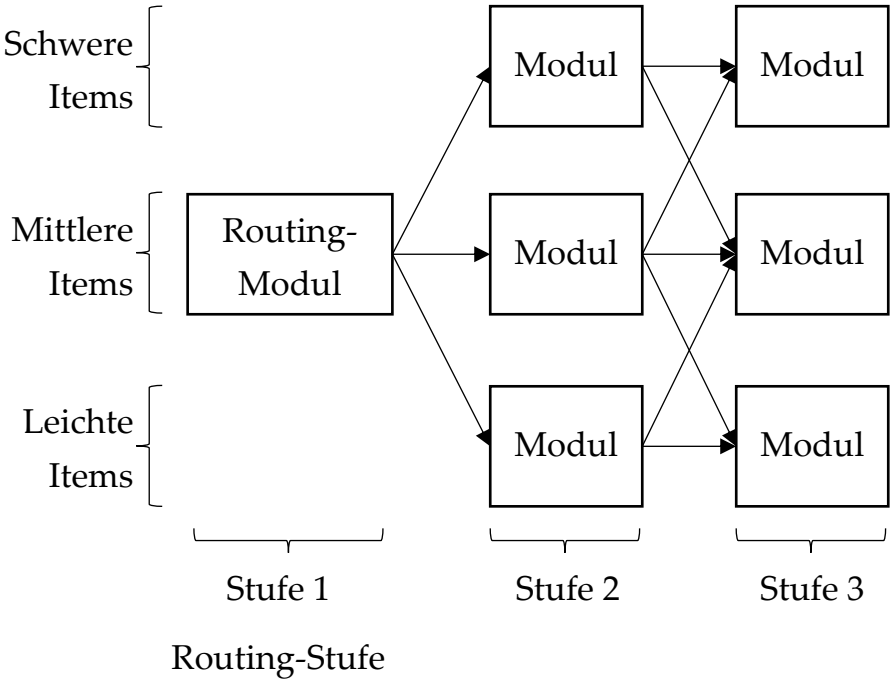

Abbildung 15: Schematische Darstellung eines Multistage-Tests (MST). Das konkrete Beispiel stellt einen MST in der Struktur 1-3-3 dar. Bei dieser Benennung nach Luecht, Brumfield und Breithaupt (2006) stellt jede Zahl die Menge der verschiedenen Schwierigkeitsbereiche und Module in einer der Stufen dar. Die Anzahl an Stufen ergibt sich aus der Länge der Zahlenreihe. Die Darstellung umfasst nur ein Panel.

Module sind in sich abgeschlossene Blöcke aus mehreren Items. Sie sind linear und beinhalten eine festgelegte Menge an Items aus einem begrenzten Schwierigkeitsbereich (Yan, Davier & Lewis, 2014). Daneben werden sie so konstruiert, dass alle Anforderungen bezüglich der content balance erfüllt werden (Hendrickson, 2007). Auch können durch die Module mehrere Dimensionen eines Konstrukts parallel gemessen werden. Damit können sie als eigenständige, abgeschlossene Kurztests betrachtet werden.

Jede Stufe enthält mehrere dieser Module, mit denen insgesamt das gesamte Fähigkeitsspektrum der erwarteten Stichprobe abgedeckt werden kann. Bei der konkreten Messung wird aber nur eines dieser Module von jeder/m Proband*in bearbeitet: Zwischen den einzelnen Stufen wird durch den Testalgorithmus genau wie bei diskreten Tests eine Fähigkeitsschätzung durchgeführt. Aus der folgenden Stufe wird dann das Modul ausgewählt, welches den höchsten Informationsgewinn verspricht. Einen Sonderfall stellt die erste Stufe dar. Sie enthält lediglich ein Modul, meist mit einem mittleren Schwierigkeitsniveau (Hendrickson, 2007; Lord, 2012), das *Routing-Modul* (Magis et al., 2017).

In Abbildung 15 ist ein Beispiel zu sehen. Der dargestellte Test besteht aus drei Stufen. Die erste Stufe enthält nur das Routing-Modul mit einer mittleren Schwierigkeit. Die beiden nachfolgenden Stufen bestehen jeweils aus drei Modulen geringer, mittlerer und hoher Schwierigkeit. Proband*innen werden immer zuerst das Routing-Modul abschließen und dann eines der drei Module der zweiten Stufe zugeteilt bekommen. Danach wird eines der in Frage kommenden Module der Stufe drei zugewiesen und der Vorgang für alle eventuell noch folgenden Stufen fortgeführt.

Die gesamte Struktur, die sich aus den Stufen, Modulen und Verzweigungen ergibt, ist ein *Panel* (Luecht & Nungester, 1998). Bei größer angelegten, standardisiert eingesetzten Multistage-Tests werden häufig mehrere Panels parallel erstellt, die alle die gleichen Kriterien erfüllen und möglichst identische psychometrische Eigenschaften aufweisen (Hendrickson, 2007). Bei der Messung wird dann für jedes Individuum eines dieser Panels zufällig ausgewählt und appliziert. So kann Erinnerungseffekten vorgebeugt werden, also die Gefahr durch Itemexposition vermieden werden (Luecht, 2003).

4.3.1 Testlänge und Stufenzahl

Der allgemeine Aufbau eines Multistage-Tests wurde im vorigen Abschnitt 4.3 beschrieben. Bei der Konstruktion eines konkreten Tests sind zu dieser Struktur einige Fragen zu klären, die in diesem und den kommenden zwei Abschnitten 4.3.2 und 4.3.3 besprochen werden. Als erstes stellt

sich die Frage nach der Länge des Tests und der Anzahl an Adaptionspunkten beziehungsweise Stufen.

Werden die Anzahl der Stufen und die gewünschte Modullänge unabhängig voneinander festgelegt, ergibt sich die Gesamtlänge des Tests daraus ganz von selbst. Gibt es hingegen eine feste Gesamtlänge als Ziel bei der Testkonstruktion, stellt sie bei den beiden anderen Entscheidungen eine Zwangsbedingung dar. Es ist dann zwischen der notwendigen Länge einzelner Module und dem erreichbaren Grad der Adaption durch mehrere Stufen abzuwägen.

Somit stellt die Testlänge eine natürliche Obergrenze für die Zahl der Stufen dar, es können aber auch nur zwei Stufen konstruiert werden. Eine allgemeingültige Regel für die Menge der sinnvollen oder notwendigen Stufen existiert nicht, die meisten tatsächlich eingesetzten Instrumente verwenden aber nur drei oder vier Stufen (Hendrickson, 2007; Yan et al., 2014). Die Gründe hierfür werden nicht immer explizit dargelegt. Einer wird aber sicherlich in der Komplexität der Testzusammenstellung liegen, die mit einer größeren Anzahl an Stufen wächst. Zudem konnten vergleichende Studien bisher nicht eindeutig zeigen, dass eine höhere Anzahl von Stufen beziehungsweise Adaptionspunkten zwingend zu einer besseren Messgenauigkeit führt (Luecht & Nungester, 1998).

Die Länge einzelner Module ist grundsätzlich, ebenso wie die Zahl der Stufen, frei wählbar. Ein großer (potenzieller) Vorteil von Multistage-Tests liegt jedoch in der Verknüpfung von mehreren Items, um zum Beispiel für content balance zu sorgen oder die Mehrdimensionalität eines Konstruktes ohne mathematisches Modell zu berücksichtigen. Im Kontext einer inhaltlich validen Testzusammenstellung kann daher die inhaltliche Breite eine Mindestlänge der Module darstellen.

Daneben muss auch berücksichtigt werden, dass in jedem Modul eine ausreichende Messgenauigkeit erreicht werden muss, um ein erfolgreiches Routing zur nächsten Stufe ausführen zu können. Werden alle Proband*innen nach einem Modul in eines von nur zwei weiteren geschickt, sind die Anforderungen relativ gering; im ersten Modul muss nur ausreichend

Information über alle Teilnehmer gewonnen werden, um sie sicher in einen der beiden in Frage kommenden Fähigkeits-/Schwierigkeitsbereiche zu sortieren. Soll hingegen dieselbe Stichprobe in vier Module aufgespalten werden, sind die jeweiligen Fähigkeitsintervalle deutlich kleiner. Entsprechend genauer muss die Einteilung erfolgen, was bei geringer Itemzahl unzuverlässig wird (Cheng & Liou, 2000). Vor allem beim Routing-Modul kann es daher von Vorteil sein, die Länge zu erhöhen, um eine Fehleinschätzung zum Start der Messung zu vermeiden (Kim & Plake, 1994). Für die darauffolgenden Stufen/Module des Tests lässt sich in vergleichenden Studien kein weiterer Einfluss von variierender Modullänge mehr feststellen (Patsula & Hambleton, 1999).

Für keinen dieser Konstruktionsaspekte gibt es eine einfache Regel, die sicher zum Erfolg führen würde. Es ist daher anzuraten, für jeden vorliegenden Fall alle inhaltlichen und messtheoretischen Anforderungen zusammenzustellen. Ergeben sich daraus mehrere mögliche Teststrukturen, können vergleichende Studien zu einer eindeutigen Entscheidung führen.

4.3.2 Anzahl der Module

Neben den Entscheidungen bezüglich der Länge und Aufteilung des Tests in Stufen muss festgelegt werden, in wie viele verschiedene Schwierigkeitsbereiche Proband*innen und Aufgaben eingeteilt werden sollen. Daraus ergibt sich die Anzahl der Module pro Stufe. Ebenso wie bei einer höheren Anzahl der Stufen folgen aus einer höheren Aufspaltung innerhalb jeder Stufe Vor- und Nachteile. Einerseits kann durch die feinere Einteilung in immer kleinere Intervalle eine große Messgenauigkeit erreicht werden (Patsula, 1999). Daraus ergibt sich aber auch direkt eine höhere Anforderung an die Modul- und Testlänge. Außerdem wird die Testzusammensetzung in der Praxis schwieriger (Yan et al., 2014), da für jedes der Module wieder alle gestellten Kriterien zu erfüllen sind.

In der Literaturrecherche zeigten sich verschieden starke Aufspaltungen in eingesetzten Instrumenten. Obwohl sich in Simulationen theoretische Vorteile feiner Aufspaltungen darstellen lassen (Patsula, 1999), weisen gezielte Studien zur Thematik nicht darauf hin. So verweisen beispielsweise

Armstrong, Jones, Koppel und Pashley (2004) darauf, dass in der Praxis oft schon eine Aufspaltung der Stufen in drei Module ausreichend ist und eine Erhöhung auf vier oder mehr Module keinen nennenswerten Effekt erzielt.

4.3.3 Scoring und Routing

Als nächster Aspekt ist zu klären, wie die abgeschlossenen Module beurteilt (*scoring*), die Fähigkeit geschätzt und das nächste Modul ausgewählt (*routing*) werden.

Der gesamte Vorgang hängt direkt mit dem gewählten IRT-Modell zusammen. Wird als Grundlage der Fähigkeitsschätzung ein 1pl- oder Rasch-Modell gewählt, gestaltet sich die Beurteilung der Antworten relativ einfach. Da in diesen Modellen der Informationsgewinn aller Items gleich gewichtet ist, kann anstelle einer numerischen Schätzung der Fähigkeiten auch einfach die Menge der korrekten Antworten gezählt werden – unabhängig davon, welche Items des Moduls tatsächlich korrekt oder falsch gelöst wurden, denn aus mathematischer Sicht spielt dann alleine der Anteil der richtigen Antworten an der Menge der Items eine Rolle (Wright, 1989). Für diesen *Rohscore* können simple Schwellenwerte bestimmte werden, die über das weitere Routing bestimmen. In einem simplen 1-2 Multistage-Test könnte so beispielsweise bestimmt werden, dass Proband*innen mit weniger als der Hälfte der korrekten Antworten im Routing-Modul in das leichtere der beiden nächsten Module geschickt werden, ansonsten in das schwerere.

Im Fall komplexerer IRT-Modelle reicht solch ein Vorgehen nicht mehr aus, da die Information der unterschiedlichen Antworten je nach Item verschieden gewichtet werden muss. Rohwerte tun dies nicht (siehe Abschnitt 3.3), die Bestimmung der Personenfähigkeiten muss daher über ein geeignetes Maß (EAP oder WLE, siehe Abschnitt 3.4.3) numerisch erfolgen. Eine naheliegende Methode ist es, alle Items im Modul dabei einzeln zu behandeln und die Schätzung anhand eines dichotomen Modells (wie in Abschnitt 3.3 vorgestellt) vorzunehmen. Jede Antwort auf ein Item im Modul stellt dann ein unabhängiges Stück Information dar.

Alternativ gibt es Ansätze, bei denen das gesamte Modul mathematisch als ein Item modelliert wird. Dazu wird aus den einzelnen Modulen je ein polytomes Pseudo-Item konstruiert, für das jede Kombination von Antworten auf die verschiedenen Items des Moduls als eine neue Antwort betrachtet wird (van der Linden & Hambleton, 1997). Der Vorteil liegt darin, dass manche der strengen Voraussetzungen für IRT-Modelle innerhalb der Module umgangen werden können. Beispielsweise darf in solchen Fällen die stochastische Unabhängigkeit verletzt werden, was die Konstruktion der Aufgaben erleichtert: Es ist also möglich, einzelne Module zu erstellen, in denen alle Items denselben Itemstamm oder Kontext haben und damit zwangsweise Korrelationen aufweisen.

In jedem Fall ist das Ziel die Bestimmung eines individuellen Schätzwertes für die Personenfähigkeiten, basierend auf allen bisher im Testverlauf gegebenen Antworten. Mit diesem kann das weitere Routing vorgenommen werden. Wie bei Rohwerten ist es üblich, Schwellenwerte zu bestimmen. Da jedes Modul für die Messung in einem beschränkten Fähigkeitsbereich der Stichprobe konstruiert wird, können diese Grenzen als Einordnungspunkte für Proband*innen genutzt werden (Weissman, Belov & Armstrong, 2007; Yan et al., 2014).

Unter besonderen Umständen wird hierbei noch eine Einschränkung vorgenommen: Yan et al. (2014) verweisen darauf, dass Fehleinschätzungen von Versuchsteilnehmer*innen zu extremen Sprüngen zwischen den Stufen eines MST führen können. Das kann vor allem bei Personen in den Extrembereichen des Fähigkeitsspektrums auftreten, da hier die gewonnene Informationsmenge pro Item auch in einem adaptiven Testformat gering ausfällt. So könnte in einem 1-3-3-Design eine extrem leistungsstarke Person zunächst korrekt in das schwere Modul der zweiten Stufe geschickt werden. Bedingt durch den geringen Informationsgewinn an den Randbereichen der Fähigkeitsverteilung und die damit insgesamt niedrige Informationsmenge am Ende der zweiten Stufe kann es dann passieren, dass bereits wenige falsche Antworten im zweiten Modul zu einer Übersteuerung führen und die Person in das leichte Modul der dritten Stufe geschickt würde. Um solchen Ereignissen entgegenzuwirken, ist es gängige

Praxis, nur Übergänge in benachbarte Module zu erlauben, also beispielsweise *leicht-mittel*, und größere Sprünge wie *schwer-leicht* zu verbieten (Sari, Yahsi-Sari & Huggins-Manley, 2016).

4.4 Vergleiche zwischen linearen und adaptiven Formaten

In den vorangegangenen Abschnitten (vgl. 4.3.1 bis 4.3.3) ist zu erkennen, dass die Konstruktion eines adaptiven Tests, ob nun CAT oder MST, einen Mehraufwand im Vergleich zu linearen Tests darstellt. Dieser Aufwand wird bei der Implementation adaptiver Verfahren vor allem durch die höhere Messgenauigkeit bei gleicher Testlänge, also eine höhere *Effizienz* (nach Segall, 2005), gerechtfertigt. Während dieser Vorteil adaptiver Verfahren im Allgemeinen in der Literatur kaum noch angezweifelt wird, ist aber vor allem der Vergleich von CATs und MSTs untereinander noch Bestandteil anhaltender Diskussion. Zudem bestehen Spekulationen über den Einfluss der unterschiedlichen Testformate auf die Motivation von Testteilnehmer*innen.

4.4.1 Effizienz

Zum Vergleich zwischen linearen Tests und CATs gibt es seit dem vermehrten Aufkommen von CATs in der Praxis eine Vielzahl an Publikationen (vgl. Frey & Ehmke, 2008; Green, 2012; Olsen, 1986; Schnipke & Reese, 1997; Weiss, 1982). Sie zeigen, dass bei der Messung durch CATs im Mittel etwa halb so viele Items benötigt werden, um die gleiche Messgenauigkeit zu erreichen wie lineare Tests.

Direkte Vergleiche zwischen MSTs und CATs sowie MSTs und linearen Formaten sind bisher noch seltener. Mit den Arbeiten von Patsula (1999) und Rotou, Patsula, Steffen und Rizavi (2007) liegen aber umfangreiche Vergleiche zur Effizienz der drei besprochenen Testformate innerhalb derselben Studien vor. Auch hier zeigt sich ein eindeutiger Vorteil von beiden adaptiven Formaten gegenüber linearen Tests. Allerdings ist auch eine höhere Effizienz von CATs gegenüber MSTs festzustellen. Dabei hängt es von der Struktur des konkreten MSTs ab, wie groß der Abstand zu den jeweils anderen Testformaten ausfällt. Patsula (1999) schließt aus ihren Daten, dass sich MSTs mit höherer Aufspaltung in Stufen und/oder Module nicht

nur strukturell immer weiter einem CAT annähern, sondern gleichzeitig in ihrer Effizienz. Dabei waren auch die strukturell simpelsten MSTs ihrer Studie (1-3-Designs) noch um 10-40% effizienter als der entsprechende lineare Test, wenn auch signifikant ineffizienter als der CAT.

Allerdings konnte Wang (2017) in seiner Dissertation auch simple MSTs konstruieren (1-2-3- und 1-3-3-Designs), die bei gleicher Testlänge noch kleine, aber nicht signifikante Effizienzeinbußen gegenüber CATs zeigten. Die zwingende Bedingung hierfür war die „forward assembly“ (Wang, 2017) der Testmodule, also die gezielte Verwendung der psychometrisch betrachtet besten (informativsten) Items zu Beginn des Tests.

4.4.2 Testmotivation

Bezüglich der Auswirkungen adaptiver Formate auf die Motivation von Proband*innen beschreiben Frey, Hartig und Moosbrugger (2009) einen Umstand, der sehr kritisch betrachtet werden sollte: Bereits seit mehreren Jahren hält sich die Meinung, adaptive Formate hätten einen motivationssteigernden Effekt. Sie führen das hauptsächlich auf eine Arbeit von Betz und Weiss (1976) zurück, da a) dieser Effekt dort berichtet wird, b) sich die Behauptung seit dieser Publikation hält und c) kaum andere Evidenz hierfür zu finden ist.

In einer eigenen Studie weisen sie allerdings einen gegenteiligen Effekt nach: Die Motivation zur Testbearbeitung fiel dabei im CAT signifikant niedriger als im linearen Test. Insbesondere scheint sich dieser Effekt durch die von Proband*innen subjektiv eingeschätzte Erfolgschance zu begründen, nicht etwa durch die empfundene Herausforderung. Der Befund reiht sich in weitere Arbeiten aus vorangegangenen Jahren ein (Bergstrom, Lunz & Gershon, 1992; Eggen, 2004; Eggen & Verschoor, 2006), in denen dieser Umstand auf theoretischer Basis vermutet wurde. Ebenfalls vermutet, aber leider ungeprüft, bleibt die Hypothese, dass eine transparente Darstellung der adaptiven Testfunktion vor Beginn der Messung den negativen Effekten entgegenwirken könnte (Frey et al., 2009).

5 Problemstellung und Forschungsfragen

In den Kapiteln 3 und 4 wurde der theoretische Rahmen der Arbeit dargestellt.

In Abschnitt 5.1 wird nun beschrieben, welchen Problematiken den Ausgangspunkt der Studien bilden. Danach werden in Abschnitt 5.2 die Forschungsfragen formuliert.

5.1 Problemstellung

Das Problemfeld wurde in Kapitel 1 bereits grob umrissen. Der Ko-WADiS-Test ist ein fertig entwickeltes und einsatzbereites Testinstrument: Die im Rahmen der Entwicklungs- und Validierungsphase durchgeführten Studien weisen alle auf eine valide Testwertinterpretation als Maß für die Kompetenz naturwissenschaftlichen Denkens hin. Der zur Messung genutzte Itempool deckt die Facetten des Konstrukts ab und ist auch in Bezug auf die Kontexte individueller Aufgaben breit gefächert. Durch den Einsatz verschiedener Testhefte werden Erinnerungseffekte und ähnliche Komplikationen in wiederholten Messungen verhindert.

Es bestehen dennoch je nach Testeinsatz zwei Probleme, um die es nun gehen soll. Für das erste Problem liegt nur anekdotische Evidenz[12] vor, die jedoch nicht ignoriert werden sollte. Die eingesetzten Fragebögen scheinen eine sehr starke kognitive Belastung aufseiten der befragten Proband*innen zu verursachen. Sie haben einen Umfang von 21 Items, wofür im Laufe des Projekts der pauschale Zeitaufwand von 45 Minuten veranschlagt und meist auch benötigt wurde. Da es sich bei den einzelnen Items um recht komplexe Aufgaben im Bereich logischen Denkens handelt, ist die geforderte Konzentration entsprechend hoch einzuschätzen. Es muss aber vermerkt werden, dass hierzu nie explizite Studien durchgeführt wurden: Die Feststellung des Problems beruht alleine auf Beobachtungen seitens der Testleiter*innen und vereinzelter Rückmeldungen von Proband*innen.

[12] Als anekdotische Evidenz werden im Folgenden vereinzelte, nicht gezielte Beobachtungen beschrieben, die im Rahmen der Projektarbeit gemacht wurden. Sie sind qualitativer Art und nicht empirisch gesichert, allerdings decken sich die angeführten Erfahrungen bei allen Beteiligten.

Weiterreichende Folgen, beispielsweise Testabbrüche oder ähnliches, gab es nicht.

Das zweite Problem ist die geringe Reliabilität des Tests. Sie liegt im Mittel für den gesamten Datensatz bei .544 (Hartmann, Mathesius et al., 2015). Für die Beobachtung von langfristigen Trends scheinen diese Werte auszureichen. Dies war auch der hauptsächliche Einsatzzweck des Tests im Rahmen der Forschungsprojekte Ko-WADiS und ValiDiS. Für detailliertere Untersuchungsszenarien, beispielsweise Vergleichsstudien in kleinen Gruppen oder über kurze Zeiträume, ist die Messgenauigkeit aber als kritisch zu bezeichnen. Hierfür werden in der Literatur als Mindestmaß Werte zwischen .6 (vgl. Blömeke, 2008) und .8 (vgl. Bortz & Döring, 2006) genannt.

Von den beiden genannten Problemen ist das zweite definitiv als schwerwiegender zu beurteilen. Zum einen liegt das an der sicheren Datenlage, mit der es begründet werden kann. Es kann damit eindeutig festgestellt und in seiner Schwere beurteilt werden; im Gegensatz zu der scheinbar kritischen Belastung der Proband*innen. Zum anderen sind die Konsequenzen für den Testeinsatz größer. Ein anstrengender Test, der dafür genaue Ergebnisse liefert, kann im Einsatz sinnvoll und begründbar sein. Ein leichter Test, der dafür keine verwertbaren Aussagen liefert, nicht.

Eine übliche Möglichkeit, die Reliabilität eines Messinstruments zu erhöhen, ist die Erweiterung des Itemkatalogs und Verlängerung des Testheftes. Mehr Items bedeuten mehr Antworten der Proband*innen und damit mehr Informationen über das untersuchte Konstrukt. Die Messgenauigkeit kann auf diese Weise theoretisch beliebig erhöht werden. Der umfangreiche Pool an konstruierten Items würde diese Lösung im Fall des Ko-WADiS-Tests leicht machen.

Leider würde durch eine Verlängerung des Tests die Problematik der Probandenbelastung verstärkt. Wo es bisher bei vereinzelten Rückmeldungen und einer beobachtbaren Unruhe gegen Ende der Erhebungen blieb, müsste mit sichtbaren Effekten gerechnet werden. Um allein durch höhere Itemzahlen in den Bereich der nach Literatur gewünschten

Messgenauigkeit zu kommen, müssten die einzelnen Testhefte auf eine Länge von mindestens 30 Items vergrößert werden. Der Zeitaufwand würde hier eine Stunde übersteigen. Dieser Ansatz erscheint also nicht praktikabel.

Grundsätzlich wäre eine weitere Möglichkeit die Optimierung der Items selbst. Gäbe es auf einer theoretischen Basis stehende Kritik an der Form der Items, gehäufte Mängel in der Formulierung des Itemstamms oder ähnliche Angriffspunkte, müsste eine Überarbeitung der Items durch Experten*innengruppen in Betracht gezogen werden. Solche Überarbeitungen stellen jedoch einen hohen Aufwand dar, zudem wurde genau dieser Schritt schon als Teil der ersten Projektphase und Testentwicklung vorgenommen. Zum jetzigen Zeitpunkt gibt es keine Kritik, die ernsthaften Zweifel an der Güte der einzelnen Items aufkommen ließe. Die Erhöhung der Messgenauigkeit einzelner Items durch eine inhaltliche Überarbeitung des Instruments scheint also keine Option zu sein.

Damit wird im Übrigen eine Eigenschaft angesprochen, die bereits im Abschnitt 4.4.1 dieser Arbeit beschrieben wurde: Die Effizienz des Messinstruments, nach Segall (2005) definiert als die Messgenauigkeit einer Messung über ihre Länge:

$$Effizienz = \frac{Messgenauigkeit}{Testlänge}$$

Diese Gleichung fasst die beide Probleme des Ko-WADiS-Tests und ihren widersprüchlichen Charakter noch einmal anschaulich zusammen: Solange die Effizienz konstant bleibt, kann entweder die Messgenauigkeit erhöht oder die Länge und damit die Belastung der Proband*innen reduziert werden. Beides ist aber nicht sinnvoll umsetzbar, da es das jeweils andere Problem in gleichem Maß verstärken würde.

Die Effizienz des Testinstruments muss also erhöht werden, ohne dass die Items verändert werden können. In Kapitel 4 wurde bereits gezeigt, wie solch eine Verbesserung von Testinstrumenten möglich ist: Adaptive Testverfahren haben im Allgemeinen eine höhere Effizienz als ihre linearen

Gegenstücke, auch wenn sie aus den exakt selben Itempools konstruiert werden. In Abschnitt 4.4.1 wurde dieser Umstand beschrieben.

In der Konstruktion einer adaptiven Version des Ko-WADiS-Tests besteht deshalb potenziell die Möglichkeit, die gewünschten Verbesserungen in beiden Aspekten gleichzeitig vorzunehmen. Das Ziel der Arbeit ist daher die Entwicklung und Erprobung einer solchen Version. Dazu ist zu klären, welche Schritte hierzu notwendig sind und welche Fragen sich in Bezug auf den Konstruktionsprozess und die finale Testversion stellen.

5.2 Forschungsfragen

Die Konstruktion eines neuen Tests ist üblicherweise ein sehr zeitaufwändiges Verfahren (Terzer, Hartig & Upmeier zu Belzen, 2013). Da in diesem Fall bereits inhaltstheoretische Grundlagen und geprüfte Items zur Verfügung stehen, kann ein großer Teil der sonst notwenigen Schritte im Prozess ausgelassen werden. Die Entwicklung der adaptiven Testversion wird daher nach größtenteils psychometrischen beziehungsweise statistischen Kriterien vorgenommen. Hierfür stellen die bereits gewonnen Daten aus den projektinternen Studien (Ko-WADiS und ValiDiS, vgl. Abschnitt 2.4) eine sichere Basis dar. In Bezug auf die Frage der Testvalidierung wurden in Kapitel 2 die schon durchgeführten Studien diskutiert. Das Validitätsargument wird außerdem in der abschließenden Diskussion der Arbeit in Kapitel 9 noch einmal aufgegriffen.

So groß der Vorteil des vorher erstellten Itempools sein mag, so stellt er eine Besonderheit dar und verursacht potenzielle Einschränkungen. Wie in der zuvor zitierten Literatur beschrieben, startet die Konstruktion von adaptiven Tests mit einer theoretischen Betrachtung des gewünschten Testszenarios. Hierbei werden elementare Fragen geklärt, wie beispielsweise

a) Wie häufig werden die selben Proband*innen befragt?
b) Wie wahrscheinlich sind Betrugsversuche?
c) Sind besonders die Mittelpunkte oder Ausreißer von Stichproben interessant?
d) Welche individuellen Konsequenzen resultieren aus der Befragung?

Aus den Antworten zu diesen Fragen werden dann Anforderungen an den Test entwickelt, aus denen sich wiederum logische Konsequenzen für die Gestaltung und Umsetzung ableiten lassen. Das betrifft alle in Kapitel 4 beschriebenen Bestandteile des Tests, also die Abschlusskriterien, die Struktur des Tests und nicht zuletzt den Testalgorithmus (also das verwendete IRT-Modell, vgl. Abschnitt 4.2.1) sowie den Umfang und Aufbau des notwendigen Itempools. Erst nach all diesen Überlegungen beginnt die Entwicklung von Items, bis der geforderte Itempool vollständig mit – zum entsprechenden IRT-Modell passenden – Aufgaben gefüllt wurde.

Es liegt aber eine genau umgekehrte Situation vor, denn die Items wurden bereits erstellt. Die Abwägungen der diversen Konstruktionsaspekte bekommen damit durch die existierenden Grenzen des Itempools eine Zwangsbedingung auferlegt. Dadurch werden selbstverständlich nicht die Vor- und Nachteile der unterschiedlichen Verfahren verändert, wie beispielsweise von Konfidenz- und Präzisionskriterium zur Beendigung einer Messung (vgl. Abschnitt 4.2.3). Aber einige der Optionen werden durch die Beschränkungen von vornherein ausgeschlossen sein. Besonders zu erwarten ist dieser Ausschluss bei allen Möglichkeiten zur Verringerung der Itemexposition, wofür die Konstruktion paralleler Testversionen und entsprechend viele Items nötig sind.

Deswegen sind im Vorfeld aller Entscheidungen mehrere Schritte wichtig. Zunächst sollte noch einmal eine grundlegende psychometrische Überprüfung des Itempools auf Grundlage der bisher erfassten Daten des linearen Instruments erfolgen. Selbstverständlich ist das im Rahmen der Testentwicklung schon früher geschehen. Seitdem wurden allerdings über einen längeren Zeitraum große Mengen an Daten gewonnen. Einerseits ist es damit möglich, die Güte der einzelnen Items noch genauer zu untersuchen und so die Qualität des Testinstruments durch eventuell strengere Selektion zu erhöhen. Andererseits erlauben höhere Datenmengen typischerweise die Arbeit mit komplexeren mathematischen Modellen, was einen Einfluss auf die Messgenauigkeit haben kann. Es ist festzustellen, ob durch den aktuellen Datensatz mehr IRT-Modelle zur Auswahl stehen als zum

Zeitpunkt der Testentwicklung. Aus diesen kann dann der vielversprechendste Testalgorithmus des adaptiven Verfahrens gebildet werden.

> **FF 1**: Welches der ausgewählten IRT-Modelle (vgl. Abschnitt 3.3) ist für die Beschreibung der Ko-WADiS Daten am besten geeignet?

Diese Frage wird durch keine eigene Studie, sondern allein durch mathematische Analysen der schon vorliegenden Daten beantwortet. Als Maß für die Güte sowie die Auswahl des Modells dienen folgende Kriterien:

a) Das Modell muss nach allgemeinen Kriterien der Modellgüte die Daten gut beschreiben (siehe dazu die methodischen Ausführungen im Abschnitt 6.2).
b) Das Modell muss einen im Vergleich zu den Alternativen möglichst großen und diversen Itempool ermöglichen. Anders formuliert: Es sollen möglichst wenige verfügbare Items durch schlechte Passung zum jeweiligen Modell ausgeschlossen werden.
c) Von allen Modellen, die nach Kriterien a) und b) gleichwertig sind, ist das Modell mit der größten Messgenauigkeit auszuwählen (siehe dazu die methodischen Ausführungen in Abschnitt 6.3).

Nach Beantwortung der ersten Forschungsfrage wird nicht nur das verwendete Modell feststehen, sondern auch der Umfang und die Zusammensetzung des verfügbaren Itempools. Somit können im Anschluss alle Überlegungen zur Ausgestaltung der adaptiven Testversion stattfinden. Besondere Relevanz hat hier die Frage des effizientesten Formats, das praktisch umsetzbar ist, ins besonders sind CATs und MSTs zu vergleichen.

> **FF 2:** Welches adaptive Testverfahren eignet sich am besten für die Effizienzsteigerung des Ko-WADiS-Tests?

Da diese Fragestellung für sich gesehen äußerst umfangreich ist, soll sie in mehrere Bestandteile aufgespalten werden. Jede der folgenden Fragen wird für sich getrennt bearbeitet und stellt einen Teilaspekt von FF 2 dar; durch die Beantwortung aller Unterpunkte kann in einer abschließenden Diskussion dann FF 2 beantwortet werden.

FF 2.1: Ist ein CAT oder ein MST zur Effizienzsteigerung des Ko-WADiS-Tests besser geeignet?

FF 2.2: Welcher Testalgorithmus (siehe Abschnitt 4.2) verspricht die größte Effizienzsteigerung bei der Messung mit einem Ko-WADiS-CAT?

FF 2.3: Welche Teststruktur und Routing-Regeln (siehe Abschnitt 4.3) versprechen die größte Effizienzsteigerung bei der Messung mit einem Ko-WADiS-MST?

Zur Beantwortung von FF 2.1 muss in einem ersten Schritt diskutiert werden, welche Anforderungen an den späteren Einsatz des Instruments gestellt werden und inwieweit – bezogen auf diese Anforderungen – eine Umsetzung beider Alternativen praktikabel ist. Sofern durch theoretische Vorüberlegungen keine Entscheidung getroffen werden kann, müssen beide Verfahren zunächst für sich optimiert werden (siehe FF 2.2 und FF 2.3). Falls aber schon im Voraus eine Wahl getroffen werden kann, wird nur eine der beiden Forschungsfragen 2.2 und 2.3 bearbeitet. Bei der Beantwortung von FF 2.2 und/oder 2.3 ist davon auszugehen, dass rein theoretische Überlegungen nicht zielführend sein werden. Wie in Abschnitt 4.4.1 beschrieben, konnte bisher keine allgemein effizienteste Form von adaptiven Verfahren identifiziert werden. Besonders verschiedene Teststrukturen von MSTs zeichnen hier kein allgemeingültiges Bild, es wird also die Konstruktion verschiedener Testversionen sowie eine vergleichende Untersuchung notwendig sein.

Nach der Bearbeitung von FF 2 wird die praktische Umsetzung und Implementierung der adaptiven Testversion erfolgen. Im Anschluss daran können beide Versionen, die neue adaptive und alte lineare, in einer Vergleichsstudie verglichen werden. Ziel ist die Überprüfung der Effizienz beider Instrumente und damit die Beantwortung der letzten Frage:

FF 3: Ist die adaptive Version des Testinstruments signifikant effizienter als die lineare Version?

Die mathematische Formulierung des zugehörigen Effizienzkriteriums erfolgt in den methodischen Erläuterungen zu dieser Studie in Kapitel 8 beziehungsweise in Anteilen auch zuvor in Abschnitt 6.3.

6 Auswahl des IRT-Modells (FF 1)

Mit dem vorigen Kapitel (5) wurden die Problemstellung der Arbeit und die Formulierung der Forschungsfragen festgelegt. Die verschiedenen Forschungsfragen werden nun jeweils in einem eigenen Kapitel beantwortet.

Es wurde schon darauf hingewiesen, dass für die Beantwortung von FF 1 keine empirische Studie mehr durchgeführt werden muss. Die notwendigen Daten wurden schon im Vorfeld dieser Arbeit gewonnen und sollten für den Zweck mehr als ausreichen (vgl. Kapitel 2). Um dies sicherzustellen, werden die vorhandenen Daten im ersten Abschnitt bereinigt und in Bezug auf die späteren Studien eingegrenzt.

Als grundsätzlich geeignete Familie von Modellen wurden die bereits in Kapitel 3 beschriebenen logistischen Modelle ausgesucht. Es handelt sich dabei um die verbreitetste Methode, um Daten dichotomer Items mittels IRT zu beschreiben. Sie sind gut dokumentiert und in verschiedener Software zu Auswertungen verfügbar. Hinzu kommt, dass sie auch schon in der bisherigen Testentwicklung verwendet wurden.

Danach werden der Modellierungsprozess und die verwendeten Kriterien zur Modellgüte in einem methodischen Abschnitt erläutert. Auf die dazu notwendigen Konzepte und Formeln aus der IRT, die in Kapitel 3 noch nicht betrachtet wurden, wird hier kurz eingegangen. Der zweite Abschnitt dieses Kapitels wird damit einen mathematisch-theoretischen Schwerpunkt haben. Er kann auch als eine Vertiefung des Theoriekapitels 3 betrachtet werden.

Die Ergebnisse der tatsächlichen Berechnungen werden für alle Modelle in aufeinanderfolgenden Abschnitten einzeln dargestellt, um die Übersicht zu wahren. Danach folgt erneut eine methodische Erläuterung, diesmal für die Vergleichsmaße der verschiedenen Modelle untereinander.

Den Abschluss des Kapitels werden die endgültige Auswahl des IRT-Modells und die Darstellung des daraus resultierenden Itempools bilden.

Sämtliche Berechnungen wurden mittels der Software R (R Core Team, 2020, Version 4.0.0) in der Entwicklungsumgebung RStudio (Version

1.3.959) durchgeführt. Das speziell zur Schätzung von IRT-Modellen verwendete Paket ist TAM (Robitzsch, Kiefer & Wu, 2020, Version 3.5-19). Das dabei eingesetzte Maximum-Likelihood-Verfahren ist das der Marginal Maximum Likelihood (MML, siehe Abschnitt 3.4.1).

6.1 Datenbereinigung

Zum Zeitpunkt dieser Auswertung umfasste der zur Verfügung stehende Datensatz Einträge von 191 Items und 8873 Testheften. Hierin sind auch frühere Entwicklungsversionen von Items enthalten, die nach Pilotierungsphasen verworfen wurden.

Insgesamt stammen die Daten von 6792 Proband*innen. Davon wurden 1190 im Rahmen von Längsschnittstudien mehrfach befragt und haben daher mehrere Einträge im Datensatz. Im Umgang damit gibt es verschiedene Möglichkeiten. Sofern aus der Datenanalyse Rückschlüsse auf die einzelnen Proband*innen gezogen werden sollen, müssen eindeutige Identifikatoren beibehalten oder eines der bearbeiteten Testhefte ausgewählt werden, damit keine widersprüchlichen Beurteilungen derselben Personen getroffen werden. Das Ziel der kommenden Betrachtungen ist aber lediglich die Suche nach einem passenden Modell und die Beurteilung der Items, solche Bedenken bestehen daher nicht.

Zudem wurden die Mehrfachmessungen zu unterschiedlichen Zeitpunkten durchgeführt. Die Arbeit von Straube (2016) weist auf ein Ansteigen der Kompetenz im Verlauf des Studiums hin, dabei handelt es sich um das einzige modellierte Personenmerkmal. Im Sinne des Testinstruments ist also eine Person zu einem späteren Zeitpunkt leistungsfähiger als bei einer frühen, ersten Messung – und damit eine andere Person. Es werden daher für die Modellberechnung alle Datensätze so behandelt, als würden sie von unterschiedlichen Personen stammen.

Auf dieser Grundlage muss nun aufgebaut werden. Im Laufe der Zeit wurden in den Datensatz verschiedene Probandengruppen integriert, darin eingeschlossen sind verschiedene Universitätsstandorte und unterschiedliche Fächergruppen aus dem Bereich der MINT-Fächer im Lehramts- sowie Monostudium.

In den Kapiteln 3 und 4 wird der Zusammenhang zwischen Item- und Personenparametern und dem möglichen Informationsgewinn beschrieben. Wie dort ersichtlich wird, muss die Passung möglichst gut sein, um die erreichte Information der Messung zu maximieren. Adaptive Testverfahren versuchen diese Passung im Laufe der Erhebung herzustellen. Aber natürlich ist es einfacher, wenn wenigstens die mittlere Schwierigkeit des Testinstruments und die mittlere Personenfähigkeit der befragten Stichprobe übereinstimmen.

Daher soll der Test im Hinblick auf die späteren Befragungen, vor allem die kommenden Vergleichsstudien, konstruiert werden. So kann der tatsächlich erreichbare Unterschied zwischen der Effizienz beider Testversionen möglichst gut dargestellt werden.

Am Standort der Testentwicklung (Freie Universität Berlin) umfasst die erreichbare Probandengruppe die der Lehramtsstudierenden in den Fächern Physik und Biologie. Grundsätzlich wären auch Monostudierende erreichbar, die Absprache ist hier aber komplizierter und eine erfolgreiche Durchführung der Studien bei ihnen nicht garantiert. Zudem gibt es zwischen beiden Studiengruppen je nach Studienphase deutliche Unterschiede in der Kompetenzausprägung und Leistungsfähigkeit (Straube, 2016). Um eine homogene Stichprobe garantieren zu können, muss also eine dieser Gruppen ausgesucht werden. Die Wahl fällt auf die der Lehramtsstudierenden, da sie der Fokus bei der bisherigen Testkonstruktion waren (N = 5548).

Die erste Einschränkung des Datensatzes sollen also darin bestehen, dass das Instrument nur auf Lehramtsstudent*innen der Naturwissenschaften an der Freien Universität Berlin zugeschnitten wird. Diese werden die Datenlage bilden und als Normierungsgruppe dienen.

Dazu kommen noch ein paar weitere Einschränkungen aus psychometrischen Gründen. Erstens ist es wichtig, dass alle Items anhand des numerischen Modellierungsverfahrens (welches auch immer ausgesucht wird, dazu mehr im kommenden Abschnitt) genau geschätzt werden können. Ob dies funktioniert ist davon abhängig, ob für jedes einzelne Item genug

Antworten zur Verfügung stehen. Linacre (1994) nennt 100 Antworten pro Aufgabe als Anforderung, um Aufgaben mit einer Wahrscheinlichkeit von 95% im Bereich eines halben Logits (siehe 3.3.1) um ihren wahren Wert einschätzen zu können.

Zweitens ist die Güte (im Mangel eines besseren Worts soll dieser Begriff hier reichen) der einzelnen Antworten im Datensatz sicherzustellen. Wird eine Befragung unter starkem Zeitdruck oder unmotiviert ausgeführt, kann das zu unüberlegten oder geratenen Antworten führen. Diese sind im Nachhinein nur schwer anhand des Datensatzes zu erkennen. Es lässt sich aber wenigstens argumentieren, dass unvollständig ausgefüllte Testhefte (besonders wenn sie abgebrochen wurden oder viel übersprungen wurde) ein Indikator für eine unsaubere Befragung sein können. Garantiert ist dieser Rückschluss nicht, es kann auch andere Gründe für ein solches Verhalten geben. Bei der Größe des Datensatzes wird an dieser Stelle aber entschieden, lieber einen kritischen Fragebogen zu viel als einen zu wenig auszuschließen. Daher werden alle Daten aus unvollständig bearbeiteten Testheften ausgeschlossen.

Der Datensatz für die ausgewählte Population umfasst mit diesen Einschränkungen noch 141 Items mit mindestens 100 Bearbeitungen und 4073 Personen mit jeweils 21 erfassten Antworten.

6.2 Methodik – Individueller Modellfit

Nach der ersten numerischen Schätzung eines IRT-Modells stellt sich die Frage, wie gut das verwendete Rechenmodell die beobachteten Daten beschreibt. Diese Passung zwischen Beobachtung und numerischen Daten wird als *Modellkonformität* oder als *Fit* bezeichnet. Um die Modellkonformität zu beurteilen, gibt es diverse verschiedene Testverfahren, Kennwerte und zugehörige Wertgrenzen, innerhalb derer Modelle als akzeptabel betrachtet werden.

Die meisten Tests für Modellkonformität basieren auf Abstandsmaßen. Es wird zuerst bestimmt, welche Antworten nach den numerisch bestimmten Parametern auftreten sollten. Danach wird der Abstand beziehungsweise die Abweichung dieser Ergebnisse zu den real beobachteten Daten

errechnet. Jedes dieser Verfahren ist grundsätzlich geeignet, den Fit einer einzelnen Person, Aufgabe oder des gesamten Datensatzes zu überprüfen. Dennoch gibt es einzelne Verfahren, die typischerweise nur für die Prüfung einer dieser Spielarten herangezogen werden. Das kann sowohl echte mathematische als auch rein traditionelle Gründe haben.

Des Weiteren ist zu erwähnen, dass der Modellfit als Ganzes geprüft werden kann, indem ein Testverfahren über den gesamten Datensatz oder auch in Einzelschritten spalten- oder zeilenweise (also personen- oder aufgabenbezogen) angewandt wird. Das bedeutet, dass nicht immer zwingend ein Test für das gesamte Modell erfolgen muss, sofern der Fit jeder einzelnen Person auf je alle Items oder aller einzelnen Items mit den Daten jeweils aller Probanden erfolgt ist.

Hier ist es so, dass die einzelnen Items von besonderem Interesse sind. Sie bilden die Grundlage für die weitere Konstruktion des adaptiven Tests, einzelne Proband*innen sind hierfür nicht von Belang. Deshalb findet die Prüfung des Fits anhand der einzelnen Items statt.

Bei den in den folgenden Abschnitten genannten Fit-Statistiken handelt es sich um diejenigen, die in der IRT-Literatur am besten etabliert sind und gleichzeitig keine Verzerrungen bei der vorliegenden Stichprobengröße zeigen.

6.2.1 Infit und Outfit

Zwei verbreitete Fit-Statistiken im Rahmen der IRT sind *Infit (inlier-sensitive fit)* und *Outfit (outlier-sesitive fit)*. Es handelt sich grundlegend um Chi-Quadrat-Statistiken, geteilt durch die Anzahl der jeweils relevanten Freiheitsgrade (Wright & Masters, 1982). Mit ihnen wird die Passung beobachteter und erwarteter Daten auf Basis der Residuen aller Beobachtungen geprüft.

Das gewichtete Residuum z_{ij} ist die Abweichung zwischen erwarteter und beobachteter Antwort x_{ij} von Person i auf Aufgabe j, geteilt durch ihre Standardabweichung (der modellierten Werte um den Erwartungswert) (Wright & Masters, 1982):

$$z_{ij} = \frac{x_{ij} - E(x_{ij})}{\sqrt{Var(x_{ij})}}$$

Mit dem Outfit wird nach Ausreißern gesucht, also nach Aufgaben oder Personen, bei denen die Residuen im Mittel sehr groß sind. Er ist definiert als der Durchschnitt der quadratischen Residuen (Wright & Masters, 1982):

$$Outfit_j = \frac{\sum_I z_{ij}^2}{I}$$

wobei I in diesem Fall für die Menge an Personen steht, die Item j bearbeitet haben.

Mit dem Infit wird dagegen nach Aufgaben oder Personen gesucht, deren Residuen fälschlicherweise innerhalb eines erwarteten Bereichs liegen. Hiermit können beispielsweise *Guttman-Muster* aufgedeckt werden, welche die rechnerische Auswertung stark erschweren. Infit ist definiert als (Wright & Masters, 1982):

$$Infit_j = \frac{\sum_I Var(x_{ij}) * z_{ij}^2}{\sum_I Var(x_{ij})}$$

Die Verteilung beider Teststatistiken ist jeweils die Chi-Quadrat-Verteilung. Die Erwartungswerte der Statistiken sind 1. Werte über dem Erwartungswert weisen auf nicht modellierte, aber relevante Variablen und damit einen *underfit* des Modells hin. Werte unter 1 sind hingegen Anzeichen dafür, dass neben den tatsächlich relevanten Variablen auch irrelevante Größen modelliert werden. Dieser *overfit* ist nicht direkt schädlich für die Messung, er kann aber zu einer fälschlichen Erhöhung der bestimmten Messgenauigkeit führen.

Es finden sich in der Literatur verschiedene Akzeptanzbereiche für beide Statistiken (vgl. Ayala & Kenny, 2009), häufig wird aber auf Smith, Schumacker und Bush (1998) verwiesen. Sie empfehlen für Infit und Outfit Akzeptanzbereiche von jeweils $1 \pm 2/\sqrt{N}$ und $1 \pm 6/\sqrt{N}$.

Infit und Outfit sind Maße für die Größe von zufälligen Modellabweichungen. Sie gehen der Frage nach, ob die Abweichung der modellierten von den empirischen Daten in einem für die Testauswertung praktikablen Rahmen liegen.

In der Gegenrichtung könnte aber auch geprüft werden, ob das Modell signifikant von den Daten abweicht. Es werden also beidseitige Signifikanztests zur Hypothese durchgeführt, dass Infit oder Outfit gleich ihrem Erwartungswert 1 sind. Zu diesem Zweck werden Infit und Outfit zuerst z-transformiert[13], also von der Chi-Quadrat-Verteilung in eine Standardnormalverteilung überführt.

Die Teststatistiken der t-Tests berechnen sich wie folgt (Wright & Masters, 1982):

$$Outfit_{j,zstd} = \left(Outfit_j^{\frac{1}{3}} - 1\right)\left(\frac{3}{Var\left(Outfit_j\right)}\right) + Var\left(\frac{Outfit_j}{3}\right)$$

$$Infit_{j,zstd} = \left(Infit_j^{\frac{1}{3}} - 1\right)\left(\frac{3}{Var\left(Infit_j\right)}\right) + Var\left(\frac{Infit_j}{3}\right)$$

Die z-standardisierten Werte sind mit einem Erwartungswert von 0 und einer Standardabweichung von 1 verteilt. Sie eignen sich damit für eine direkte Beurteilung der Signifikanz von Modellabweichungen durch p-Werte (Schulz, 2002). Negative Werte weisen auf einen overfit, positive Werte auf einen underfit hin.

Die Berechnung aller besprochenen Maße ist für Personen und Aufgaben identisch, es müssen für die Bestimmung von Personen-Infit und -Outfit lediglich alle Indizes der gezeigten Formeln getauscht werden.

6.2.2 RMSD

Die *Root Mean Square Deviation* (RMSD), teilweise auch als *Root Mean Square Error* (RMSE) oder in älteren Quellen *Root Mean Square Error of Approximation* (RMSEA) bezeichnet, ist eine weitere Fit-Statistik, die sowohl innerhalb der IRT als auch der KTT häufige Anwendung findet. Sie beruht auf der Chi-Quadrat-Statistik (X^2) und berechnet sich wie folgt (Tennant & Pallant, 2012):

[13] Die Übertragung von Werten einer Chi-Quadrat-Verteilung in die Standardnormalverteilung kann mittels der Wilson-Hilferty-Transformation (Wilson & Hilferty, 1931) vorgenommen werden.

$$RMSD = \sqrt{\frac{\mathrm{X}^2 - df}{df(N-1)}}$$

dabei ist df die Menge der Freiheitsgrade und N die Menge der Beobachtungen. Wird die RMSD nur für ein einzelnes Item berechnet, entspricht df der Anzahl an geschätzten Itemparametern und N der Menge an tatsächlich gegebenen Antworten, nicht der Größe der Gesamtstichprobe.

RMSD ist genau ebenso wie Infit und Outfit ein absolutes Maß für den Fit von Modelldaten und lässt sich ebenfalls auf die Chi-Quadrat-Statistik zurückführen. Im Gegensatz zu den anderen beiden Maßen wird bei der Berechnung der RMSD jedoch die Größe der Stichprobe berücksichtigt. Damit wird dem Umstand entgegengewirkt, dass Chi-Quadrat-Tests sensitiv sind für die Stichprobengröße: Je höher N ist, desto kleiner sind die absoluten Abweichungen von beobachteten und modellierten Daten, die als statistisch signifikant beurteilt werden. Infit und Outfit haben dieses Problem und identifizieren bei großen Probandenzahlen auch praktisch irrelevante Modellabweichungen. RMSD hingegen ist gegenüber der Stichprobengröße invariant.

Der Nachteil des RMSD wiederum ist, dass man hiermit nicht getrennt nach Über- oder Unterfit suchen kann. Es kann daher sinnvoll sein, mehrere Maße zu betrachten – sofern die Stichprobe unverzerrte Betrachtungen erlaubt.

Der Erwartungswert der RMSD bei perfektem Modellfit ist 0. Als Grenzen für guten und akzeptablen Fit gelten 0.5 und 0.8 (MacCallum, Browne & Sugawara, 1996).

6.2.3 Anwendung der Kriterien

Der Modelltest jedes einzelnen Modells wurde in einem festen Schema durchgeführt. Zuerst wurde folgende Schleife so lange durchlaufen, bis durch die verschiedenen Kriterien keine Änderungen des Datensatzes mehr vorgenommen wurden:

1. numerische Modellschätzung
2. Ausschluss aller Items, deren RMSD-Wert oberhalb von 0.8 liegt
3. Ausschluss aller Items, deren Infit eine größere Abweichung als $\frac{2}{\sqrt{N}}$ aufweist
4. Ausschluss aller Items, deren Outfit eine größere Abweichung als $\frac{6}{\sqrt{N}}$ aufweist
5. Ausschluss aller Items, die nicht mindestens 100 Mal bearbeitet wurden
6. Ausschluss aller Personen, die nicht mindestens 10 Aufgaben beantwortet haben
7. Neuanfang bei Punkt 1.

Danach findet ein Ausschluss von Items anhand ihrer empirischen und erwarteten itemcharakteristischen Kurve (ICC) statt. Alle Items mit Steigungen/Trennschärfen kleiner als 0 werden entfernt. Außerdem werden sämtliche Items entfernt, bei denen trotz angemessener Fitwerte Unterschiede zwischen beiden Kurven zu erkennen sind, die sich in Abstand oder Abweichung der Kurvenverläufe zeigten. Zu diesem Zweck werden die beobachtete mittlere Lösungswahrscheinlichkeit der Proband*innen in regelmäßigen Fähigkeitsintervallen berechnet. Daraus ergibt sich eine empirische ICC, die im Vergleich zu den Erwartungen aus dem Modell betrachtet werden kann. Mit dieser Art graphischer Überprüfung kann überprüft werden, ob Items in Fähigkeitsbereichen mit nur wenigen Bearbeitungen auffälliges Verhalten zeigen, dass sich aufgrund der kleinen Fallzahl in Fitindizes nicht zeigt (Hambleton, Swaminathan & Rogers, 1995). Im Abschnitt 6.4.1 wird ein Beispiel gegeben.

Der Vergleich der Modelle untereinander wird erst nach dieser individuellen Anpassung gezogen, die dazu verwendeten Kriterien werden nun beschrieben.

6.3 Methodik – Modellvergleich

Die in den vorigen Abschnitten 6.2.1 und 6.2.2 betrachteten Fitmaße können verwendet werden, um die Passung eines einzelnen Modells zu den Daten zu prüfen, die es beschreiben soll. Dazu wird die absolute Abweichung zwischen Beobachtung und Vorhersage berechnet, je nach Fitstatistik mit unterschiedlichen Berechnungsgrundlagen und Gewichten. Daher bezeichnet man diese Form von Überprüfung eines einzelnen Modells in sich auch als *absoluten Fit.*

Gibt es mehrere Modelle, die eine gute absolute Passung für den gleichen Datensatz aufweisen, helfen solche Maße bei der Auswahl nicht weiter. Werden zwei konkurrierende Modelle betrachtet, die unterschiedlich viele Freiheitsgrade (also geschätzte Parameter) besitzen, werden dementsprechend auch unterschiedliche absolute Abweichungen zu erwarten sein. In den konkret betrachteten Modellen wird das schnell anschaulich, da hier unterschiedliche Eigenschaften der Items modelliert werden. Die Güte von Modellen per Augenmaß vergleichen zu wollen, von denen eines nur Schwierigkeitsparameter und das andere auch Trennschärfen berücksichtigt, scheint aussichtslos.

Es müssen daher noch statistische Maße für den *relativen Fit* verwendet werden, mit denen Modelle untereinander vergleichen werden können.

6.3.1 AIC und BIC

Das *Akaike-Informationskriterium* (AIC), benannt nach Hirotsugu Akaike (Akaike, 1974), ist das älteste und eines der am häufigsten verwendeten relativen Fitmaße. Es schätzt den Informationsverlust eines Modells ein. Das geschieht im Vergleich zur echten Datenverteilung, also zum perfekten Modell, auf Grundlage der Kullback-Leibler-Divergenz ein – diese wird in der vor allem in der Informationstheorie verwendet und beschreibt den mittleren Informationsverlust pro Modellparameter (Kullback &

Leibler, 1951). Somit ist auch die Beschreibung der Modellgüte durch das AIC ein Mittelwert über alle Modellparameter. Die Gesamtzahl der Parameter wird in der errechneten Teststatistik nicht mehr berücksichtigt, aus diesem Grund können durch das AIC unterschiedlich komplexe Modelle untereinander verglichen werden.

Die generelle Definition des AIC ist (Burnham & Anderson, 2010):

$$AIC = 2 * k - 2 * \ln \hat{L}$$

wobei k die Anzahl der geschätzten Modellparameter und $\hat{L}$ das Maximum der Likelihoodfunktion (siehe 3.4.1) ist, also die Wahrscheinlichkeit der beobachteten Daten unter Annahme der verwendeten Modellparameter.

Für kleine Stichproben (was hier N < 40*k bedeutet) weist das AIC ein Bias für Modelle mit hoher Parameterzahl auf, wählt diese also bevorzugt aus. Für diesen Fall gibt es die Möglichkeit der Korrektur durch (Burnham & Anderson, 2010):

$$AICc = AIC + \frac{2k^2 + 2k}{n - k - 1}$$

Hierbei wird also als Ausgleich für den Bias ein von der Parameterzahl abhängiger Strafterm hinzugefügt, um einen overfit zu vermeiden.

Da das AIC den Informationsverlust des Modells beschreibt, sind kleinere Werte vorzuziehen.

Sehr ähnlich zum AIC kann das *Bayesianische Informationskriterium* (BIC) verwendet werden. Es ist definiert als (Burnham & Anderson, 2010):

$$BIC = 2 * \ln n - 2 \ln \hat{L}$$

Mathematisch unterscheiden sich beide Kriterien nur im ersten Term (Strafterm), der beim AIC nur von der Parameterzahl und beim BIC von der Stichprobengröße abhängt. In der Herleitung und Anwendung besteht der Unterschied in der Herangehensweise an die Problematik konkurrierender Modelle: Das AIC beurteilt alle Modelle im Vergleich zu einem angenommenen perfekten Modell, das BIC beurteilt sie im Vergleich

untereinander. Für eine detaillierte Beschreibung, Herleitung und Diskussion siehe Burnham und Anderson (2010) oder Vrieze (2012).

Ebenso wie das AIC stellt das BIC den Informationsverlust des Modells dar, weshalb im Vergleich Modelle mit geringeren Werten zu bevorzugen sind.

6.3.2 Testinformation und Reliabilität

Mit AIC und BIC kann untersucht werden, welches der Modelle den geringsten Informationsverlust pro Parameter erreicht. Andersherum ist aber auch interessant zu betrachten, wie viel Information insgesamt über die Stichprobe gewonnen werden konnte beziehungsweise wie genau die Fähigkeitsmessung mit den jeweiligen Modellen erfolgt ist. In den Abschnitten 3.3.1 bis 3.3.3 wurde bereits dargestellt, wie sich für eine gegebene Messung die gewonnene Information berechnen lässt. Dabei ist noch einmal festzuhalten, dass die bei einer gegebenen Antwort gewonnene Information von Person und Item abhängt, also für alle Personen individuell zu bestimmen ist. Um die Genauigkeit der Messung insgesamt zu beurteilen, kann aber einfach die im Verlauf des Tests gewonnene Information über alle Proband*innen gemittelt werden.

Hiermit könnte ein Vergleich zur Messgenauigkeit der Modelle vorgenommen werden, sehr anschaulich wäre er aber nicht – typischerweise wird hierfür die Reliabilität der Messung als Maß verwendet und daher auch als Maß der Testgüte erwartet. Sie kann definiert werden als (Bortz & Döring, 2006):

$$r = \frac{\sigma_T^2}{\sigma_X^2} = \frac{\sigma_X^2 - \sigma_E^2}{\sigma_X^2}$$

wobei $\sigma_T^2 = \sigma_X^2 - \sigma_E^2$ die wahre Varianz der Messgröße ist, σ_X^2 die beobachtete Varianz der Messwerte und σ_E^2 der Standardfehler der Messung. Letzterer kann nach Christensen et al. (2013) anhand der Information I bestimmt werden:

$$\sigma_E^2 = \frac{1}{I}$$

Damit ergibt sich für die Reliabilität die Formel:

$$r = \frac{\sigma_X^2 - \frac{1}{I}}{\sigma_X^2}$$
$$= 1 - \frac{1}{\sigma_X^2 * I}$$

Es kann also auf diese Weise ein Reliabilitätsmaß bestimmt werden, das die mittlere Genauigkeit der Messung anhand des jeweiligen IRT-Modells darstellt. Es hat ähnliche Eigenschaften wie Reliabilitäten in der klassischen Testtheorie und kann mit den gleichen Standards beurteilt werden.

Als Besonderheit ist abschließend festzustellen, dass für jede Messung mehrere unterschiedliche Werte bestimmt werden müssen. In Abschnitt 3.4 wurden als mögliche Schätzer für Personenfähigkeiten sowohl WLE-Werte als auch EAP-Werte beschrieben. Es können somit in jeder Messung pro Person zwei verschiedene Fähigkeitsschätzungen zustande kommen. Da die Information und somit auch die eben hergeleitete Reliabilität abhängig sind von diesem Schätzwert, müssen für jedes Modell sowohl die WLE-Reliabilität als auch die EAP-Reliabilität berichtet werden[14]. Je nach vorliegender Datenlage können sich beide Verfahren als das genauere herausstellen, der Vergleich zwischen den Modellen wird anhand des jeweils genauesten Schätzers vorgenommen.

6.4 Ergebnisse

In diesem Abschnitt werden zunächst die Ergebnisse der Fitprüfungen für die einzelnen Modelle dargestellt. Das geschieht gleichzeitig mit den Angaben zu ausgeschlossenen Items und Personen, da das einzige Ausschlusskriterium negative Auswirkungen auf die Modellpassung sind (siehe Abschnitt 6.2.3). Da Ergebnisdarstellung und Diskussion für die

[14] Die Information ist daneben auch abhängig von den Itemparametern. Für sie wird aber nur jeweils ein Schätzwert erstellt, weshalb sich keine komplexeren Kombinationen mit den Personenschätzern ergeben.

jeweiligen Modelle einen sehr geringen Umfang haben, werden beide Punkte zusammengefasst.

Der Vergleich der Modelle untereinander sowie die Auswahl des für die weitere Testentwicklung verwendeten Modells erfolgt danach in Abschnitt 6.4.4 anhand der dazu beschriebenen Gütemaße sowie der verbleibenden Datenmenge (Personen und Items, die nach Modellanpassung im Datensatz verbleiben).

Die Betrachtung des Itempools selbst erfolgt nicht in diesem, sondern im nächsten Abschnitt (6.5). Es handelt sich dabei zwar um eines der Resultate aus der Modellanpassung, allerdings hat der Itempool direkte Konsequenzen für das spätere Testinstrument und sollte daher in einem eigenen Bereich besprochen werden.

6.4.1 Modellfit 1pl

Für das einparametrisch-logistische Modell ergab sich nach dem geschilderten Ausschlussverfahren ein verbleibender Datensatz von 2325 Personen und 59 Items. Es wurden 51 Items aufgrund mangelhafter Fitwerte und 31 Items durch die anschließend manuelle Überprüfung der ICCs ausgeschlossen. Bei 1748 Personen blieben dadurch weniger als 10 registrierte Antworten im Datensatz übrig, was zum Ausschluss führte.

Abbildung 16 stellt eines der Items dar, die durch die Betrachtung der ICCs ausgeschlossen wurden. Es handelt sich dabei um Item BZ_Museum_01. Es weist 206 individuelle Bearbeitungen auf und hat die Überprüfung durch Infit, Outfit und RMSD mit den sich ergebenden Grenzen bestanden. Tatsächlich handelt es sich bei den berichteten Fitwerten des Items um hervorragende Ergebnisse (siehe Tabelle 2).

Tabelle 2: Fitwerte des manuell ausgeschlossenen Items BZ_Museum_01 im 1pl-Modell

N = 206	Akzeptanzbereich	Beobachtet
Infit	1 ± 0.41	1.009
Outfit	1 ± 0.13	1.002
RMSD	< 0.008	0.004

Abbildung 16 zeigt aber ein Problem im Antwortverhalten, dass durch die Fit-Statistiken nicht erfasst wurde. Im Fähigkeitsbereich zwischen -0.5 und 0.5 weicht die empirisch beobachtete Lösungswahrscheinlichkeit in einem irregulären Muster von der erwarteten ab. Bei kleinen Proband*innenzahlen könnte das erklärt werden durch statistische Ausreißer, die bei probabilistischem Testen zu erwarten sind. In den Fähigkeitsintervallen, aus denen die beiden mittleren und besonders kritischen Punkte des Graphs errechnet wurden, befinden sich aber jeweils ca. 70 Proband*innen. Das rein zufällige Auftreten der Ergebnisse im Rahmen des 1pl-Modells wird daher ausgeschlossen.

Eine mögliche Erklärung liegt in stark ausgeprägtem Rateverhalten. Es ist denkbar, dass für Proband*innen mit Fähigkeiten im Bereich der Itemschwierigkeit eine der inkorrekten Antworten ebenso plausibel erscheint wie die richtige Lösung, also als ein starker Distraktor funktioniert. Das kann zu bewusstem Rateverhalten oder von Person zu Person zufällig scheinenden Antworten führen, beides kann durch das 1pl-Modell nicht modelliert werden. Aufschluss hierüber könnte eine Distraktorenanalyse schaffen, bei der die Wahrscheinlichkeiten aller vier Antworten betrachtet werden. Im Rahmen dieser Arbeit ist aber eine Itemüberarbeitung nicht vorgesehen (vgl. Abschnitt 5.1), die Konsequenz ist deshalb in jedem Fall der Ausschluss des Items.

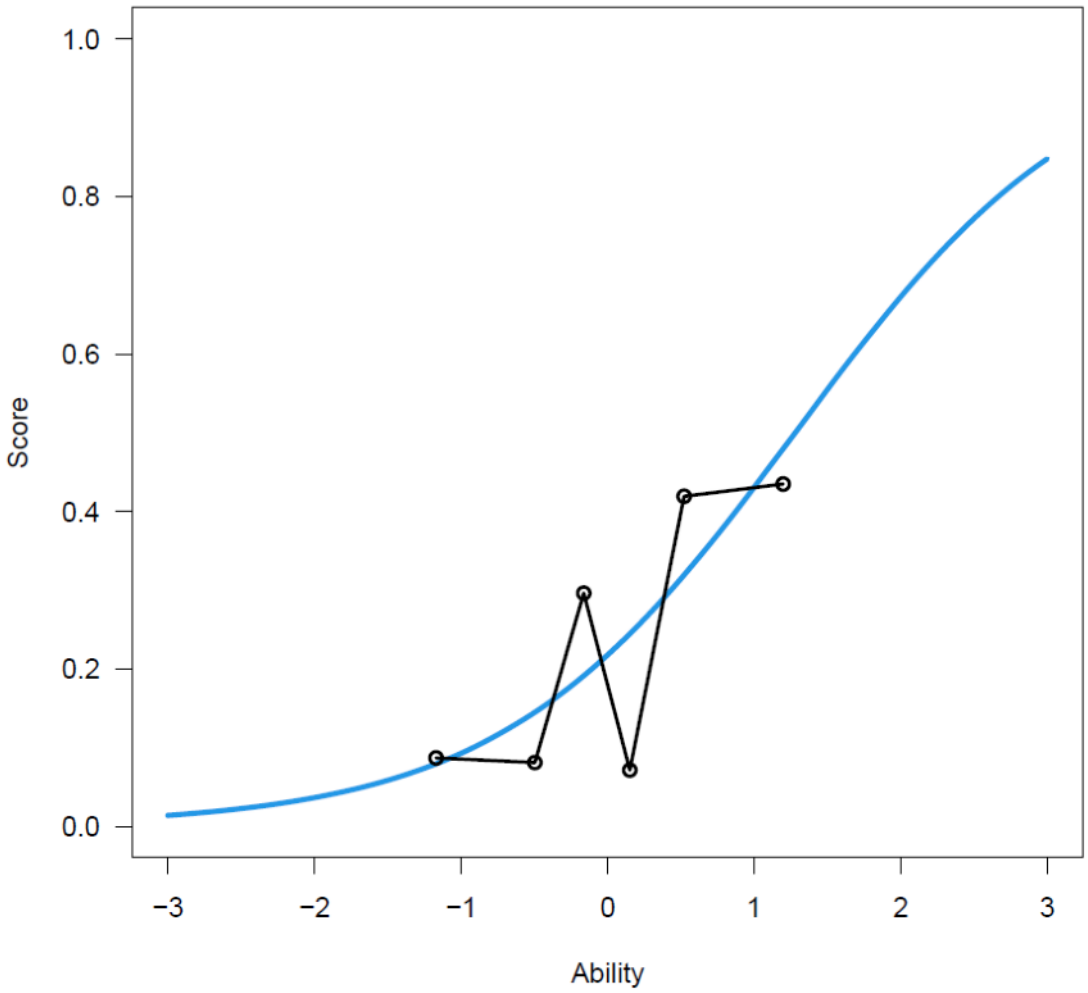

*Abbildung 16: Erwartete und beobachtete ICC des ausgeschlossenen Items BZ_Museum_01 im 1pl-Modell. Die Lösungswahrscheinlichkeit ist über die Fähigkeiten der Proband*innen aufgetragen. In blau ist die nach den Itemparametern erwartete Kurve zu sehen, in Schwarz die tatsächlich in verschiedenen Leistungsgruppen beobachtete Lösungswahrscheinlichkeit.*

Die Tabellen mit Infit, Outfit sowie RMSD Fitwerten der verbliebenen Items können im Anhang eingesehen werden.

Als Reliabilitäten der Personenschätzung ergaben sich Rel(WLE) = .423 und Rel(EAP) = .464.

6.4.2 Modellfit 2pl

Für das zweiparametrisch-logistische Modell ergab sich nach dem geschilderten Ausschlussverfahren ein verbleibender Datensatz von 2584 Personen und 62 Items. Es wurden 59 Items aufgrund mangelhafter Fitwerte und 19 Items durch die anschließend manuelle Überprüfung der ICCs

ausgeschlossen. Bei 1489 Personen blieben dadurch weniger als 10 registrierte Antworten im Datensatz übrig, was zum Ausschluss führte. Die Tabellen mit Infit, Outfit sowie RMSD Fitwerten der verbliebenen Items können im Anhang eingesehen werden.

Als Reliabilitäten der Personenschätzung ergaben sich Rel(WLE) = .501 und Rel(EAP) = .573.

6.4.3 Modellfit 3pl

Das dreiparametrische Modell konnte nicht erfolgreich geschätzt werden und wird daher aus den weiteren Betrachtungen ausgeschlossen.

Der verwendete Algorithmus zur MLE-Schätzung konvergierte nicht und brach nach der maximalen Anzahl an Iterationen ab. Als Lösungsansatz wurde zunächst die Zahl der erlaubten Iterationen erhöht, was aber auch bei einer Steigerung von 1000 auf 5000 keinen Erfolg brachte, da es während des Vorgangs immer wieder Phasen mit divergierenden Iterationen gab. Auch die Wahl von Startwerten der Rateparameter ergab keine Verbesserung (es wurde als Startwert 0.25 gewählt, da bei allen Items eine der vier Antwortmöglichkeiten korrekt war). Umgehen ließ sich das Problem, indem die Rateparameter praktisch aller Items auf 0 festgesetzt wurden, was allerdings einer Annäherung an das 2pl-Modell entspricht und damit keinen Mehrwert für weitere Modellvergleiche liefert.

Da für die numerische Bestimmung komplexerer Modelle auch mehr Datenpunkte benötigt werden besteht der Verdacht, dass die vorliegende Datenmenge nicht ausreichend für eine sichere Bestimmung aller Parameter ist. Ein Softwarefehler kann ebenfalls nicht ausgeschlossen werden, konkrete Anhaltspunkte für diesen Verdacht bestehen allerdings nicht.

6.4.4 Modellvergleich

Wie in Tabelle 3 zu sehen ist, war in beiden Fällen die Personenschätzung durch EAP-Werte reliabler als durch WLEs. Das zweiparametrische Modell konnte hier eine höhere Messgenauigkeit erreichen.

Tabelle 3: Reliabilitäten der angepassten IRT-Modelle im Vergleich. Der höchste Wert ist grau gekennzeichnet.

IRT-Modell	Rel(EAP)	Rel(WLE)
1pl	0.464	0.423
2pl	0.573	0.501

Tabelle 4: Relative Fitmaße der angepassten IRT-Modelle im Vergleich. Der jeweils von AIC und AICc zu verwendende Wert (basierende auf df und N) ist grau gekennzeichnet.

IRT-Modell	df	AIC	AICc	BIC
1pl	59	42564	42567	42904
2pl	124	43776	43788	44502

Im Vergleich der beiden Modelle anhand von AIC und BIC schnitt das einparametrische etwas besser ab. Der Unterschied zwischen beiden Modellen ist aber bei allen Maßen gering. Betrachtet man diese Werte in Kombination mit den Parameterzahlen und der erreichten Messgenauigkeit, also der pro Proband*in gewonnenen Information, zeigt sich insgesamt: Bei der Datenbeschreibung durch das 1pl-Modell kann mit jedem einzelnen Parameter mehr Information gewonnen werden als mit dem 2pl-Modell, es ist informationstheoretisch betrachtet effizienter. Der Vorsprung ist aber gering und spielt an dieser Stelle keine Rolle, da bei beiden Modellen nicht der Verdacht eines möglichen Overfits durch zu viele Parameter besteht. Das 2pl-Modell kann dafür durch die größere Menge an Parametern die Daten genauer beschreiben und insgesamt mehr Information gewinnen. Da beide Modelle in sich akzeptable Fitwerte aufwiesen, wird das 2pl-

Modell für die weitere Auswertung bevorzugt. Der verbliebene Itempool ist zudem beim 2pl-Modell mit 62 statt 59 verbleibenden Items leicht größer.

FF 1 kann somit als beantwortet betrachtet werden: Aus der Familie der logistischen Modelle wurde das 2pl-Modell als das Modell identifiziert, das die Ko-WADiS Daten am besten beschreibt. Das geschah unter Berücksichtigung der überprüften Fitmaße, der erreichbaren Messgenauigkeit sowie der Abdeckung des Kompetenzkonstrukts durch die in das Modell passenden Items.

6.5 Beschreibung des Itempools

Für die Konstruktion des adaptiven Tests verbleibt nach Modellanpassung und
-auswahl ein stark eingeschränkter Satz an Items, verglichen zur Ausgangslage nach der ersten Datenbereinigung (vgl. Abschnitt 6.1). Von der Menge und Verteilung der übrigen Items hängen zwei wichtige Punkte ab:

a) Die Zahl und Schwierigkeitsverteilung der Items gibt Obergrenzen für die Gesamtgröße des Tests vor, da je nach gewähltem Abschlusskriterium in allen anvisierten Fähigkeitsbereichen ausreichend Items verfügbar sein müssen.
b) Die inhaltliche Breite der verfügbaren Items und ihrer fachlichen Kontexte ist entscheidend für die Frage nach einer validen Testwertinterpretation.

Deshalb erscheint es sinnvoll, die Items in Bezug auf beide Aspekte zu betrachten und in einer Übersicht darzustellen.

Die Wrightmap (Abbildung 17) zeigt grundsätzlich eine erwünschte Verteilung von Fähigkeiten und Itemparametern. Die Proband*innen zeigen eine Normalverteilung um 0 Logits mit einer Standardabweichung von 0.75, es befinden sich also rund 95% der Stichprobe in einem Fähigkeitsbereich zwischen 1.5 und -1.5. In diesem Bereich können wir auch praktisch alle verbliebenen Items finden, die Passung zwischen der Gesamtverteilung der Stichprobe und dem Itempool ist also insgesamt gut.

Im unteren Randbereich der Verteilung, also für Fähigkeitswerte unter -1.5, sind jedoch kaum mehr geeignete Items vorzufinden. Ein ähnliches Bild zeigt sich im oberen Randbereich, auch wenn hier etwas mehr Items zur Verfügung stehen. Es wird also in den Extrembereichen der Personenverteilung nicht möglich sein, eine genaue Messung durchzuführen – unabhängig vom gewählten Testverfahren. Welche weiteren Auswirkungen dies auf die Möglichkeit der Testgestaltung hat, wird im folgenden Kapitel diskutiert.

Die inhaltliche Breite des Testinstruments ist auch nach dem drastischen Ausschluss von Items erhalten geblieben. Den Idealfall würde eine

Gleichverteilung der Items jeweils auf Kompetenzfacetten und Fachbereiche darstellen. In Bezug auf die Facetten wurde es fast erreicht, im Bereich der Modelltests gibt es ein leichtes Überangebot im Vergleich zum Rest des Instruments.

Die Aufspaltung der Iteminhalte auf die drei Fachbereiche ist weniger gut gelungen; es wurden überproportional viele Items aus dem Bereich Chemie ausgeschlossen. Auch hier wird im nächsten Kapitel diskutiert, ob sich Konsequenzen für die geplante Testgestaltung ergeben.

Die Übersicht über Itemparameter, Fitwerte und ICCs findet sich, wie in Abschnitt Modellfit 2pl6.4.2 erwähnt, im Anhang.

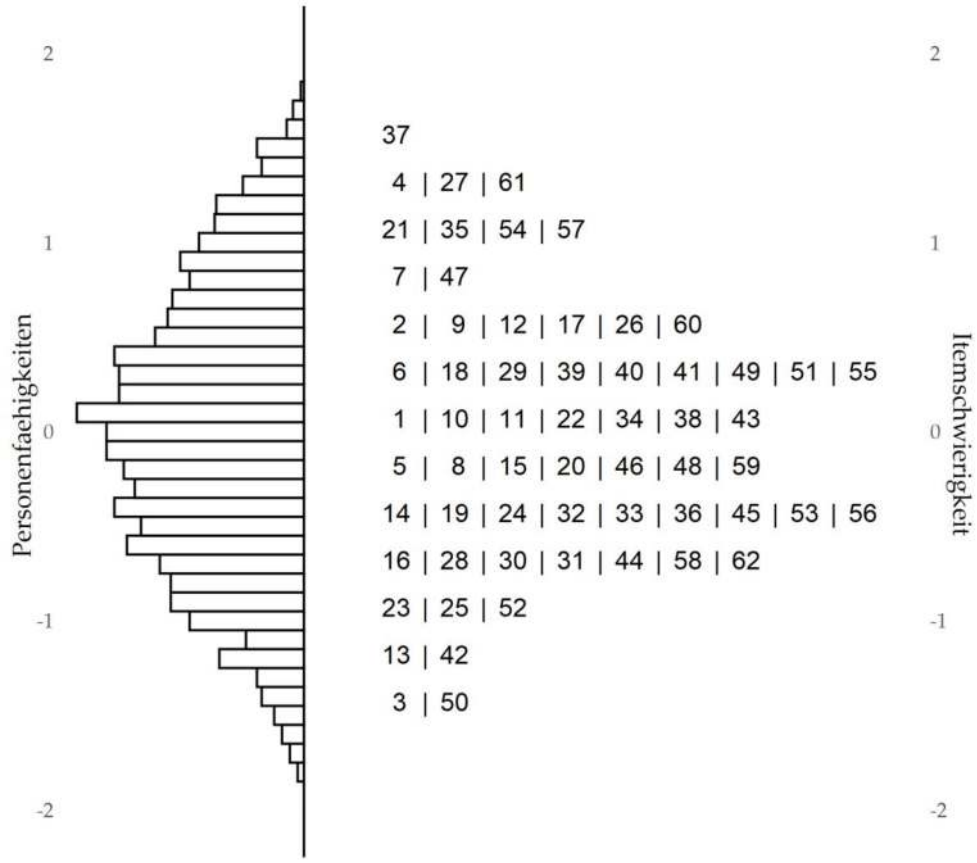

Abbildung 17: Wrightmap des 2pl-Modells. In dieser Art von Grafik werden die Verteilung der Personenfähigkeiten in der Stichprobe (links) der Verteilung der Itemschwierigkeiten im Testinstrument (rechts) gegenübergestellt. Die Items werden ihrer alphabetischen Nummerierung nach mit einer Ziffer dargestellt. Die

gemeinsame Skala ist auf beiden Seiten markiert. Somit entsteht ein visueller Überblick nicht nur über beide Verteilungen in sich, sondern auch über die Abdeckung der zu messenden Fähigkeiten durch die verfügbaren Items. Für die Grafik wurden die oberen und unteren 0,5% der noch verbliebenen Stichprobe als Ausreißer entfernt, um eine bessere Auflösung zu ermöglichen.

Tabelle 5: Häufigkeit der verschiedenen Kompetenzfacetten im Itempool.

Fragen	**Hypothesen**	**Planung**	**Auswertung**
7	8	11	9

Modellzweck	**Modelltest**	**Modelländerung**
8	12	7

Tabelle 6: Häufigkeit der verschiedenen Fachbereiche im Itempool.

Biologie	**Chemie**	**Physik**
24	13	25

7 Testkonstruktion (FF 2)

In diesem Kapitel werden die Forschungsfragen 2.1 bis 2.3 behandelt. Als erstes wird anhand theoretischer Überlegungen entschieden, welches Testverfahren (CAT oder MST) für die Konstruktion des neuen Testinstruments Verwendung findet, also FF 2.1 geklärt (Abschnitt 7.1). Dazu werden, wie in Kapitel 5 beschrieben, Einschränkungen durch die Modellauswahl und den Itempool berücksichtigt, die Einsatzziele des Instruments sowie praktische Aspekte der Testgestaltung.

Der Auswahl entsprechend wird danach in den Abschnitten 7.2 und 7.3 die Testkonstruktion beschrieben. Hier bedeutet das, dass verschiedene Entscheidungen in Bezug auf den Umfang des Tests und den Testalgorithmus getroffen werden müssen (vgl. Abschnitte 4.2 und 4.3, je nach dem gewählten Testverfahren).

Die praktische Umsetzung und Implementierung des Instruments werden nicht in diesem Kapitel erfolgen, sondern direkt im Zusammenhang mit der späteren Pilotierung in Kapitel 8.

7.1 Auswahl des Testverfahrens

Grundsätzlich erlaubt der vorhandene Pool an Items die Konstruktion sowohl eines CATs als auch eines MSTs. Es stellt sich also nach der Modellauswahl weiterhin die Frage, welches der beiden Verfahren zu wählen ist.

Zunächst ist relevant, wozu das Instrument in Zukunft eingesetzt werden kann und soll. Bisher wurde der Test nur für Gruppenvergleiche herangezogen, es gibt allerdings keinen eindeutig festgelegten Einsatzbereich für den Zeitraum nach den laufenden Forschungsvorhaben. Es erscheint also sinnvoll, alle Möglichkeiten in Betracht zu ziehen. Dabei sollte vor allem unterschieden werden zwischen der Messung und Beschreibung von Gruppen und der Diagnose einzelner Personen.

Die Beschreibung von Gruppen erfordert eine genaue Ermittlung des mittleren Fähigkeitsniveaus und der Streuparameter der Stichprobe. Die Fähigkeiten der einzelnen Individuen sind hierbei nicht mehr von Relevanz,

sobald diese Werte bestimmt wurden. Auch ist es nicht wichtig, eine genaue Messung in den Extrembereichen des latenten Merkmals zu erreichen – kleine Anteile der Stichprobe, die extreme Merkmalsausprägungen haben, haben nur einen geringen Einfluss auf die Kennwerte der Gesamtverteilung. Es ist also für ein Messinstrument besonders wichtig, die Mitte der Verteilung genau zu erfassen.

Für die Individualdiagnostik muss das Messprofil über den gesamten Fähigkeitsbereich verteilt sein, damit auch Grenzfälle einzeln untersucht werden können. Zudem lässt sich argumentieren, dass die Anforderungen an die Messgenauigkeit insgesamt höher sind, sofern einzelne Personen betrachtet werden. Bei geringer Reliabilität steigt die Wahrscheinlichkeit einer falschen Einschätzung von Individuen sehr schnell an, damit auch die mögliche falsche Benotung oder Förderung/Intervention im Anschluss an den Test. Die Kollektivdiagnostik ist hier robuster, es sind lediglich mehr Daten für die Auflösung von Gruppenunterschieden notwendig (Moosbrugger & Kelava, 2012).

Betrachtet man die bisher erreichten Messgenauigkeiten sowie die Verteilung der Itemschwierigkeiten und Personenfähigkeiten in Abbildung 17, lässt sich den vorigen Abschnitten nach eindeutig folgern: Der Itempool ist zwar für die Beobachtung von Gruppen gut geeignet, da er den mittleren Fähigkeitsbereich der Population gut abdeckt, weist aber erhebliche Schwächen bei der Messung in Randbereichen auf. Zusammen mit der allgemein niedrigen Messgenauigkeit kann daher die Individualdiagostik als Einsatzzweck verworfen werden. Das schließt noch keines der beiden möglichen Verfahren aus, hat aber Konsequenzen für die Ausgestaltung des Tests. Besprochen wurde dies bereits bei möglichen Abschlusskriterien von CATs (siehe Abschnitt 4.2.3).

Neben dem Einsatzzweck können auch praktische Aspekte der Testgestaltung zu einer Entscheidung zwischen den beiden Verfahren führen. Hiervon werden im Folgenden zwei betrachtet.

Wie in den theoretischen Ausführungen erklärt, werden bei einem CAT mit einfachem Algorithmus die Items nach rein psychometrischen

Kriterien, meist dem anhand des IRT-Modells erwarteten Informationsgewinn, ausgesucht (vgl. Abschnitt 4.2.1). Im verwendeten Modell werden die unterschiedlichen Fachrichtungen und Kompetenzfacetten der Items nicht berücksichtigt. Eine der Folgen kann daher sein, dass Proband A kein einziges Item aus dem Bereich der Modelltests vorgelegt bekommt, während Proband B nur anhand von zwei der drei möglichen Fächer geprüft wird. Aus informationstechnischer Sicht ist das kein Problem, da keine Mehrdimensionalität des latenten Merkmals vorliegt.

Es besteht aber die Gefahr von Motivationseffekten, die auf diese Weise verursacht werden könnten. Während der Entwicklung der Testaufgaben wurden Proband*innen mit gemischten Testheften befragt, die Items aus allen drei Fächern in gleichem Anteil enthielten, und mit fachspezifischen Testheften, die nur Aufgaben des gewählten Studienfachs der jeweiligen Personen enthielten. Hierbei zeigte sich keine Aufspaltung der Kompetenz in den Fachbereichen, weshalb sie eindimensional modelliert wird (Straube, 2016).

Es wurde leider nie geprüft, was passiert, wenn die Proband*innen mit gezielt fachfremden Aufgaben konfrontiert werden. Aufgrund vereinzelter Gespräche mit Testteilnehmer*innen gibt es die Befürchtung, so eine Messung könnte zu einer verringerten Teilnahmebereitschaft und Motivation führen, was die gezeigte Leistung in der Messung beeinflussen würde.

Im reduzierten Itempool ist das Fach Chemie deutlich unterrepräsentiert. Die Wahrscheinlichkeit des obigen Szenarios bei der Messung von Chemiestudierenden erscheint damit sehr hoch. Es ist abzuwägen, ob die grundsätzliche Gefahr von Motivationsverlusten akzeptiert werden kann, auch wenn solche Effekte bisher nicht gezielt beobachtet oder empirisch nachgewiesen werden konnten. Eine Literatursuche ergab leider keine allgemeingültigen Ergebnisse zur Frage, ob die Befürchtung realistisch ist.

Denkbar wären zusätzliche Zwangsbedingungen für den Auswahlalgorithmus in einem CAT, durch die eine breite Verteilung der gewählten Itemkontexte erreicht wird. Das Verfahren wird dadurch allerdings sehr schnell komplizierter. Zudem ist schwer abzuschätzen, wie groß die

Auswirkungen dieser Bedingungen auf die Messgenauigkeit wären. Bei MSTs hingegen kann eine ausgewogene Repräsentation aller inhaltlichen Aspekte des Tests bei der Konstruktion der Module berücksichtigt werden, weshalb MSTs an dieser Stelle vorteilhaft erscheinen.

Ein Argument für die Wahl eines CATs könnte noch die Effizienzsteigerung sein, die, wie in Kapitel 5 besprochen wurde, das Ziel der gesamten Arbeit ist. In Abschnitt 4.4.1 wurden die beiden Formate untereinander verglichen, die herangezogenen Studien verweisen meist auf CATs als die effizientere Möglichkeit. Allerdings ist der Unterschied bei optimaler Struktur eines MSTs und unter Verwendung der *forward assembly* bei der Modulerstellung nicht mehr signifikant, was das Argument stark entkräftet. Zudem ist nicht klar, ob der eventuelle Vorteil des CATs in einer noch höheren Messgenauigkeit überhaupt notwendig ist: Ausgehend von der mittleren Effizienzsteigerung in den zitierten Studien erscheint es realistisch, die Messgenauigkeit von .573 (Abschnitt 6.4.4) durch einen MST auf einen Wert von .7 oder höher zu treiben. Damit würde den Ansprüchen an die Gruppendiagnostik in einschlägiger Literatur genüge getan. Eine weitere Erhöhung wäre wünschenswert, muss aber gegen die praktischen und inhaltlichen Bedenken verrechnet werden.

Unter Einbeziehung der genannten Punkte wird **FF 2.1** daher wie folgt beantwortet: Ein MST erscheint der Zielstellung vollständig angemessen und in Konstruktionsaspekten praktischer beziehungsweise unbedenklicher als die Erstellung eines CATs. Für die Optimierung des Ko-WADiS-Tests wird daher ein MST als die geeignetere Wahl betrachtet.

Damit entfällt **FF 2.2** für die weitere Bearbeitung. In den kommenden Abschnitten ist allerdings noch **FF 2.3** zu beantworten. Dazu gehören folgende Aspekte:

1. Es muss entschieden werden, welche Struktur für den MST verwendet werden soll, also die Anzahl der Stufen und die Aufspaltungen in Schwierigkeitsbereiche innerhalb der Stufen.
2. Für die einzelnen Module müssen Konstruktionsvorschriften aufgestellt werden.
3. Scoring- und Routingregeln müssen definiert werden.

7.2 Regeln für Modul- und Strukturaufbau

Auf den ersten Blick würde es sinnvoll erscheinen, sich zuerst für eine Teststruktur zu entscheiden und diese danach mit passenden Modulen zu füllen. Das macht besonders dann Sinn, wenn inhaltliche Eigenschaften des Testinstruments eine bestimmte Teststruktur vorgeben. Soll eine Kompetenz mit unterschiedlich modellierten Kompetenzniveaus gemessen werden, liegt beispielsweise die Aufspaltung der letzten Teststufe in jeweils entsprechende Schwierigkeitsniveaus nahe.

Eine solche Vorgabe gibt es für den Ko-WADiS-Test nicht, bei der Testkonstruktion geht es einzig um die maximal erreichbare Effizienz des Instruments. Die Literatur gibt leider keine eindeutige Lösung dafür vor, weshalb verschiedene Strukturen erstellt, getestet und verglichen werden müssen.

Damit der Vergleich und die Auswahl der endgültigen Struktur nicht durch die Qualität der einzelnen Module verfälscht wird, sollen und können im Voraus gemeinsame Regeln für die Konstruktion der Module aufgestellt werden, die für alle Strukturen in gleicher Weise gelten. Sie werden nun der Reihe nach dargestellt und begründet.

Die erste Stufe jedes Tests enthält nur ein Routing-Modul. Der Grund hierfür ist der Mangel an Kovariaten, die eine grobe Einschätzung und Einteilung der Proband*innen in verschiedene Fähigkeitsbereiche erlauben würden.

Das Routing-Modul deckt mit den enthaltenen Items einen Schwierigkeitsbereich ab, der dem erwarteten Fähigkeitsbereich von Mittelwert ± 0.5 Standardabweichungen der Stichprobe entspricht. Die Vorhersage über die Fähigkeitsverteilung erfolgt anhand der bisher beobachteten Personendaten, die bereits im vorigen Kapitel verwendet wurden. Für diese Festlegung gibt es zwei Gründe. Es wurde zunächst entschieden, das Routing-Modul nicht über das gesamte Fähigkeitsspektrum zu strecken, sondern nur über das mittlere Drittel der Verteilung. Durch eine breite Streckung der Itemschwierigkeiten könnte erreicht werden, dass nach dem Routing-Modul über alle Proband*innen gleich viel Information vorliegt und die Weiterleitung aller Personen aus dieser Hinsicht gleich verlässlich ist. Die Informationsmenge wäre aber in diesem Fall sehr gering, die Weiterleitung also für die gesamte Stichprobe fehleranfällig.

Das gewählte Vorgehen sorgt dafür, dass für das zentrale Drittel der Stichprobe eine genauere Schätzung vorliegt als für die Personen im oberen und unteren Bereich der Verteilung. Um letztere weiterleiten zu können, ist allerdings keine exakte Einschätzung notwendig: Bearbeiten in Realität leistungsstarke/schwache Teilnehmer*innen Items mit mittleren Schwierigkeiten, werden sie die meisten davon richtig/falsch beantworten. Die geringe Menge an gewonnener Information führt dann zu einer Über/Unterschätzung der realen Kompetenzausprägung. Wer also die meisten Items des Routing-Moduls korrekt löst, wird in seiner Kompetenz zwar überschätzt, aber folgerichtig in das schwerere Modul der nächsten Stufe geleitet (und für leistungsschwache Personen umgekehrt). So eine Weiterleitung funktioniert bei mittleren Fähigkeitsniveaus nicht, aber diese werden durch das Vorgehen exakter eingeschätzt und können so genauer verortet werden, als es bei einem breit gesteckten Routing-Modul möglich wäre.

Ein Problem würde entstehen, wenn die Stufe nach dem ersten Routing in fünf oder mehr Niveaus aufgeteilt würde. Es könnte dann durch die Über/Unterschätzung passieren, dass Personen anstelle einer korrekten Einstufung in das zweit/viertschwerste Modul in das erst/fünftschwerste geschickt würden. So hohe Aufspaltungen sind aber nicht vorgesehen (siehe dazu das Ende dieses Abschnitts).

Der zweite Bestandteil der Regel ist die Verwendung von Standardabweichungen zur Festlegung der Grenzen des Moduls. Der Prozentrang, also der prozentuale Anteil von Personen in einer Verteilung, ausgedrückt durch deren Abstand vom Mittelwert in Standardabweichungen, ist nicht abhängig von der absoluten Streuung in der Verteilung. Mit ihm kann eine Verteilung in gleich große Intervalle eingeteilt werden, egal wie stark die Fähigkeiten in der Stichprobe wirklich streuen. Deshalb ist er am besten geeignet, um einen festen Anteil der Stichprobe anzuvisieren, im Gegensatz zur Verwendung der mittleren Itemschwierigkeiten aus der dort verfügbaren Verteilung. Im Idealfall würden sich Schwierigkeiten und Fähigkeiten natürlich decken und es gäbe keinen Unterschied zwischen den Verfahren, davon kann aber nicht immer ausgegangen werden.

In den restlichen Stufen werden die Module so erstellt, dass sie möglichst lückenlos den Fähigkeitsbereich von zwei Standardabweichungen um den Mittelwert der Verteilung abdecken. Dass die Stufen lückenlos gefüllt werden sollen, ist wohl nicht weiter erklärungsbedürftig. Die äußeren Grenzen der Stufen wurden auf zwei Standardabweichungen beschränkt, um Extremfälle und Ausreißer der Verteilung bei der Testkonstruktion zu ignorieren. Es werden so 98% der Population anvisiert.

Die Aufspaltung innerhalb der Stufen erfolgt nach Prozenträngen. Das bedeutet, es werden nach Möglichkeit gleich viele Personen in alle Module einer Stufe einsortiert, anstelle einer Einteilung in gleich breite Fähigkeitsintervalle. Zum einen begründet sich diese Regel ebenso wie die Eingrenzung des Routing-Moduls mit einer funktionalen Betrachtung des Routings nach dem Modul. Zum anderen soll so erreicht werden, dass alle Items des Testinstruments gleich oft beantwortet werden. Sollten zu einem späteren Zeitpunkt noch einmal Itemanalysen auf Basis von Daten des MST vorgenommen werden, kann so eine mangelnde Datenlage bei einzelnen Items ausgeschlossen werden. Daneben hat es auch einen Vorteil für die Testpraxis: Sollte eine Stichprobe über einen längeren Zeitraum beobachtet, also mehrfach mit dem Instrument gemessen werden, wird die Problematik der Itemexposition durch das Verfahren verringert.

Das gewählte Vorgehen kann dann als kritisch betrachtet werden, wenn durch eine feine Aufspaltung die Breite der anvisierten Fähigkeitsbereiche kleiner ist als die Fehlermarge bei der Personenschätzung. Ob dieses Problem besteht, wird sich in der Erprobung zeigen, es wird aber zu diesem Zeitpunkt nicht davon ausgegangen.

Alle Module einer Struktur haben die gleiche Länge von Testlänge in Items/Stufenzahl. Wie in den theoretischen Ausführungen schon erwähnt, kann im Allgemeinen kein struktureller Vorteil von variierenden Modullängen festgestellt werden. Die einzige mögliche Ausnahme ist das Routing-Modul, das manchmal für eine genauere Ersteinschätzung im Vergleich zu den anderen Modulen eines Tests verlängert wird. Das ist vor allem relevant bei der Individualdiagnostik, da hier nach einem Routing-Modul alle Individuen für einen meist folgenden CAT exakt zugeteilt werden sollen. Wie vorher besprochen ist das Ziel des Routing-Moduls aber bei diesem Test keine direkt exakte Einteilung der Proband*innen, sondern lediglich die Zuteilung zum am besten geeigneten Modul. Zudem ist die Gesamtlänge des Testinstruments durch den verfügbaren Itempool und die angesprochene Kritik in Bezug auf Probandenbelastung stark beschränkt. Ein langes Routing-Modul mit mittlerer Schwierigkeit könnte so bei Proband*innen mit niedriger/hoher Fähigkeit zu einer anfänglichen Über/Unterforderung führen, was die Motivation beeinflussen würde. Außerdem wären dann durch Vorgaben der Gesamttestlänge die anschließend folgenden Module mit passender Schwierigkeit kürzer, die Gesamtmessung also ungenauer. Es wurde deswegen entschieden, das Routing-Modul nicht zu verlängern, auch wenn solch ein Vorgehen bei vielen anderen adaptiven Tests beobachtet werden kann.

Mit diesen Regeln ist festgelegt, wie abhängig von der Struktur die Längen und Schwierigkeitsintervalle aller Module ausfallen. Es fehlen nur noch Vorschriften dafür, wie diese Module mit Items aus dem verfügbaren Pool zu füllen sind.

Die Auswahl der Items folgt dem Prinzip der forward assembly. Grund ist der Vorteil des Vorgehens bezogen auf die Effizienz des Testinstruments (siehe Abschnitt 4.4.1 oder Wang, 2017). Es werden also die Module der

Reihenfolge im Testverlauf nach mit den informativsten Items zuerst aufgefüllt, sodass die ersten Module die „besten“ Items enthalten.

Zu diesem Zweck wird, nachdem die Schwierigkeitsintervalle der Module festgelegt wurden, der mögliche Informationsgewinn jedes Items innerhalb der jeweiligen Grenzen über das Intervall des Moduls integriert. Das Integral drückt dann den Informationsgewinn für das gesamte Modul aus und ist so ein Gütemaß für das einzelne Item. Von den besten Items werden der Modullänge entsprechend viele aus dem Pool gezogen und in das Modul eingesetzt. Hierbei muss noch beachtet werden, dass die Schwierigkeit dieser Items über das gesamte Zielintervall des Moduls gestreut ist, und nicht beispielsweise nur Items am oberen Rand des Moduls ausgewählt werden. Da trennscharfe Items einen engen Bereich mit hohem Informationsgewinn aufweisen, könnten sonst innerhalb des Moduls Bereiche mit geringem Gesamtinformation entstehen.

Diesem Auswahlverfahren wird aber noch eine Beschränkung auferlegt: *Solange nicht bereits alle Facetten und Fachbereiche im Modul repräsentiert sind, werden nicht mehr als zwei Items mit der gleichen Facette und dem gleichen Fachkontext in ein Modul eingefügt.* Durch diese Regel soll den inhaltlichen Bedenken Folge getragen werden, die bei der Testauswahl als Argument gegen CATs und für MSTs verwendet wurden. Über eine noch strengere Beschränkung wurde nachgedacht, wie beispielsweise die Auswahl aller drei Fächer in den ersten drei Items, die in das Modul eingefügt würden. Der Versuch, diese Idee praktisch umzusetzen, scheiterte aber an den verfügbaren Items. Grund hierfür ist schlicht der Umfang des Itempools (vgl. Abschnitt 6.5) und damit die Freiheit bei der Auswahl psychometrisch guter Items in den verschiedenen Schwierigkeitsintervallen (besonders bei leichten Items).

Es stehen nun alle notwenigen Regeln für den Aufbau der verschiedenen Teststrukturen fest. Zwei Anmerkungen sollen aber noch gemacht werden.

Erstens erlaubt der Umfang des Itempools nicht den Aufbau verschiedener Panels pro Struktur, dafür ist er zu klein. Für den prinzipiellen Vergleich der Strukturen untereinander ist das nicht von Relevanz, es spielt nur eine

Rolle bei möglichen Erinnerungseffekten im denkbaren Einsatz des fertigen Tests bei Langzeitstudien.

Untersuchungen im Projekt ValiDiS deuten an, dass es sich bei dem gemessenen Konstrukt um eine Kompetenz handelt, die sich kurzzeitig nur gering und auch nur durch eine sehr intensive Intervention messbar steigern lässt. Für den zukünftigen Testeinsatz folgen daraus eher Testabstände im Rahmen von ganzen Semestern als im Wochenabstand. Werden die Testergebnisse nicht an den Erfolg in einer Lehrveranstaltung oder ähnliche Erfolgskriterien gebunden, also keine speziellen Anreize für das Einprägen der Items gegeben, sollten Erinnerungseffekte damit klein ausfallen.

Zweitens gibt der Itempool starke Einschränkungen in der möglichen Testlänge vor: So sind im Bereich von -1 bis -2 Logits überhaupt nur 12 Items im Pool vorhanden. Wird eine Struktur erstellt, bei der dieser Bereich ein Modul-Intervall darstellt, ist diese Itemzahl (plus Items aus dem Routing-Modul) eine natürliche Grenze der Testlänge. In der Praxis werden die Konsequenzen vermutlich nicht groß sein. Da einer der möglichen Kritikpunkte am Testinstrument die scheinbar zu große Belastung mit 21 Items im linearen Format ist, sollte dieser Wert sowieso nach Möglichkeit unterschritten werden.

7.3 Regeln für Scoring und Routing

Die Regeln zur Weiterleitung in Module der nächsten Stufe werden zum größten Teil durch das gewählte 2pl-Modell vorgegeben (vgl. Abschnitt 4.3.3). Wie im vorigen Abschnitt 7.2 diskutiert, werden alle inhaltlichen Anforderungen durch die Konstruktion der einzelnen Module erfüllt. Es ist daher nicht notwendig, weitere inhaltliche Zwangsbedingungen für das Routing vorzugeben, die Auswahl der Module erfolgt allein anhand der Informationsmaximierung.

Zunächst muss nach Abschluss jeder Stufe eine Einschätzung der Personen erfolgen. Bei der Auswahl und Anpassung des geeigneten IRT-Modells erwiesen sich EAP-Werte als die genaueren Schätzwerte für Personenfähigkeiten, daher werden sie als die Fähigkeitswerte für alle Berechnungen

verwendet. An jedem Routing-Punkt wird eine aktuelle EAP-Einschätzung auf Grundlage aller bisher im Test gegebener Antworten vorgenommen.

Hierbei ist noch wichtig zu erwähnen, wie nicht bearbeitete Items beurteilt werden. Es gibt zwei Optionen: Erstens können unbearbeitete Items aus der Berechnung ausgeschlossen werden, da technisch gesehen keine falsche Antwort gegeben wurde und somit nicht sicher ist, ob die Person das Item wirklich falsch gelöst hätte oder aus einem anderen Grund keine Antwort gegeben hat. Dieses Vorgehen vermeidet eine Fehleinschätzung der einzelnen (fehlenden) Antwort. Es bedeutet aber auch einen Informationsverlust für die Messung, da somit die Itembearbeitung vollständig verloren geht. Die Alternative ist, alle nicht gegebenen Antworten als falsche Antworten einzustufen. Auf diese Weise geht technisch gesehen keine Information verloren, da alle Items einfließen, allerdings könnte der Grund für eine ausgelassene Antwort auch ein anderer sein als mangelnde Fähigkeit. So ist es denkbar, dass unter Zeitdruck zunächst schwer erscheinende Items übersprungen werden, um sie vielleicht später zu lösen. Reicht dann die Gesamtzeit nicht aus, wurden möglicherweise für die Person lösbare Items ausgelassen. Da es im Ko-WADiS-Test in linearer Form ebenso wie im MST kein Zeitlimit gibt und die Korrektur übersprungener Items möglich ist[15], werden nicht beantwortete Items mit falschen Lösungen berechnet, anstatt sie als fehlende Information zu verwerfen.

Nach der Fähigkeitsschätzung muss die erwartete Information des kommenden Moduls bestimmt werden. Sie wird im Test ebenso wie in den bisherigen theoretischen Ausführungen durch die *Fisher- Information* bestimmt:

$$I(\theta) = a^2 PQ$$
$$= a^2 \frac{e^{a(\theta-b)}}{(1+e^{a(\theta-b)})^2}$$

[15] Mit Einschränkungen im MST, siehe Abschnitt 4.3.

An dieser Stelle sei noch einmal erwähnt, dass die Information mehrerer Items addiert werden kann, die erwartete Information eines Moduls ist also die Summe der erwarteten Informationen seiner Items.

Somit steht fest, wie nach der Bearbeitung eines Moduls alle Personen und der Informationsgewinn in den noch zur Verfügung stehenden Modulen eingeschätzt werden können. Die Weiterleitung erfolgt immer zu dem Modul mit dem höchsten erwarteten Zuwachs an Information.

Als einzige weitere Beschränkung wurde entschieden, nur Sprünge in benachbarte Module zu erlauben, wie es bereits in Abschnitt 4.3.3 diskutiert wurde. Der Grund hierfür ist die geringe Menge an Items in jedem einzelnen Modul. Da die gewonnene Information nach den ersten Modulen in allen denkbaren Strukturen noch gering sein wird, besteht die Gefahr einer Falscheinschätzung und Übersteuerung. Die Folge wäre ein weiterer Informationsverlust in den folgenden Modulen, was so verhindert werden soll.

Zusammengefasst erfolgt das Routing nach Abschluss eines Moduls wie folgt:

1. Alle nicht korrekt gelösten Items werden als falsch bewertet, auch ausgelassene.
2. Die Fähigkeit der Person wird durch EAP mit allen bisher im Test gegebenen Antworten eingeschätzt.
3. Der erwartete Informationsgewinn aller zur Verfügung stehenden Module wird anhand des EAP-Werts, der normierten Itemkennwerte und der Fisher-Information für das 2pl-Modul berechnet.
4. Die Weiterleitung erfolgt in das Modul mit dem größten erwarteten Informationsgewinn. Hierbei sind nur Sprünge zwischen Modulen mit überlappenden oder angrenzenden Schwierigkeitsbereichen erlaubt.

7.4 Auswahl der Teststruktur durch Simulationsstudien

In den vorigen Abschnitten 7.2 und 7.3 wurde geklärt, wie die einzelnen Bestandteile des MSTs konstruiert und das Routing gesteuert werden. Was fehlt, ist die Auswahl der endgültigen Struktur, die nach diesen Regeln zusammengesetzt und angewendet wird. Diese ist allein danach

auszuwählen, ob sie im Vergleich zu Alternativen die effizienteste Umsetzung des Ko-WADiS-Tests darstellt.

Da die Literaturrecherche keine eindeutige Lösung liefert, müssen die denkbaren Alternativen jeweils erstellt und verglichen werden. Aus den betrachteten Metastudien in Abschnitt 4.3.1 bis 4.4.1 ergeben sich ein paar Richtlinien, die sich als verlässlich erwiesen haben und die Auswahl einschränken. Eine Aufspaltung in a) mehr als drei Schwierigkeitsbereiche und/oder b) mehr als vier aufeinanderfolgende Stufen (Routing-Stufe/Modul eingeschlossen) hat vermutlich keine relevanten Auswirkungen mehr auf die Effizienz. Das deckt sich auch mit der Problematik, dass bei einer feineren Aufspaltung in einer der beiden Dimensionen der vorhandene Itempool zu dünn gestreckt würde.

Es wurde entsprechend entschieden, die Strukturen 1-2 bis 1-2-2-2 sowie 1-3 bis 1-3-3-3 zu testen. Für alle Strukturen werden verschieden lange Testversionen gebaut. Da entschieden wurde, die Module alle gleich lang zu gestalten, ist die Testlänge jeder einzelnen Version immer ein Vielfaches der Stufenzahl. Die maximale Länge der Tests wird praktisch durch den Itempool vorgegeben, der besonders bei einer Aufspaltung in drei Schwierigkeitsniveaus nur noch eine begrenzte Menge an Items im leichten Fähigkeitsbereich bietet. Eine Übersicht über alle 22 Variationen wird im Ergebnissteil dieser Studie gegeben (Abschnitt 7.4.2).

Jede Testvariante wurde per Hand nach den zuvor gelisteten Regeln konstruiert. Grund für den manuellen Aufbau war die Abwägung des Zeitaufwands im Gegensatz zur Programmierung eines vollautomatischen Algorithmus zur Modul- und Testkonstruktion. Dieses Vorgehen findet sich häufig in internationalen Publikationen (vgl. Hendrickson, 2007; Zheng & Chang, 2015). Die entsprechenden Testinstrumente beinhalten aber auch typischerweise eine wesentlich höhere Menge an Items, womit sie nicht mehr realistisch per Hand erstellt werden könnten.

7.4.1 Methodik

Ein praktisches Hindernis beim Vergleich unterschiedlicher Testvarianten liegt in der Auswahl und Gewinnung von ausreichend vielen Proband*innen. Für jeden einzelnen Testdurchlauf muss eine Teilstichprobe aus der Zielpopulation gezogen werden, die a) repräsentativ für die Population, b) groß genug für eine Analyse des Testverhaltens und c) von allen anderen Teilstichproben unabhängig ist. Für Punkt b) kommt im Fall von adaptiven Formaten erschwerend hinzu, dass jeder einzelne Pfad im Test von ausreichend vielen Teilnehmer*innen durchschritten werden muss – im Gegensatz zu nur einem Testpfad bei linearen Formaten.

Um alle vorliegenden Strukturen und Testlängen in dieser Art prüfen zu können, wären 25 unabhängige Stichproben im Umfang von jeweils minimal 100 bis 200 Personen, je nach Teststruktur, notwendig. Um zu garantieren, dass jeder Testpfad von wenigstens 50 Personen durchlaufen wurde, wäre die notwendige Gesamtzahl an Teilnehmer*innen mindestens n = 2200[16]. Da zur Zeit der geplanten Studie in den Berliner Hochschulen nur etwa 1800 Personen für einen Lehramtsabschluss in der MINT-Fächergruppe eingeschrieben waren (Amt für Statistik Berlin-Brandenburg, 2019), war das Vorhaben so nicht realisierbar.

Es wurde deshalb entschieden, die verschiedenen Varianten anhand von Simulationen zu vergleichen, die Testdurchläufe also mit virtuellen Proband*innen durchzuführen.

Die im Ko-WADiS-Datensatz erfasste Stichprobe umfasst 2500 Personen. Sie weist eine Normalverteilung der Personenfähigkeiten auf und enthält Messdaten aus allen relevanten Studienfächern und Studienphasen. Zudem wurde sie über einen Zeitraum von mehreren Jahren aggregiert. Aus diesen Gründen wird sie als repräsentativ für die Zielpopulation

[16] Diese Zahl stellt das absolute Minimum dar. Es gibt 11 Testvarianten mit zwei und 11 Testvarianten mit drei Schwierigkeitsbereichen, woraus sich allein schon die 2200 notwendigen Teilnehmer*innen ergeben, wenn je genau 50 Personen in eines der Niveaus eingeordnet werden. Streuungen und Sprünge zwischen den Bereichen sind hier also noch nicht einkalkuliert, andernfalls wäre die Zahl weitaus höher. Im Sinne des Arguments reicht aber schon diese sehr unrealistische Schätzung.

angesehen und kann verwendet werden, um die Verteilungsparameter der Zielpopulation sowie die normierten Parameterwerte aller Items festzulegen. Somit ist es auch möglich, die gewünschten Stichproben virtuell zu generieren.

Die einzelnen Simulationen verliefen in drei Schritten (siehe auch Abbildung 18):

1. Generierung der virtuellen Stichprobe und ihres Antwortverhaltens
2. Generierung der simulierten Beobachtungen
3. Auswertung der Beobachtung durch 2pl-Modellberechnung

Der erste Schritt ähnelt einer vollständigen Imputation der echten Beobachtungen, mit denen zuvor gearbeitet wurde. Allen Personen wurde einmalig ein fixer EAP-Schätzwert für ihre Fähigkeit zugeschrieben, der auf den Ko-WADiS-Daten basiert. Dieser wird im Rahmen der Simulation als der „echte" Fähigkeitswert betrachtet. Danach wurde die Lösungswahrscheinlichkeit für jedes einzelne Item im Test mit diesem Wert sowie den Itemparametern gemäß dem verwendeten 2pl-Modell bestimmt. Anhand dieser Lösungswahrscheinlichkeit wurde dann eine zufällige Antwort aus den beiden Möglichkeiten „richtig" und „falsch" gezogen. Auf diese Weise entstand eine Matrix mit allen Proband*innen, Items und gemäß dem 2pl-Modell plausiblen Antworten.

Der Unterschied zu einer normalen Imputation ist, dass hier auch die real beobachteten Antworten verworfen und neu generiert wurden. Für das Verwerfen echter Daten gibt es zwei Gründe. Zum einen entsprechen so alle Einträge der Matrix einer perfekten Modellpassung – sie wurden schließlich rein anhand der Modellparameter und -gleichung generiert. Die echten Daten wurden durch weitere nicht modellierte Parameter beeinflusst. Würden sie im Datensatz bleiben, gäbe es eine interne Verzerrung der Daten in Bezug auf die Modellpassung, das macht eine spätere Beurteilung und Diskussion schwierig. Zum anderen kann so sichergestellt werden, dass die Stichproben mehrerer Simulationen voneinander unabhängig sind. Alle Antworten sind zufällig generiert, der

einzige Zusammenhang zwischen den Stichproben zweier Durchläufe ist die identische Verteilung der als echt angenommenen Fähigkeiten.

Der gesamte erste Schritt fand im Vorfeld der simulierten Messung statt und diente allein der Steuerung des Antwortverhaltens der virtuellen Stichprobe. Dem Messinstrument selbst standen in der Simulation weder die echten Fähigkeiten noch die vollständige plausible Antwortmatrix zur Verfügung, da dies in einer realen Messung auch nicht der Fall wäre.

Der zweite Schritt entspricht der Messung. Allen Proband*innen wurde das erste Modul des Tests zugewiesen. Die Beobachtungen, also die im simulierten Testdurchlauf gegebenen Antworten, wurden für jede Item-Proband*innen-Kombination aus der Matrix mit plausiblen Daten ausgelesen und in einem zweiten, zu Testbeginn noch leeren Datensatz gespeichert. Dieser wurde dann genutzt, um eine Fähigkeitsschätzung für das Routing vorzunehmen. Den Regeln folgend wurden die nächsten Blöcke zugewiesen und diese Schritte bis zum Testende wiederholt. Es entstand so eine unvollständige Datenmatrix mit Beobachtungen, ebenso wie es auch bei einer echten Messung der Fall wäre.

Den dritten Schritt und Abschluss jeder Simulation bildete die Auswertung. Ebenso wie es auch bei einer realen Messung der Fall wäre, wurde dies anhand einer 2pl-Modellberechnung mit den Beobachtungen und festgesetzten Itemparametern durchgeführt. Abzusichern ist bei der Durchführung von Simulationsstudien, welchen Einfluss die Annahme der perfekten Modellpassung auf die Ergebnisse hat. In Wirklichkeit weichen Modellannahmen und Beobachtungen voneinander ab, es gibt also unbekannte Einflüsse auf das Antwortverhalten der Proband*innen. Diese wurden in der Modellgleichung nicht berücksichtigt und flossen damit auch nicht in die Ergebnisse der Simulationen ein – hier wurde vorausgesetzt, dass einzig und allein der Fähigkeitswert das Antwortverhalten steuert.

Eine bessere Modellpassung bedeutet allgemein genauere Vorhersagen. Das bedeutet, dass die in den Simulationen erreichten Messgenauigkeiten vermutlich höher als die in Realität erreichbaren Werte sind.

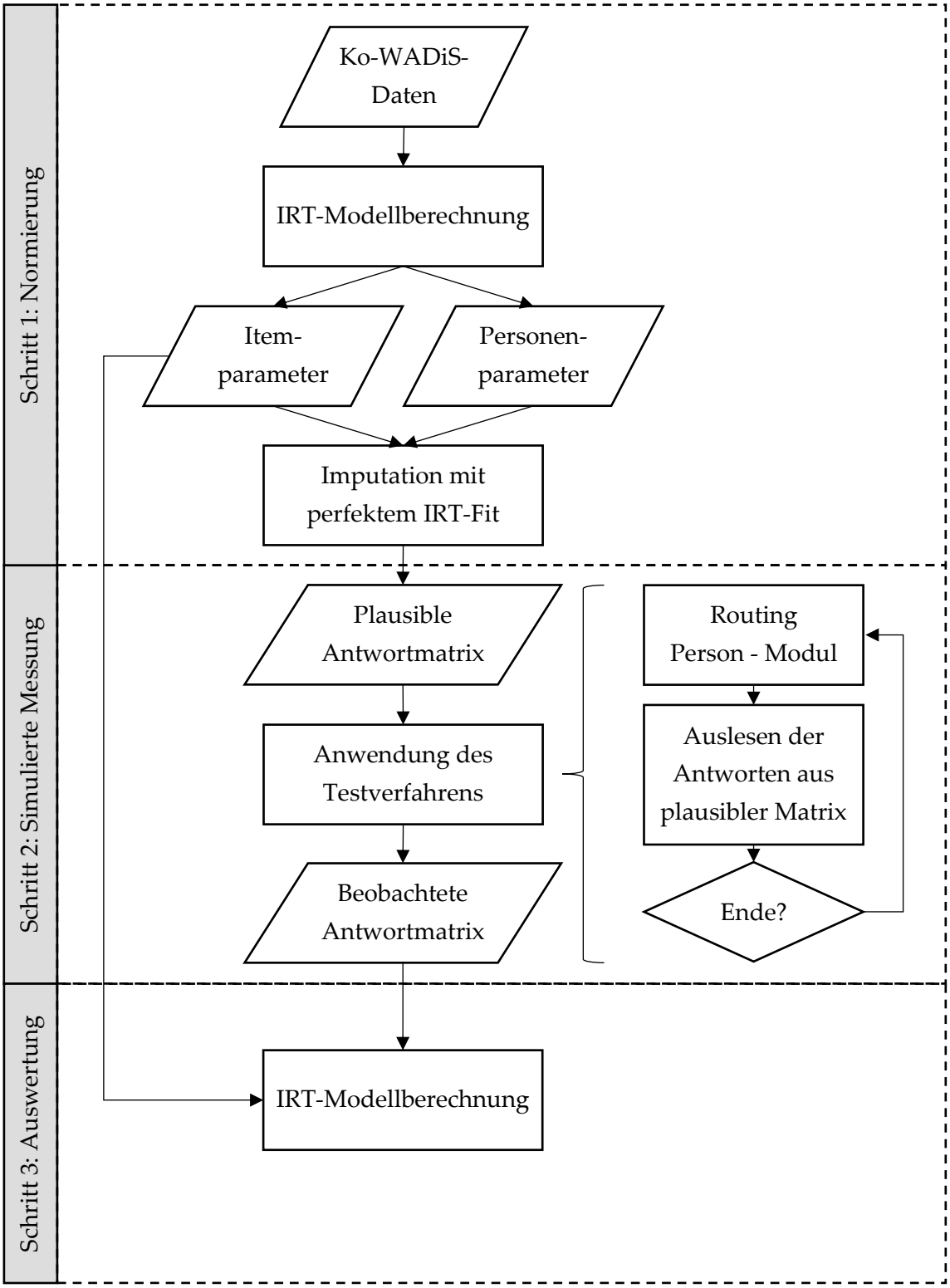

Abbildung 18: Ablauf der einzelnen Simulationsstudien. Prozesse sind als Rechtecke, Datensätze als Parallelogramme dargestellt. Zum Testverfahren in Schritt 2 Siehe auch Abschnitt 4.2 für allgemeine Fragen zu MSTs und Abschnitt 7.3 für die konkreten Routing-Regeln.

Es stellt sich die Frage, wie groß diese Abweichung ist. Grundsätzlich wurde die Abweichung von modellierten und realen Daten schon durch die Maße für Modellfit besprochen und ausgewertet. Leider lässt sich aus diesen Werten nur schwer herauslesen, welcher Einfluss in Bezug auf die Reliabilität besteht – für diesen Zweck wurden die Fitmaße nicht erstellt. Um den Einfluss dieses Fehlers einschätzen zu können, wurden stattdessen bereits mit echten Proband*innen gewonnene Daten reproduziert: Die linearen Testhefte wurden unter den gleichen Bedingungen wie die MST-Varianten simuliert.

Hierfür wurde für alle Proband*innen im Ko-WADiS-Datensatz zuerst geprüft, welches Testheft von ihnen bearbeitet wurde. Mit den geschätzten Personenfähigkeiten wurden dann für alle im Testheft vorkommenden Items Antworten unter Annahme eines perfekten Modellfits generiert.

Im Anschluss wurden für jedes Testheft die vorliegenden Ko-WADiS-Daten sowie die simulierten Antworten in zwei Datensätzen zusammengefasst. Diese wurden anhand der fixierten Itemkennwerte in getrennten 2pl-Modellen ausgewertet und die Messgenauigkeiten verglichen.

Die notwendigen Skripte für die Simulationen wurden alle selbst in R (R Core Team, 2020, Version 4.0.0) erstellt, es wurde keine vorgefertigte Software verwendet. Neben den Basispaketen und TAM (Robitzsch et al., 2020, Version 3.5-19) wurden die Pakete doSNOW (Microsoft Corporation & Weston, 2019b, Version 1.0.18), snow (Tierney, Rossini, Li & Sevcikova, 2018, Version 0.4-3) und doParallel (Microsoft Corporation & Weston, 2019a, Version 1.0.15) verwendet, um die Skripte zu parallelisieren[17].

7.4.2 Ergebnisse

Die Ergebnisse der Simulationen zu den MSTs sind in Abbildung 19 und Abbildung 20 sowie in Tabelle 7 dargestellt. Wie auch beim Modellvergleich wird die Messgenauigkeit als eines der Vergleichsmaße verwendet.

[17] Parallelisierung beschreibt in diesem Fall Methoden, um die Simulationen in mehreren Stücken parallel, aber synchronisiert auf mehreren Prozessoren ausführen zu können, um die Berechnungen schneller auszuführen.

Bei allen durchgeführten Simulationen erwies sich die EAP/PV-Reliabilität als besser als die der WLEs, weshalb diese im Vergleich genutzt wird.

Bei Betrachtung der Grafiken zeigen sich mehrere Trends: Erstens scheint in beiden Gruppen, also sowohl bei einer Aufspaltung in zwei als auch in drei Schwierigkeitsniveaus, jeweils die Struktur mit drei aufeinanderfolgenden Stufen die effizienteste zu sein. Zweitens liegt die Gruppe der Testvarianten mit nur zwei Schwierigkeitsniveaus leicht über der mit dreien. Weiterhin steigt die Messgenauigkeit der verschiedenen Teststrukturen mit einer höheren Itemzahl an, was zu vermuten war. Ein Ausreißer ist Teststruktur 1-3-3-3 in Länge 8[18] (Abbildung 20), der erreichte Wert liegt höher als die restlichen Trendlinien der Gruppe es erwarten lassen würden. Der Einfluss der Testlänge ist insgesamt größer als die der verschiedenen Aufspaltungen der Teststruktur und besonders stark bei geringen Itemzahlen.

Die exakten Werte der Reliabilitäten und Effizienz aus allen Simulationen sind in Tabelle 7 dargestellt. Insgesamt erweist sich die Teststruktur 1-2-2 in Länge 15 mit einem Wert von Rel(EAP) = .644 als die genaueste. Dieses Ergebnis steht in Einklang mit den beobachteten Trends in den restlichen Daten. Allerdings kann aus der Tabelle auch ausgelesen werden, dass die kürzesten Testvariationen die höchste Effizienz aufweisen. Die 1-3-3-3-Struktur hat bei einer Länge von 8 Items den größten Wert von allen Variationen (E = .066), wobei es sich wie bereits besprochen um einen Ausreißer handelt. Wird dieser verworfen, ist stattdessen die Struktur 1-2 in Länge 8 mit einem Wert von E = .063 an erster Stelle, siehe dazu die Diskussion (Abschnitt 7.4.3).

Von den tendenziell messgenauesten Tests, also den längsten Varianten aller Strukturen, weist wiederum Teststruktur 1-2-2 in Länge 15 den Höchstwert von E = .043 auf.

Die Ergebnisse der zweiten Simulationsreihe sind in Tabelle 8 zu sehen. Wie gut die Daten von echten Messungen anhand der Simulationen

[18] Hiermit wird die Testvariante bezeichnet, die eine 1-3-3-3-Struktur sowie eine Gesamttestlänge von 8 Items hat.

reproduziert werden konnten, schwankt je nach Testversion. Im Mittel wird die Messgenauigkeit in den simulierten Daten um 0.02 überschätzt, die stärksten Abweichungen sind eine Überschätzung um 0.064 und eine Unterschätzung um -0.056 Punkte.

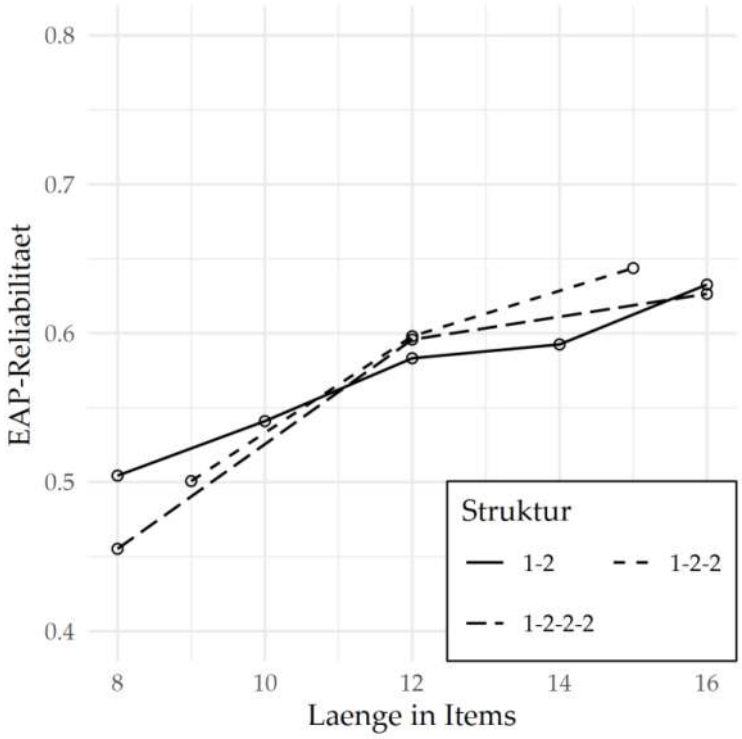

Abbildung 19: Dargestellt sind die in der Simulationsstudie erreichten EAP-Reliabilitäten der MST-Versionen mit einer Aufspaltung in zwei Schwierigkeitsbereiche. Die Linien zeigen die Ergebnisse der Versionen, Gruppiert nach der Anzahl der Stufen im Test (siehe Legende) und aufgetragen über die Gesamttestlänge in Items.

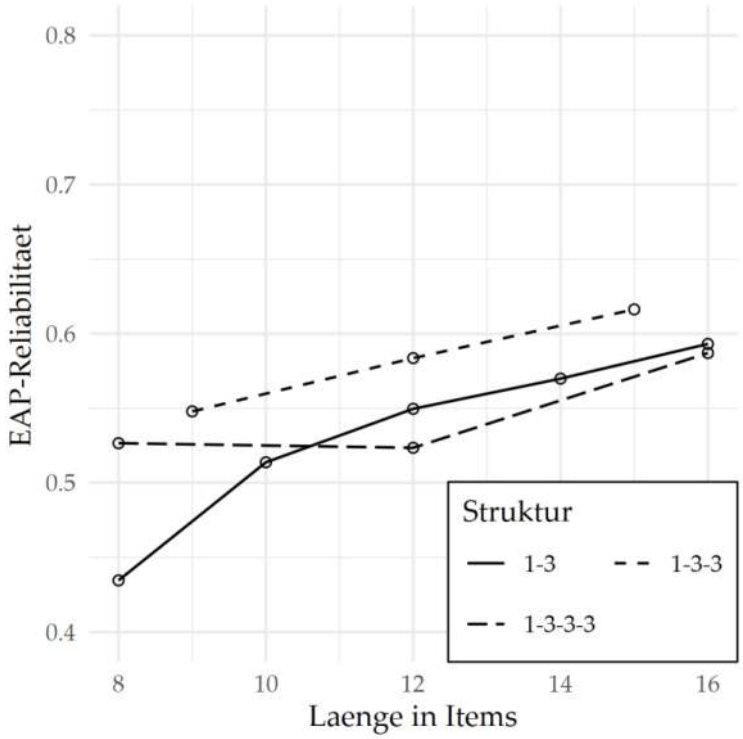

Abbildung 20: Dargestellt sind die in der Simulationsstudie erreichten EAP-Reliabilitäten der MST-Versionen mit einer Aufspaltung in drei Schwierigkeitsbereiche.

*Tabelle 7: Überblick der Simulationsergebnisse zu verschiedenen MST-Strukturen. Dargestellt sind die Modulanordnung und Gesamttestlänge (in Anzahl der zu bearbeitenden Items) zur Kennzeichnung der einzelnen Tests. Daneben sind die in der Messung erreichte mittlere EAP/PV-Reliabilität eingetragen sowie die Korrelation zwischen den gemessenen Personenfähigkeiten und denjenigen, mit denen die virtuellen Proband*innen gesteuert wurden. Der höchste Wert ist gekennzeichnet. Als letztes ist die Effizienz jeder einzelnen Testvariante eingetragen. Auch hier ist der höchste Wert in Dunkelgrau gekennzeichnet, zusätzlich ist der Wert der längsten Testversion aus jeder Struktur mit einem Stern versehen.*

Struktur	**Länge in Items**	**EAP/PV-Reliabilität**	**Korrelation echte Fähigkeiten/Messung**	**Effizienz (Reliabilität/Länge)**
1-2-2-2	12	.596	.792	.050
1-2-2-2	16	.626	.817	.039*
1-2-2-2	8	.455	.703	.057
1-2-2	12	.598	.815	.050
1-2-2	15	.644	.834	.043*
1-2-2	9	.501	.74	.056
1-2	10	.541	.708	.054

Struktur	**Länge in Items**	**EAP/PV-Reliabilität**	**Korrelation echte Fähigkeiten/Messung**	**Effizienz (Reliabilität/Länge)**
1-2	12	.583	.732	.049
1-2	14	.593	.716	.042
1-2	16	.633	.765	.040*
1-2	8	.504	.652	.063
1-3-3-3	12	.524	.735	.044
1-3-3-3	16	.587	.789	.037*
1-3-3-3	8	.527	.762	.066
1-3-3	12	.584	.789	.049
1-3-3	15	.616	.802	.041*
1-3-3	9	.548	.765	.061
1-3	10	.514	.681	.051
1-3	12	.55	.73	.046
1-3	14	.57	.717	.041
1-3	16	.593	.748	.037*
1-3	8	.435	.675	.054

Tabelle 8:Überblick über die Simulationsergebnisse zu linearen Testheften. Dargestellt sind die in realen und in simulierten Messungen erreichten EAP-Reliabilitäten der neun aktuell eingesetzten Testhefte des linearen Ko-WADiS-Tests. Als Überschätzung ist die Differenz des simulierten und realen Wertes eingetragen. Positive Werte bedeuten eine zu hohe Einschätzung der Messgenauigkeit in der Simulation, negative Werte eine zu niedrige.

	EAP real	**EAP simuliert**	**Überschätzung EAP**
Heft 1	.521	.551	.031
Heft 2	.606	.633	.028
Heft 3	.635	.663	.028
Heft 4	.602	.666	.064
Heft 5	.618	.681	.063
Heft 6	.615	.559	-.056
Heft 7	.573	.597	.024
Heft 8	.674	.652	-.022
Heft 9	.551	.57	.019
Mittelwert			.02

7.4.3 Diskussion

Abgleich der Simulation mit Realdaten

Als erstes sollte die Güte der Simulationen beurteilt werden, da darauf die Aussagekraft aller weiteren Schlussfolgerungen basiert: Die erreichten Messgenauigkeiten bei echten Stichproben werden durch das Simulationsvorgehen um 0.02 ± 0.04 überschätzt.

Es kann leider nicht geprüft werden, ob dieser Effekt mit der Testlänge skaliert, da für einen solchen Vergleich keine entsprechenden Realdaten vorliegen. Zu vermuten wäre, dass es sich um einen relativen Effekt handelt, der das Maß der Abweichungen zwischen Realität und Modell in der Stichprobe widerspiegelt. In diesem Fall wären kürzere Tests, was die absoluten Abweichungen betrifft, weniger stark betroffen.

Die Streuung der Effektgröße zwischen den verschiedenen Testheften wird sich vermutlich durch eine verschieden gute Modellpassung der enthaltenen Items begründen. Hierzu ist anzumerken, dass in den betrachteten Testheften auch Items enthalten sind, die in der Modellanpassung aus Abschnitt 6.4.2 ausgeschlossen wurden. Sofern dies zutrifft, sind die besonders starken Abweichungen auf einen schlechteren Modellfit zurückzuführen, als er bei den Items im MST auftritt. Da Itemausschluss und -auswahl bei der Konstruktion der MSTs strenger war, sind im Mittel geringere Fehleinschätzungen zu vermuten. Auch diese Vermutung kann aber nicht nachgewiesen werden, da die notwenigen Realdaten nicht vorliegen.

Für die weitere Betrachtung der MST-Simulationen wird davon ausgegangen, dass die Reliabilität grundsätzlich um einen Wert von 0.02 überschätzt wird, da es sich hierbei um den beobachteten Mittelwert handelt und eine eventuelle Systematik der Abweichungen vermutet, aber nicht belegt werden kann. Es werden sich damit für den Vergleich der MST-Strukturen untereinander keine Konsequenzen ergeben.

Vergleich der MST-Strukturen

Bei der Interpretation der beobachteten Unterschiede zwischen verschiedenen Strukturgruppen ist Vorsicht angeraten. Zum einen gibt es innerhalb jeder Gruppe nur zwischen drei und fünf Datenpunkte, womit rechnerische Vergleiche ausgeschlossen sind. Zum anderen ist zu vermuten, dass die Gesamtlänge der Tests im betrachteten Bereich (die Maximallänge ist 16) kombiniert mit den Strukturen einen deutlichen Einfluss auf die Messgenauigkeit hat.

So ist denkbar, dass die Messgenauigkeit aller simulierten Testvarianten einfach nicht hoch genug ist, um Proband*innen sicher in mehr als zwei getrennte Fähigkeitsniveaus einzustufen – unabhängig von der gewählten Struktur. Ein scheinbarer Vorteil der Aufspaltung in zwei Schwierigkeitsbereiche wäre dann in Anteilen durch die Länge des Testinstruments begründet. Beim Einsatz längerer Versionen, beispielsweise mit 25 Items Gesamtlänge, könnte eine feinere Aufspaltung von Vorteil sein. Da der Itempool die Konstruktion längerer Tests nicht ermöglicht, kann das aber in keiner Weise geprüft werden. Weitere Spekulationen sollen hier vermieden werden.

Zudem zeigte sich auch in der zitierten Literatur (vgl. Abschnitt 4.3.2 und 4.4.1) keine allgemein beste Struktur, weshalb die Simulationen überhaupt erst notwendig waren. Aus diesem Grund verbleibt es bei einer rein deskriptiven Ergebnisdarstellung ohne rechnerische Vergleiche der Gruppen untereinander. Trotzdem gibt es ein paar Punkte zu diskutieren.

Eine der Strukturgruppen weisen unerwartetes Verhalten auf, betroffen sind die Tests mit 1-3-3-3 Struktur. Hier sinkt die Messgenauigkeit, wenn der Tests von acht auf zwölf Items verlängert wird. Da keine Items ausgetauscht, sondern nur neue Items in den verlängerten Test eingefügt wurden, ist dieses Ergebnis irritierend: Auch wenn die zusätzlichen Items einen geringeren Informationsgewinn ermöglichen als die bereits in der kurzen Version enthaltenen, sollte jeder Mehrgewinn an Information auch eine Steigerung der Messgenauigkeit darstellen.

Grundlegend wäre es vorstellbar, dass der probabilistische Charakter des 2pl-Modells, das der Simulation und Proband*innen-Steuerung zugrunde liegt, fehlerhafte Ergebnisse verursachen kann. Da die Antworten den errechneten Lösungswahrscheinlichkeiten nach zufällig gezogen werden, könnte ein insgesamt unwahrscheinliches Proband*innenverhalten generiert werden und die Messung beeinflussen. Für einzelne Personen wird dies sicherlich in jeder Simulation gelten, da die virtuelle Stichprobe einen Umfang von N = 2584 hat. Der Fall, dass ein signifikanter Anteil dieser Stichprobe ein rein zufällig atypisches Verhalten aufweist, kann aber allein durch den Stichprobenumfang ausgeschlossen werden. Um sicher zu gehen, wurden die Simulationen trotzdem mehrfach durchgeführt, wobei jedes Mal ein neues Antwortmuster für die gesamte Stichprobe generiert wurde. Das Testverhalten blieb über alle Durchläufe stabil bei den in Tabelle 7 berichteten Reliabilitäten. Ein Zufallseffekt wird damit endgültig als Erklärung verworfen.

Der Ansatz des Zufallseffekts wurde explizit geprüft, weil dieser auch Implikationen für alle anderen Strukturen gehabt hätte (es hätten dann weitere Zufallseffekte und Verschiebungen der Strukturgruppen gegeneinander ausgeschlossen werden müssen). Es gibt daneben noch weitere Ansätze, die denkbar sind:

- Eine Übersteuerung im Routing-Modul des kürzesten 1-3-3-3-Tests, begründet durch nur zwei vorhandene Items im Modul und dadurch extrem geringe Informationsmengen. Unabhängig von der Gültigkeit der ersten Schätzwerte könnte diese Übersteuerung zu einem zufällig korrekten Routing führen.
- Ein sprunghafter Abfall der Itemgüte zwischen einer der hinteren drei Stufen im längeren Test, da gute Items bei der Verlängerung des Tests in die vorderen Stufen gezogen wurden.

Bei Betrachtung der einzelnen Module sowie bei der Ausführung des Routings gibt es aber keine Hinweise darauf, dass einer dieser Ansätze die Beobachtung erklären kann. Die Ursache bleibt damit an dieser Stelle ungeklärt. Der Datenpunkt zur Struktur 1-3-3-3 mit Länge 8 wird als Ausreißer verworfen, um keine Rückschlüsse aus unverständlichen Daten zu ziehen.

Als letztes gilt es in diesem Abschnitt, Forschungsfrage 2 abschließend zu beantworten: Welche der betrachteten Strukturen ist am besten geeignet, um die Effizienz des Ko-WADiS-Tests zu steigern?

Tabelle 7 lässt vermuten, dass die Antwort in Struktur 1-2 liegt, da diese den größten Effizienzwert aufweist. Hier ist aber wichtig, die Art der Testkonstruktion und notwendige Absolutwerte der Messgenauigkeit zu berücksichtigen. Da die Module nach dem Prinzip der forward-assembly mit Items gefüllt wurden, ist der Informationsgewinn pro Item bei kurzen Testversionen zwangsweise höher. Diese werden mit den besten Items gefüllt, für längere Testversionen muss auf weniger trennscharfe und informative Items zurückgegriffen werden. Eine logische Folge davon ist das Absinken der Effizienz innerhalb einer jeden Struktur bei Verlängerung des Tests, da die mittlere Güte der Items abnimmt. Trotzdem steigt durch die hinzugefügten Items die Messgenauigkeit des Tests. Da die Items sich bei weiterer Verlängerung immer mehr den Mindestanforderungen bei der Itemauswahl annähern, im Falle des 2pl-Modells war dies vor allem die Trennschärfe von mindestens 0.4, wird sich dieser Effekt abschwächen. Es ist zu erwarten, dass bei einem ausreichend großen Itempool die Trendkurve jeder einzelnen Struktur abflacht.

Es sollte daher nicht geprüft werden, welche Testvariante insgesamt die höchste Effizienz aufweist, sondern welche dies bei einer akzeptablen Messgenauigkeit tut. Da keine Variante einen Wert von 0.7 erreicht, wird als akzeptable Grenze eine Reliabilität von 0.6 festgelegt, was den Minimalstandard der bereits in Kapitel 5 zitierten Literatur darstellt.

In diesem Bereich finden sich nur noch die jeweils längsten Versionen der verschiedenen Strukturen, davon weist die Struktur 1-2-2 sowohl die beste Effizienz als auch die höchste Messgenauigkeit auf.

FF 2 kann damit abschließend wie folgt beantwortet werden: Unter Berücksichtigung von Mindeststandards der Messgenauigkeit ist ein Multistage-Test in 1-2-2-Struktur, der nach den in Abschnitt 7.2 und 7.3 definierten Regeln konstruiert wurde, das beste Verfahren zur Effizienzsteigerung des Ko-WADiS-Tests.

8 Testpilotierung und Vergleichsstudie (FF 3)

In den vorigen beiden Kapiteln 6 und 7 wurde der Ko-WADiS-MST konstruiert und soweit möglich schon im Vorfeld einer Datenerhebung optimiert. Im Zusammenhang damit wurden FF 1 und FF 2 beantwortet. Die praktische Umsetzung des Tests sowie der Einsatz in einer realen Befragung sind noch notwendig, um einen Vergleich zum linearen Ko-WADiS-Test zu ziehen und FF 3 beantworten zu können. In diesem Kapitel wird daher zunächst dargestellt, mit welcher Software und Infrastruktur der MST in die Praxis implementiert wurde.

Danach wird in Abschnitt 8.2 über die Pilotierungsstudie berichtet, mit der zum ersten Mal Realdaten durch den MST erhoben wurden. Diese werden im letzten Abschnitt 8.3 des Kapitels mit den Bestandsdaten aus Einsätzen des linearen Formats verglichen, um den MST in Bezug auf seine tatsächliche Effizienzsteigerung zu beurteilen.

8.1 Praktische Umsetzung

Für den ersten Einsatz des MST wurde die Plattform *tet.folio* (Technology Enhanced Textbook, https://tetfolio.fu-berlin.de/) ausgewählt.

Es handelt sich dabei um eine interaktive Lehr-Lern-Plattform sowie um eine darin integrierte Web-Applikation der Freien Universität Berlin. Ursprünglich wurde tet.folio entwickelt, um einen digitalen Lernraum zu erstellen, der durch eigenständig gesammelte Materialien sowie durch digitale Messungen und interaktive Bildschirmexperimente angereichert werden kann (Haase, Kirstein & Nordmeier, 2016). Im Verlauf der Entwicklung wurden aber immer mehr Möglichkeiten der Medieneinbindung und Gestaltung integriert, sodass inzwischen auch Tools für Umfragen, Lehrevaluationen und Ähnliches vorhanden sind. Im Folgenden wird die Gestaltung des MST vorgestellt, dabei werden auch Vor- und Nachteile der Umsetzung aufgelistet und verglichen.

Zunächst ist festzustellen, dass der Zugriff von Proband*innen auf die Plattform sehr einfach ist. Die Webseite kann sowohl von Desktopcomputern als auch mobilen Geräten mit allen gängigen Betriebssystemen und

Webbrowsern aufgerufen werden[19]. Eine Anmeldung ist für die Proband*innen nicht notwendig. Zudem kann die Befragung auch im Privatmodus des Browsers ausgeführt werden, da die notwendigen Dateien für die Testdurchführung serverseitig gespeichert werden. Somit kann auch von Seiten der Proband*innen der Schutz persönlicher Daten gewährleistet werden.

Das Design der einzelnen Bestandteile des Testinstruments wurde möglichst nah an dem der Papierversion gehalten. Damit ist gemeint, dass alle Items sowie die Datenschutzbelehrungen, Kovariaten zu Studienfach und Personencode und ähnliches auf einzelnen Seiten erstellt wurden, die theoretisch sogar ausgedruckt und als lineares Testheft verwendet werden könnten (siehe dazu Abbildung 21 bis Abbildung 23).

In der Praxis bearbeiten die Proband*innen also seitenweise einen Test, ebenso wie sie es auch bei einer papierbasierten Messung tun würden. Der Unterschied besteht darin, dass die präsentierten Seiten im Hintergrund von der Testlogik adaptiv ausgewählt werden.

Die Designentscheidung wurde getroffen, um keine Fragen bezüglich der Validität des Instruments sowie der Vergleichbarkeit mit dem linearen Instrument aufkommen zu lassen. Eine Umgestaltung hätte das Antwortverhalten beziehungsweise die Informationsverarbeitung beeinflussen können.

[19] Das umfasst Windows, GNU/Linux, macOS sowie Firefox, Chrome, Internet Explorer, Safari und in eingeschränktem Maß Opera.

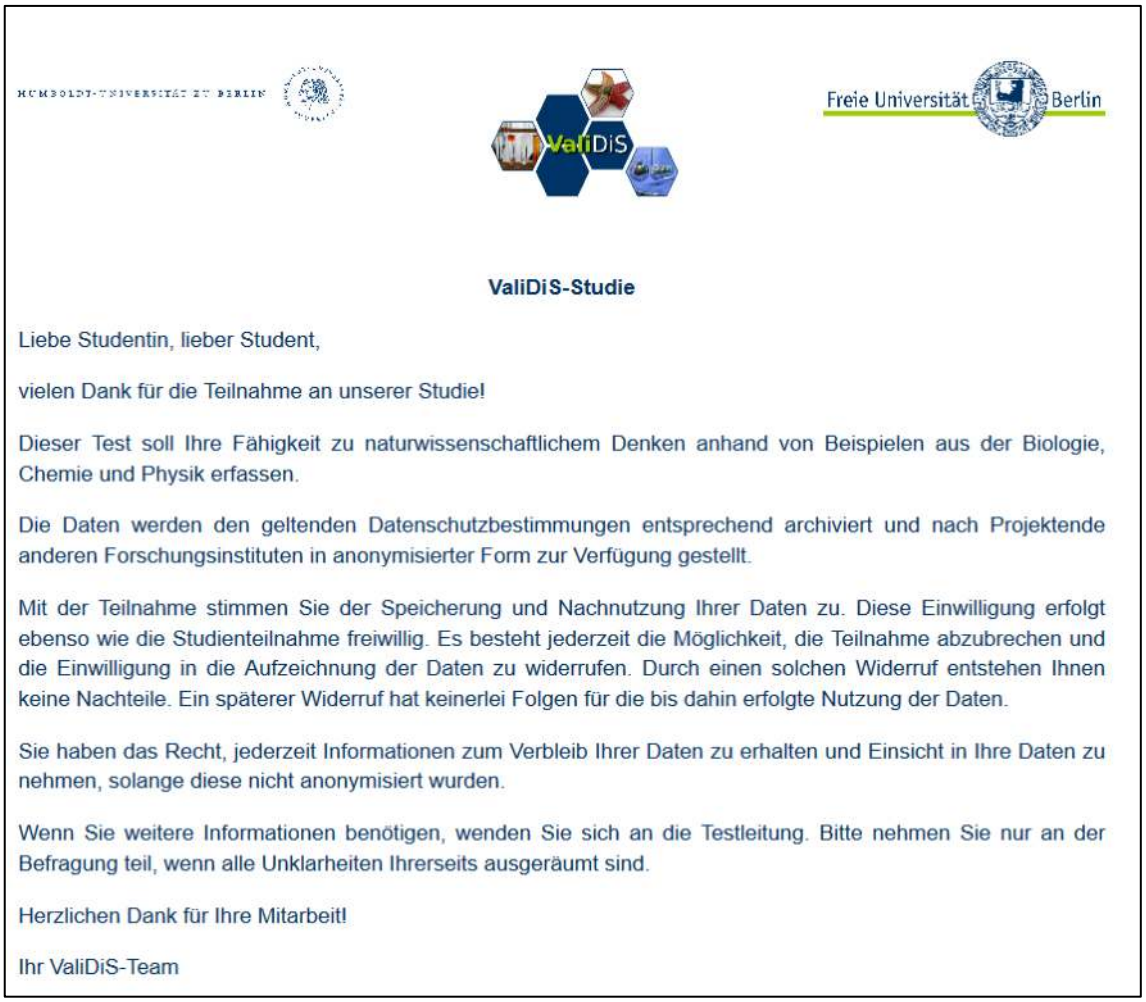

Abbildung 21: Screenshot der Startseite des MST. Aus Datenschutzgründen wurden die Kontaktdaten der Testleitung und Verantwortlichen aus dem Bild entfernt, die zum Zeitpunkt der Erhebungen noch am unteren Rand der Seite in Tabellenform aufgelistet wurden.

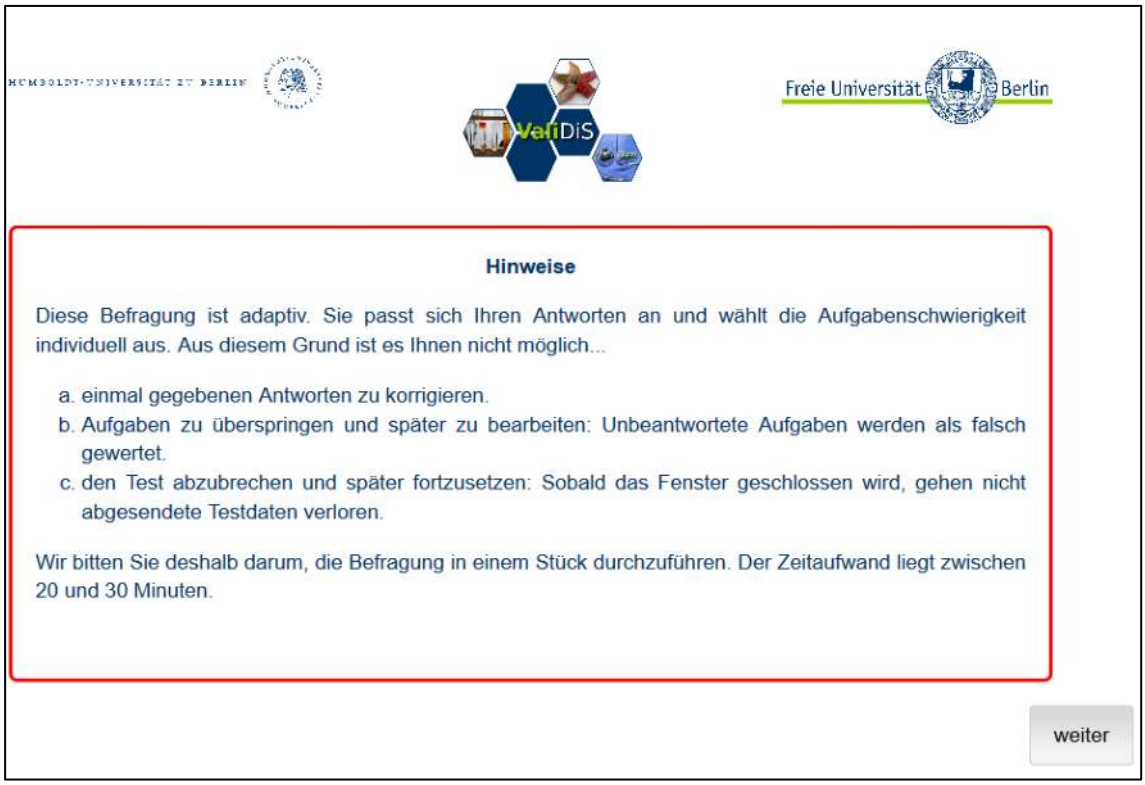

Abbildung 22: Screenshot der dritten Seite des MST; Nach der Startseite (Projektvorstellung) und den Datenschutzinformationen sowie den Kovariaten erfolgte noch eine Stichpunktartige Erinnerung an Besonderheiten der Umfrage, die zuvor von der Testleitung verbal erläutert wurden.

Eine technische Einschränkung gab es bei der Korrekturmöglichkeit von Antworten. Grundsätzlich ist es ein Vorteil von MSTs gegenüber von CATs, dass sie innerhalb von Modulen eine Antwortkorrektur erlauben. So kann auch ein im Augenblick zu schwer erscheinendes Item kurz übersprungen und später beantwortet werden, solange in der Zwischenzeit kein Modul beendet und ein Routingschritt vorgenommen wurde. Die Teilnehmer*innen einer Befragung sind so in ihrem Verhalten weniger eingeschränkt und unter Druck gesetzt.

Dieser Vorteil konnte in der Pilotierung nicht genutzt werden (siehe dazu auch die Hinweise in Abbildung 22). Aus technischen Gründen war es nicht möglich, gleichzeitig eine volle Anonymität der Proband*innen (siehe dritter Absatz in diesem Abschnitt) und diese Sprungmöglichkeit zu gewährleisten. In Absprache mit allen beteiligten Verantwortlichen wurde entschieden, das Defizit im Sinne des Datenschutzes in Kauf zu nehmen.

Abbildung 23: Screenshot eines Beispielitems. Es handelt sich um "PT_Meeresspiegel_021" aus den Bereichen "Physik" und "Testen von Modellen" (Itemschwierigkeit -0.43 und Trennschärfe 0.47). Für einen direkten Vergleich zwischen der digitalen und der Papierversion des Items siehe Abbildung 4.

Für die Testleitung hat die Umsetzung über tet.folio den Vorteil, dass die Plattform selbst die Arbeit mit sogenannten Büchern vorsieht, die als einzelne Seiten erstellt werden (siehe hierzu Abbildung 24 und die Hinweise in der Abbildungsbeschriftung). Die Bearbeitung aller Bestandteile des MST erfolgt direkt im Browser, die einzelnen Inhalte werden seitenweise in der integrierten html-Umgebung erstellt.

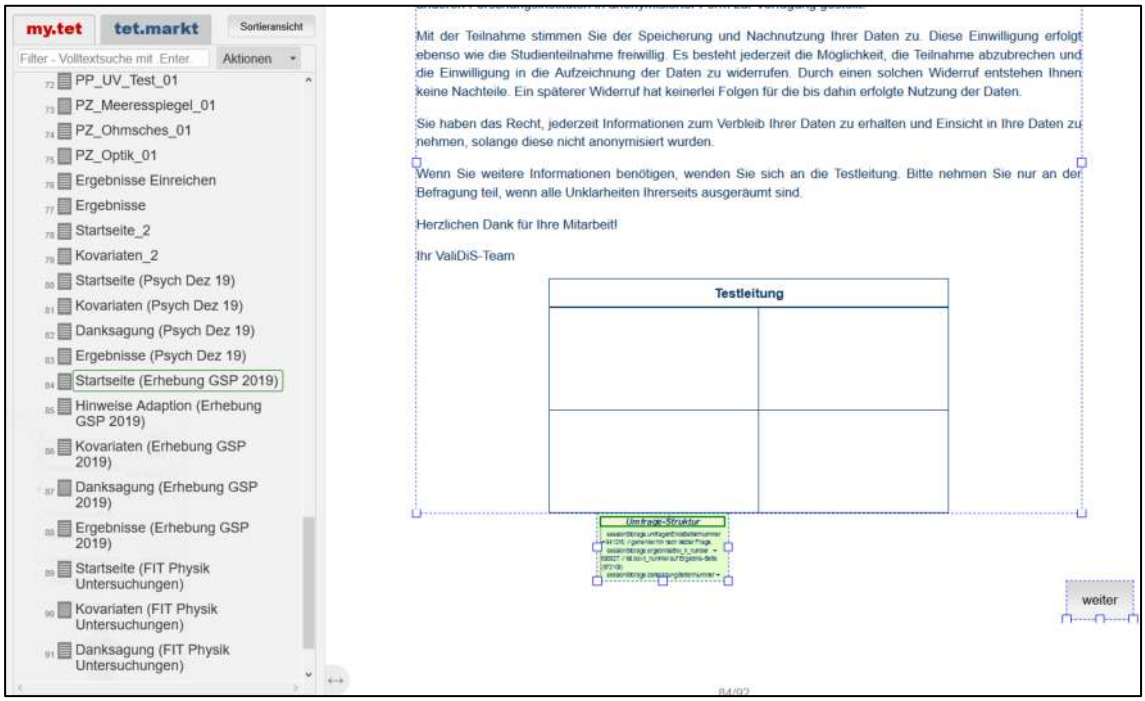

Abbildung 24: Bearbeitungsansicht des MST in der Entwicklungsumgebung des tet.folio. Links sind die einzelnen Seiten des tet.folio-Buches zu sehen, hier entspricht jede Seite einem „Blatt" des Tests, also beispielsweise einem Item. Auf der rechten Seite ist die aktuell ausgewählte Startseite des MST in Bearbeitungsansicht zu sehen. Aus Datenschutzgründen wurden die Kontaktdaten der Testleitung und Verantwortlichen aus dem Bild entfernt, die zum Zeitpunkt der Erhebungen noch in Tabellenform aufgelistet wurden. Am unteren Rand ist in Grün eine Box zu sehen, die außerhalb der Bearbeitungsansicht versteckt ist. Es handelt sich dabei um das Script mit der gesamten Teststruktur und -logik, das beim Aufruf der Startseite geladen wird – es ist in der Seite hinterlegt und kann direkt bearbeitet werden. Hierdurch ist es möglich, verschiedene Startseiten mit unterschiedlichen Teststrukturen in einem tet.folio-Buch zu erstellen. Durch das Aufrufen der entsprechenden Startseite wird die gewünschte Teststruktur geladen, alle MST-Versionen teilen sich aber den gemeinsamen Itempool im Buch. Der MST kann also, nachdem er nun erstellt wurde, sehr schnell angepasst und verschickt werden.

Daneben konnte es vermieden werden, einen eigenständigen Server zur Personenschätzung in der Messung verwenden zu müssen. Im Normalfall muss bei jeder Form von adaptiven Messungen im Hintergrund die Fähigkeitsschätzung durch das gewählte IRT-Modell vorgenommen werden. Hierzu könnten beispielsweise an jedem Routingpunkt die gesammelten Personendaten an einen dedizierten Server gesendet werden, auf dem die notwendigen Berechnungen durch ein vordefiniertes Skript ausgeführt und die Ergebnisse zurückgesendet werden. Dies stellt den bereits zuvor besprochenen Mehraufwand gegenüber linearen Tests dar.

Hier half die geringe Länge des Tests, die ansonsten eher eine Schwierigkeit dargestellt hatte: Das Instrument ist 15 Items lang, jedes Modul enthält 5 Items. Es gibt zwischen den drei Stufen zwei Zweigstellen, an denen ein Routing vorgenommen wird. Dabei gibt es für jedes Item zwei Antwortmöglichkeiten, richtig und falsch. Nach dem ersten Modul wurden somit 5 Items mit je zwei Möglichkeiten bearbeitet. Es gibt also $2^5 = 32$ Antwortkombinationen, die beim Routing zwischen der ersten und zweiten Stufe des Tests berücksichtigt werden müssen. Nach der zweiten Stufe wurden von jeder Person 10 Items bearbeitet, womit hier die Kombinationen auf $2^{10} = 1024$ ansteigen. Diese Zahl ist klein genug, um alle möglichen Pfade im Voraus festzulegen.

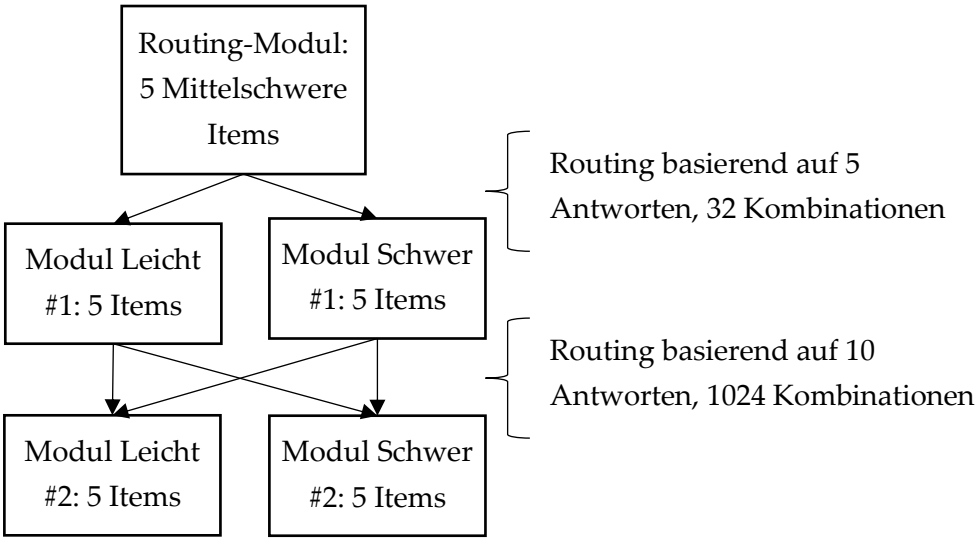

Abbildung 25: Panel-Struktur des umgesetzten MST und Anzahl der möglichen Antwortkombinationen, die zu den Zeitpunkten des Routings existieren.

Es wurden deshalb alle im Test möglichen Antwortkombinationen erstellt und für beide Routingpunkte Fähigkeitsschätzungen für jede Kombination ermittelt. Danach wurde auf Basis der Schätzwerte eine Routing-Entscheidung getroffen und dem entsprechenden Antwortmuster fest zugewiesen. Bei der Umsetzung im tet.folio konnte diese Zuordnung einfach eingespeichert werden: Das Antwortmuster wird in der Messung gespeichert und an jeder Verzweigung die entsprechende Entscheidung aus den gespeicherten Zuordnungen ausgelesen. Da die nötigen IRT-Berechnungen bereits vor der eigentlichen Messung durchgeführt wurden, besteht so kein Bedarf für weitere Infrastruktur. Es wäre leistungstechnisch betrachtet möglich, den gesamten MST auf einem lokalen Server, beispielsweise einem normalen Desktoprechner, zu hosten. Im Fall dieser Studie wurde aber keine gesonderte Lösung gesucht, da das tet.folio selbst schon auf FU-internen Servern gehostet und gewartet wird und somit alle Datenschutzanforderungen erfüllt werden konnten.

8.2 Pilotierungsstudie

Sinn der Pilotierung war es, den nun einsatzbereiten MST anhand von echten Personen zu erproben. Hauptsächlich diente dies dem Ausschluss von den nicht endgültig abschätzbaren Einflüssen der Simulationen auf die Testbeurteilung, die in Abschnitt 7.4.3 besprochen wurden. Könnten diese Einflüsse sicher identifiziert und vorhergesagt werden, hätte ein Vergleich der beiden nun vorhanden Testversionen schon durch die Simulation selbst erfolgen können.

Da dies leider nicht möglich war, mussten die Tests in realistischen Szenarien verwendet und verglichen werden. Das bedeutet, dass die hierfür ausgewählte Stichprobe dem späteren Einsatzbereich des Testinstruments entsprechen sollte. Dieser wurde bereits in Abschnitten 5 und 7.1 besprochen und im Bereich von Stichprobenvergleichen (längs- oder querschnittlich) bei Lehramtsstudierenden der Naturwissenschaften verortet.

In dieser Stichprobe lagen bereits ausreichend Daten für den linearen Test vor, um ihn in Bezug auf seine Effizienz beurteilen zu können. Aus diesem Grund, gekoppelt mit den tendenziell geringen Studierendenzahlen in der

ausgewählten Population, wurde mit dem linearen Test keine erneute Messung durchgeführt. Es wurde stattdessen entschieden, alle erreichbaren Proband*innen für die Erprobung und Beurteilung des MST einzusetzen. Andernfalls hätte die Stichprobe geteilt werden müssen.

Für diese Entscheidung ist noch einmal wichtig zu bemerken, dass keine inhaltlichen Rückschlüsse in Bezug auf einzelne Items oder Proband*innen aus der Studie gezogen werden sollten. Es ging allein um die Feststellung der realen Messgenauigkeiten, weshalb diese Entkopplung der Stichproben als legitim betrachtet wurde. Das Vorgehen wäre nicht möglich gewesen, wenn Vergleiche zwischen den beiden Stichproben das Ziel gewesen wären.

8.2.1 Proband*innenauswahl

Obwohl im Vorfeld entschieden wurde, die Stichprobe nicht weiter als nötig aufzuteilen, konnten von der eigentlich anvisierten Gruppe der Lehramtsstudierenden in Naturwissenschaften nicht genug Proband*innen gewonnen werden. Zudem gab es bei praktisch allen in Frage kommenden Proband*innen an den teilnehmenden Standorten das Problem, dass die Items des Instruments durch vorige Befragungen in den Längsschnitterhebungen des Projekts ValiDiS bekannt waren. Es bestand damit die Gefahr von Erinnerungseffekten.

Durch die Zusammenarbeit mit dem Fachbereich *Sachunterricht und seine Didaktik* der FU Berlin konnte jedoch auf eine alternative Stichprobe zurückgegriffen werden: Studierende des Sachunterrichts im Grundschullehramt arbeiten ebenfalls mit naturwissenschaftlichen Inhalten, sofern sie hier ihren Studienschwerpunkt setzen. Somit kommen sie grundsätzlich auch für den Testeinsatz in Frage und wurden bereits in früheren Studien mit dem Instrument untersucht (Straube, 2016). Hinzu kam, dass das lineare Testinstrument zwar grundsätzlich in der Population schon eingesetzt wurde, seitdem aber bereits mehrere Semester vergangen waren. Es bestanden somit keine Bedenken, dass eine Proband*innengruppe aus diesem Bereich die Items wiedererkennen und aus dem Gedächtnis heraus beantworten würde.

So wurde die Befragung einer Erstsemesterkohorte dieser Studentengruppe (Sachunterricht mit Schwerpunkt Naturwissenschaften) an der FU Berlin ermöglicht. Die Befragung fand zum Ende des Wintersemesters 2018/19, im ersten Quartal 2019 statt. In der Stichprobe befanden sich N = 283 Proband*innen.

8.2.2 Methodik

Die Pilotierung wurde am Standort in Kleingruppen von maximal 30 Personen durchgeführt. Sie fand unter Aufsicht von Testleiter*innen statt, die mit Befragungszweck und Messinstrument vertraut waren. Vor Beginn der einzelnen Erhebungen wurde jeweils explizit darauf aufmerksam gemacht, dass es sich um ein adaptives Testinstrument handelt. Dieses Vorgehen wurde gewählt, da

- entgegen üblicher Befragungsformate am Standort keine Antwortkorrektur möglich war und
- die Anpassung der Aufgabenschwierigkeiten zu Motivationsverlusten führen kann (Frey et al., 2009). Durch die explizite Information sollte der Effekt im Vorfeld erläutert und durch Bewusstmachung minimiert werden.

Vor der eigentlichen Auswertung der Daten erfolgte eine Betrachtung der ebenfalls erhobenen Bearbeitungszeiten aller Aufgaben. Ziel war die Identifikation von solchen Aufgaben, die sich durch auffällig kurze Bearbeitung auszeichneten und von Personen, die systematisch signifikant schnellere Aufgabenbearbeitungen aufwiesen als der Rest der Stichprobe. Zu diesem Zweck wurden die Zeiten z-standardisiert. Bearbeitungszeiten, die zwei Standardabweichungen schneller als das Mittel der jeweiligen Aufgabe waren, wurden als Rateverhalten interpretiert.

8.2.3 Ergebnisse und Diskussion

Bei fünf der 283 Personen wurde für mehreren Antworten (n>5) eine auffällig kurze Bearbeitungszeit festgestellt. In diesen Fällen wurde das Antwortmuster als durchgängiges Rateverhalten interpretiert, weshalb die betreffenden Proband*innen für alle weiteren Auswertungen ausgeschlossen

wurden. Auffällige Items wurden nicht identifiziert. Die Bearbeitungszeit des Tests lag im Mittel bei 22 Minuten (mit einer Standardabweichung von sechs Minuten, siehe Abbildung 26).

Bei der Messung wurde eine Messgenauigkeit von .62 (EAP/PV-Reliabilität) erreicht. Sie entspricht damit der korrigierten Prognose aus den Simulationsstudien in Abschnitt 7.4.

Die Personenfähigkeiten innerhalb der Stichprobe waren mit einer Standardabweichung von 0.62. um einen Wert von -0.7 normalverteilt. Damit ist die Stichprobe im Vergleich zur bisherigen Gesamtstichprobe des Instruments weniger leistungsstark und enger gruppiert.

Dieses Ergebnis verwundert nicht weiter, wenn man die Auswahl der Proband*innen vergleicht. Zum einen handelt es sich bei der zuletzt gemessenen Gruppe um Studierende im ersten Semester des Grundschullehramts. Straube (2016) zeigte bereits in vergleichenden Studien, dass a) Studierende des Grundschullehramts weniger leistungsstark als Monostudierende oder Lehramtsstudierende mit einem naturwissenschaftlichen Wahlfach sind und b) in allen betrachteten Gruppen eine Leistungszunahme mit höherer Semesterzahl beobachtet werden kann. Die für die Modellberechnungen herangezogene Gruppe setzte sich aus solchen Studierenden aller Semester zusammen. Letzteres begründet auch, wieso die Fähigkeitsverteilung bei dieser Gruppe breiter ist als in der Stichprobe der Pilotierung, die nur aus einem Fachbereich und einem Semester gezogen wurde.

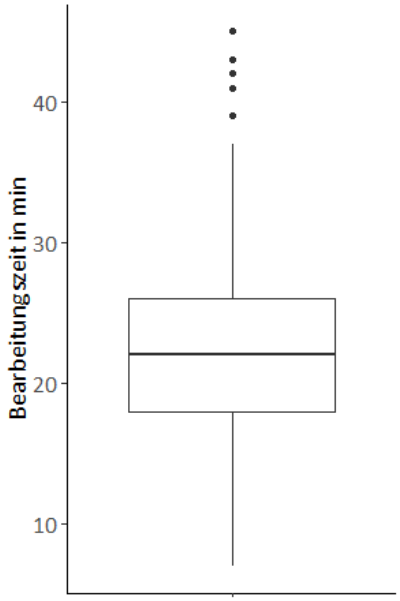

Abbildung 26: Bearbeitungszeiten des MST. Ausreißer sind als Punkt eingezeichnet. Die mittlere Bearbeitungszeit lag bei 22 Minuten.

8.3 Vergleich zwischen MST und linearem Format

Für den Vergleich der beiden Formate sind zwei Kennwerte relevant, die Testlänge und die erreichte Messgenauigkeit.

Für das lineare, papierbasierte Instrument wurden stets Bearbeitungszeiten von 45 Minuten angelegt und häufig auch vollständig benötigt, in Einzelfällen war die Befragung früher abgeschlossen (im kürzesten beobachteten Fall nach 35 Minuten). Diese Werte sind allerdings rein ‚anekdotisch' und damit möglicherweise verfälscht sowie personenbezogen. Sie entsprechen also dem Zeitpunkt, zu dem bei dem linearen Test jeweils die schnellsten oder langsamsten Proband*innen fertig waren, sie stellen nicht das Mittel der Bearbeitungszeit dar. Für den Vergleich beider Formate wurde daher die mittlere ‚anekdotische' Zeitangabe (40 min) abgeglichen mit der Zeitmarke im adaptiven Test, zu dem die Mehrheit der Stichprobe fertig war: 28 Minuten (eine Standardabweichung später als der Mittelwert)[20]. Die Bearbeitungszeit wurde durch die neue Testversion also um etwa 30% reduziert. Das Ergebnis deckt sich mit der Reduzierung der pro Person bearbeiteten Items von 21 im linearen Test zu 15 im MST.

Tabelle 9: Vergleichswerte für linearen Test und MST. Die Kennwerte für den linearen Test stammen aus den Modellrechnungen aus Kapitel 6, die Werte des MST aus der Pilotierungsstudie.

Testformat	**Linear**	**MST**
Länge in Items	21	15
EAP/PV Reliabilität	.573	.62
Effizienz	.027	.041
Effizienz %	100	153

Um Forschungsfrage 3 eindeutig beantworten zu können, wurde die Messgenauigkeit neben der direkten Betrachtung der Mittelwerte auch ein Vergleich zwischen den Verteilungen durchgeführt. Da die

[20] Es könnte argumentiert werden, dass in beiden Fällen der Mittelwert gewählt werden sollte, also 22 Minuten für den MST. Da der Wert des linearen Tests aber ‚anekdotisch' und möglicherweise zu hoch ist, wurde der strengere Vergleichswert für den MST verwendet.

Varianzhomogenität nicht erfüllt ist, wurde auf einen einseitigen *Mann-Whitney-U-Test* (Field, 2011) zurückgegriffen. Er hat auch den Vorteil, dass es sich um einen Rangsummentest handelt, aus diesem Grund ist die Teststatistik für Messgenauigkeit und Effizienz dieselbe.[21]

Die Teststatistik war signifikant mit $p < .001$, die geprüfte Gleichheitshypothese muss daher verworfen werden. Somit wird die Gegenhypothese akzeptiert und angenommen, dass sich die Messgenauigkeit und Effizienz in beiden Stichproben signifikant unterscheidet.

FF 3 kann also positiv beantwortet werden: Der MST ist signifikant effizienter als das bisherige lineare Format. Die Messgenauigkeit wurde in der Pilotierung um 8% gesteigert, die Länge des Instruments gleichzeitig um 30% verringert. Die Effizienzsteigerung beträgt 53%. Anders ausgedrückt: Durch das neue Format wird pro Item 53% mehr Information über die Proband*innen gewonnen.

[21] Beide Werte stehen in einem homogenen linearen Zusammenhang, womit sich durch Umrechnung keine Ränge verändern.

9 Zusammenfassung und Gesamtdiskussion

In dieser Arbeit wurde ein adaptiver Multistage-Test zur Messung der Kompetenz naturwissenschaftlichen Denkens entwickelt und erprobt. Als Ausgangslage wurden der vorhandene Itempool aus dem linearen Ko-WADiS-Test sowie bisher damit erhobene Daten verwendet.

In den Kapiteln 2 bis 4 wurde die theoretische Rahmung für diese Testkonstruktion gegeben. Hier wurden das Instrument sowie die relevanten mathematischen und testtheoretischen Konzepte vorgestellt.

In Kapitel 5 wurde aufbauend auf den vorherigen Ausführungen das Problemfeld der Arbeit konkretisiert. Dabei erfolgte auch die Formulierung der Forschungsfragen; hier noch einmal zur Übersicht aufgelistet:

FF 1: Welches der ausgewählten IRT-Modelle ist für die Beschreibung der Ko-WADiS Daten am besten geeignet?

FF 2: Welches adaptive Testverfahren eignet sich am besten für die Effizienzsteigerung des Ko-WADiS-Tests?

FF 2.1: Ist ein CAT oder ein MST zur Effizienzsteigerung des Ko-WADiS-Tests besser geeignet?

FF 2.2: Welcher Testalgorithmus verspricht die größte Effizienzsteigerung bei der Messung mit einem Ko-WADiS-CAT?

FF 2.3: Welche Teststruktur und Routing-Regeln versprechen die größte Effizienzsteigerung bei der Messung mit einem Ko-WADiS-MST?

FF 3: Ist die adaptive Version des Testinstruments signifikant effizienter als die lineare Version?

9.1 Modellauswahl

FF 1 wurde in Kapitel 6 beantwortet. Zu diesem Zweck wurde als erstes eine theoriegeleitete Auswahl und Bereinigung des vorhandenen Datensatzes vorgenommen (Abschnitt 6.1). Ziel war es, durch eine strenge Vorauswahl mögliche Fehlerquellen bei der Modellberechnung auszuschließen. Darunter fielen:

1. Verzerrungen der Kompetenzverteilung in der Proband*innengruppe durch die Kombination verschiedener Studiengruppen
2. Items oder Proband*innen, deren Modellparameter aufgrund geringer verfügbarer Datenmengen nur unscharf eingeschätzt werden können
3. Im Datensatz enthaltene Antworten, die durch Rateverhalten entstanden

Zusätzlich wurde die Auswahl der Stichprobe im Hinblick auf die spätere Pilotierung des adaptiven Tests und die hierfür relevante Studierendengruppe durchgeführt[22]. Insgesamt führte diese erste Einschränkung des Datensatzes zu einer Verminderung der Stichprobengröße von N = 8873 zu N = 4073. Den größten Einfluss hatte hierbei die Auswahl der Proband*innen nach ihrem Studienfach.

Im Anschluss wurden die drei ausgewählten logistischen IRT-Modelle berechnet und durch weiteren Ausschluss schlecht fittender Personen und Items optimiert (Abschnitt 6.4.1 bis 6.4.3). Die verwendeten Kriterien waren Infit, Outfit, RMSD sowie die Passung von empirischer und modellierter ICC. Bei 1pl- und 2pl-Modell verblieben nach der Optimierung des individuellen Modellfits ungefähr 70% der Proband*innen sowie die Hälfte der Items. Das 3pl-Modell konnte nicht erfolgreich berechnet werden, da der verwendete numerische Algorithmus trotz Anpassung der Startparameter nicht erfolgreich konvergierte.

Der Vergleich der angepassten Modelle untereinander (Abschnitt 6.4.4) wurde mittels AIC, BIC und der Menge der gewonnenen Testinformation durchgeführt. Bezogen auf die Fitmaße AIC und BIC unterschieden sich die Modelle nicht signifikant, weshalb im Endeffekt die höhere

[22] Die gezielte Konstruktion eines adaptiven Tests für die Verteilung der zu messenden Stichprobe soll die Messgenauigkeit erhöhen. Ist die Verteilung im Voraus bekannt, kann sie in gleich große Fähigkeitsbereiche aufgeteilt werden, für die dann jeweils Module des MST erstellt werden. Durch eine genauere Kenntnis der Stichprobe kann so die Passung zwischen Items und Proband*innen erhöht und damit der Informationsgewinn maximiert werden (siehe Abschnitte 3.3.2 und 4.3). Bei der Auswahl der Daten wurde nur eine einzelne Studierendengruppe anvisiert, damit die Verteilung möglichst gut beschrieben werden konnte.

Messgenauigkeit bei der Datenbeschreibung zur Auswahl des 2pl-Modells führte. FF 1 wurde also damit beantwortet, dass das 2pl-Modell von den praktisch verwendbaren Alternativen den ausgesuchten Datensatz am besten beschreibt.

Ein möglicher Kritikpunkt am gewählten Vorgehen liegt in der großen Datenmenge, die ausgeschlossen wurde. Die Gründe und Kriterien für diesen Ausschluss wurden in den vorigen Absätzen genannt und stellen prinzipiell keine Besonderheit dar. Auf Personenseite haben die Einschränkungen auch vermutlich keine negativen Auswirkungen auf die Testkonstruktion gehabt. Nach der Auswahl des 2pl-Modells verblieben immer noch 2485 Proband*innen mit normalverteilten Fähigkeitswerten. Zudem wurde gerade durch die strenge Auswahl dafür gesorgt, dass für jede einzelne Person ein notwendiges Mindestmaß an nicht erratenen Antworten vorlag. Damit kann die Güte der ermittelten Parameter, die für die Testkonstruktion relevant waren, als hoch angesehen werden.

Jedoch hätte beispielsweise durch das Anlegen weniger strenger Grenzwerte für die Fitmaße ein größerer Teil des Itempools bewahrt werden können. Die geringe Menge an Items im Bereich kleiner Schwierigkeiten (siehe Abbildung 17) stellte bei den in Kapitel 7 verfügbaren Alternativen der Teststruktur ein Problem dar, weil die Gesamttestlänge von 15 Items nicht überschritten werden konnte – es waren einfach nicht genug Items im Pool vorhanden, um alle Schwierigkeitsbereiche eines längeren Testinstruments füllen zu können. Eines der Resultate der stark begrenzen Gesamtlänge war damit die schlussendlich erreichte Messgenauigkeit von Rel(EAP) = .62 durch den adaptiven Test.

Dieser Umstand wurde in Kauf genommen, da die Testlänge im Vergleich zum 21 Items langen linearen Instrument sowieso verringert werden sollte. Zudem war das Hauptziel der Testkonstruktion die Steigerung der Testeffizienz, nicht das Erreichen eines bestimmten Reliabilitätswertes durch Anpassung der Testlänge. Besonders diese Messgenauigkeit steht aber oft im Fokus von Testkonstruktion und -einsatz. Im Sinne einer genaueren Messung kann die Notwendigkeit von höheren Reliabilitäten durchaus argumentiert werden. Sollte bei der Planung einer späteren Testanwendung

so eine Notwendigkeit festgestellt werden, müsste also eine erneute Modellberechnung mit weniger strengem Itemausschluss getestet werden. Garantiert wäre der Erfolg so eines Vorgehens aber nicht, da geringere Anforderungen an die Modellpassung natürlich wieder die Wahrscheinlichkeit von Fehleinschätzungen in Bezug auf Personen und Items erhöhen.

9.2 Testkonstruktion und Simulationsstudien

Nach der Modellberechnung wurde FF 2 in Kapitel 7 bearbeitet. Als erste Teilfrage musste hierbei FF 2.1 beantwortet, also die geeignetere Teststruktur ausgesucht werden. Nachdem durch die Feststellung der tendenziell geringen Messgenauigkeiten ein Einsatz des Instruments in der Individualdiagnostik ausgeschlossen wurde, blieb als entscheidendes Kriterium die inhaltliche Validität der Messwertinterpretation. MSTs haben in dieser Hinsicht den Vorteil, dass durch die gezielte Kombination von verschiedenen Itemkontexten und Kompetenzfacetten innerhalb der einzelnen Module eine gewisse Breite der beobachteten Kompetenz garantiert werden kann. Daher wurde als Antwort auf FF 2.1 der Konstruktion eines MST Vorzug gegenüber dem CAT gegeben. FF 2.2 wurde als nicht mehr relevant beurteilt und aus der weiteren Bearbeitung ausgeschlossen.

Zu FF 2.3 wurden als erstes die Regeln für Scoring, Routing und Modulzusammensatzung des MST erstellt (Abschnitt 7.2 und 7.3). Sie folgten direkt den theoretischen Grundlagen aus Kapitel 4 sowie den praktischen Notwendigkeiten, die sich aus den Einschränkungen des Itempools einerseits und der verlangten inhaltlichen Breite der Module andererseits ergaben.

Anhand dieser Regeln wurden dann verschiedene Teststrukturen und -Längen erstellt und in einer Simulationsstudie verglichen (Abschnitt 7.4). Für jede einzelne Testvariante wurde eine eigene virtuelle Stichprobe erstellt. Dazu wurde die im reduzierten Ko-WADiS Datensatz vorhandene Stichprobe anhand ihrer Parameter und des verwendeten 2pl-Modells vollständig imputiert. Damit konnten für die 25 betrachteten Testvarianten ausreichend große und unabhängige Stichproben generiert werden, die trotzdem alle die gleiche Merkmalsverteilung in Bezug auf die zu messende Fähigkeit aufwiesen. Als Beurteilungsmaß der Varianten wurden

die EAP/PV-Reliabilitäten verwendet, die in den simulierten Messungen erreicht wurden.

Als Ergebnis dieses Vergleichs fanden sich mehrere Trends in den Daten: Die kürzesten Tests zeichneten sich als die effizientesten aus, die längsten Tests erzielten die größten Messgenauigkeiten, und die Teststruktur 1-2-2 mit zwei Schwierigkeitsbereichen sowie drei Stufen war tendenziell am genauesten und effizientesten.

Um in der späteren Pilotierung ein Mindestmaß an Messgenauigkeit zu wahren, wurde die messgenaueste Testvariante für die praktische Umsetzung des adaptiven Tests ausgewählt. Dabei handelte es sich um den Test mit 1-2-2 Struktur und einer Länge von 15 Items.

Streng genommen wäre die Antwort auf FF 2.3 der nur acht Items lange Test in Struktur 1-2 gewesen, weil er die höchste Effizienz aufwies. Bei einer Messgenauigkeit von nur .5 (EAP/PV) wäre der Testeinsatz aber nicht zu rechtfertigen gewesen.

Die Güte der Simulationen sowie ihre Vorhersagekraft wurden überprüft, indem bereits im Projekt Ko-WADiS verwendete lineare Testhefte ebenfalls simuliert wurden. Für diese wurde das gleiche Vorgehen gewählt, der Testalgorithmus wurde lediglich durch die starre Aneinanderreihung der Items (in der gleichen Zusammenstellung und Reihenfolge wie in den Testheften) ausgetauscht. Die in realen Messungen erreichten Reliabilitäten wurden durch die Simulation im Mittel um einen Wert von 0.02 überschätzt. Was ihre absolute sowie relative Größe betrifft, wurde diese Abweichung als akzeptabel eingeschätzt. Zurückzuführen ist sie vermutlich auf den für die Simulationen vorausgesetzten perfekten Modellfit. Dieser ist nicht realistisch und führt zu einem besser vorhersagbaren Verhalten der virtuellen Proband*innen.

Bei der Auswahl der Teststruktur zeigte sich erneut die in Abschnitt 9.1 erwähnte Problematik. Es kann nicht ausgeschlossen werden, dass bei größeren Testlängen ein anderes Bild entstanden und damit auch eine andere Teststruktur bevorzugt worden wäre, bei der durch eine insgesamt höhere Messgenauigkeit auch eine Einteilung in mehr als nur zwei

Schwierigkeitsniveaus funktioniert hätte. Hierbei handelt es sich allerdings weniger um ein Problem innerhalb dieser Arbeit. Vielmehr stellt sich damit die Frage, ob im Vergleich von längeren Teststrukturen eine andere und allgemeingültige Aussage zum optimalen Testaufbau entstehen würde. So wäre denkbar, dass mit steigender Testlänge Grenzwerte in der Messgenauigkeit beobachtet werden, nach denen eine feinere Aufspaltung des Instruments besser funktioniert.

Im Hinblick auf die in Kapitel 4 zitierten Quellen ist jedoch nicht von der Allgemeingültigkeit solcher Beobachtungen auszugehen, jedenfalls nicht, wenn über die Grenzen eines einzelnen Testinstruments hinaus abstrahiert werden soll. Wie viel Information über eine Person gewonnen werden kann, hängt schlussendlich von den Items eines Testinstruments ab. Somit ist auch von der Gestaltung und Güte der Items abhängig, nach wie vielen Antworten eine sichere Einteilung von Proband*innen in zwei, drei oder mehr Fähigkeitsniveaus praktikabel ist.

Auch wenn diese sehr akademische Frage nach der (von der Testlänge unabhängigen) perfekten Struktur eines MST mit Ko-WADiS Items unbeantwortet bleiben musste, konnte aber wenigstens die beste praktisch umsetzbare Version bestimmt werden.

9.3 Pilotierung

Als letzter Schritt wurde FF 3 in Kapitel 8 beantwortet. Als Vorgehen wurde ein Vergleich der Messgenauigkeit und Effizienz von beiden vorliegenden Testversionen, linear und adaptiv, in einer bereits bekannten Stichprobe gewählt.

Grundsätzlich wäre für den Vergleich von mehreren Messinstrumenten eine parallele Befragung in repräsentativen Teilstichproben aus derselben Population nötig. Für den linearen Test lagen für die Zielpopulation aber schon ausreichend Daten vor, um den Test in Bezug auf die dort erreichten Werte sicher zu beurteilen (siehe Abschnitt 6.1). Somit wurde entschieden, nur den neuen MST anzuwenden, um die hierfür fehlenden Testwerte zu bestimmen.

In der angepeilten Gruppe der Lehramtsstudierenden in Naturwissenschaften konnten jedoch für die Pilotierung des MST nicht genug Proband*innen gewonnen werden. Der Grund hierfür war eine Kombination aus geringen Studierendenzahlen sowie einer starken Belastung der Studierenden durch andere Befragungen.

Um den Test dennoch in einer relevanten Gruppe mit einer realen Befragung beurteilen zu können, wurde auf eine Erstsemesterkohorte von Studierenden des Grundschullehramts ausgewichen. In dieser Gruppe wurden Proband*innen des Studienfachs *Sachunterricht mit Schwerpunkt Naturwissenschaften* befragt. Diese Studierenden beschäftigten sich ebenso wie die Student*innen aus der ursprünglich anvisierten Population in ihrem Studium mit naturwissenschaftlichen Inhalten. Zudem waren Studierende aus diesem Fachbereich bereits zuvor mit dem Ko-WADiS-Test befragt und die Testwertinterpretation auch bei ihnen als valide beurteilt worden (Straube, 2016). Die für die Pilotierung verfügbare Stichprobe umfasste N = 283 Personen.

Im realen Einsatz erreichte der MST eine Messgenauigkeit von Rel(EAP) = .62. Diese lag zwar nur um 8% über der Genauigkeit, die der lineare Test im Ko-WADiS Datensatz aufwies, allerdings war die Testlänge dabei um 30% geringer. Das bedeutet, dass der Informationsgewinn pro Item allein durch die neue Art der Testanwendung um 53% gesteigert werden konnte.

Zur Beantwortung von FF 3 wurde schließlich noch ein Mann-Whitney-U-Test durchgeführt. Die Teststatistik fiel signifikant aus ($p < .001$), weshalb der MST als signifikant effizienter angenommen und FF 3 positiv beantwortet wurde.

Die Aussagekraft dieses Ergebnisses muss aber möglicherweise eingeschränkt werden: die beiden Testinstrumente wurden nicht anhand derselben Stichprobe verglichen. Da die befragten Personen aber direkte Auswirkungen auf die Performanz des Messinstruments haben, kann eine Verzerrung des Vergleichs zugunsten einer der beiden Testvarianten nicht ausgeschlossen werden.

Die Testleistungen in der Stichprobe der Pilotierung lagen unterhalb der aus den Ko-WADiS Studien. Der MST wurde allerdings auf die zweite dieser beiden Gruppen ausgerichtet und konstruiert. Somit ist festzustellen, dass die Fähigkeitsverteilung der tatsächlich gemessenen Proband*innen von der bei der Testkonstruktion angenommenen Verteilung abweicht. Die Items und Module, aus denen der MST aufgebaut wurde, decken also die Stichprobe nicht ideal ab.

Die gewonnene Information einer Messung und Auswertung durch IRT-Verfahren und damit auch adaptiver Algorithmen hängt aber direkt von der Passung zwischen Item- und Personenparametern ab (vgl. Abschnitt 3.3.2 und 4.3.3). Es kann damit die Hypothese aufgestellt werden, dass die schlechte Passung zwischen Aufbau des MST und der Fähigkeitsverteilung der Studierenden im Grundschullehramt die Messgenauigkeit reduziert hat. Durch die vorhandenen Daten kann diese Hypothese leider nicht ausreichend geprüft werden. Weiterhin kann damit nicht untersucht werden, welchen Einfluss auf die Resultate weitere Parameter haben, wie beispielsweise die unterschiedlich breite Streuung der Fähigkeiten in beiden Stichproben.

Zusammenfassend ist die Vermutung der Ergebnisverzerrung also plausibel, da mehrere mögliche Einflussfaktoren nicht überprüft wurden. Aus Sicht der mathematischen Theorie kann vermutet werden, dass der Unterschied zwischen beiden Formaten bei einem direkten Vergleich sogar größer ausfallen würde.

Unabhängig von der zukünftigen Performanz des Instruments kann zum aktuellen Zeitpunkt festgehalten werden, dass durch das gewählte Vorgehen in der Pilotierung eine signifikante Effizienzsteigerung erzielt werden konnte.

9.4 Ausblick

Um eine endgültig sichere Aussage über den Vergleich zwischen beiden Testversionen treffen zu können, wäre die Durchführung einer weiteren Vergleichsmessung in der Zukunft notwendig. In der Zwischenzeit kann aber von einer signifikant höheren Effizienz des MST und damit einer

erfolgreichen Umsetzung ausgegangen werden. Es stellt sich damit wieder die Frage aus der Einleitung: Wie hoch ist der Aufwand zur Konstruktion eines adaptiven Tests, und ist er gerechtfertigt?

Der Hauptaufwand bei der Umsetzung war die testtheoretische Fundierung des neuen Instruments und die Suche nach einer passenden Teststruktur. Aus rein praktischer Sicht betrachtet, erscheint der Aufwand für die Konstruktion eines MST nicht hoch, sofern ein geeigneter Itempool vorhanden ist. Davon ausgehend müssen zwar die Regeln für den Aufbau des Tests und den Testalgorithmus festgelegt werden, hierfür liefern Literatur und die praktischen Umstände eines Testeinsatzes aber sehr klare Entscheidungen (vgl. Abschnitte 7.2 und 7.3).

Die Wahl der Teststruktur war in dieser Arbeit mit einem hohen Aufwand verbunden. Bei Betrachtung der Resultate (Abschnitt 7.4.2) kann aber diskutiert werden, ob in jedem Fall so umfangreiche Vergleichsstudien notwendig wären. Bei gleicher Testlänge lagen die Unterschiede der Reliabilität zwischen den geprüften Teststrukturen im Bereich von 0.05. Für den Ko-WADiS-Test und andere Instrumente, die eine eher geringe Messgenauigkeit aufweisen, kann das bereits eine Rolle spielen. Sofern aber auf einem genaueren Instrument aufgebaut würde, könnte wahrscheinlich die Auswahl einer der getesteten Strukturen nach rein inhaltlichen Argumenten riskiert werden, ohne ähnliche Simulationen vornehmen zu müssen.

Somit scheint die einzige Hürde das Vorhandensein des Itempools zu sein. Dieser muss a) umfangreich genug für den Aufbau eines MST sein, b) Items enthalten, die den Fähigkeitsbereich der Zielgruppe gut abdecken und c) IRT-konform sein.

Was den Umfang betrifft, so enthält der in dieser Arbeit erstellte MST nicht wesentlich mehr Items als die bereits zuvor benutzten linearen Testhefte. Tatsächlich enthalten diese jeweils 21 Items. Zählt man alle Items aus den Modulen des neuen 1-2-2 MST zusammen, sind es 25. Bei der Wahl einer geringen Aufspaltung steigen also die Anforderungen an den Umfang des Itempools nicht wesentlich.

Die Abdeckung der Fähigkeitsverteilung in der Zielgruppe ist ebenfalls keine Anforderung, die nicht schon bei einer linearen Messung erfüllt werden müsste. Auch hier ist es wichtig, dass zur genauen Diagnose der Proband*innen für alle betrachteten Inhalte sowohl leichte als auch schwere Items vorhanden sind. Nicht umsonst werden auch bei der Testkonstruktion mit klassischen Modellen (KTT) Schwierigkeitsanalysen vorgenommen (Moosbrugger & Kelava, 2012). Besondere Relevanz haben diese Betrachtungen bei der Modellierung von Niveaustufen, da für die unterschiedlichen Leistungsniveaus jeweils Items mit entsprechenden Schwierigkeitsmerkmalen vorhanden sein müssen.

Der letzte Punkt, die Erfüllung der Anforderung der IRT durch die Items im Testinstrument, ist weniger offensichtlich zu beurteilen. Diese Anforderungen werden oft als streng beschrieben, vor allem im Vergleich zur KTT, nach der die meisten aktuellen Testinstrumente im didaktischen Bereich entwickelt wurden (vgl. Abschnitt 3.2). Tatsächlich kann aber auch hier gezeigt werden, dass es sich im Grund um keine größere Hürde handelt als bei jeder anderen Testkonstruktion. Wright (1989) zeigt, dass die Verwendung von Rohwerten (die Summe der korrekten Antworten einer Person) informationstechnisch äquivalent ist zur Verwendung von Personenschätzern aus einem berechneten 1pl- oder Raschmodell. Sofern diese Rohwerte also als Maß der Personenfähigkeit akzeptiert werden, was in der KTT und zugehörigen Instrumenten meist der Fall ist, kann mit ihnen auch eine adaptive Messung wie mit einem 1pl-Modell vorgenommen werden.

Zusammengefasst erfüllen also vermutlich die meisten linearen Testinstrumente bereits die Anforderungen für die Anpassung zu einer einfachen adaptiven Testvariante, wie sie in Kapitel 7 und 8 dargestellt wurde.

Selbstverständlich sind nicht alle Instrumente dazu geeignet, in einem adaptiven Format eingesetzt zu werden. Je breiter beispielsweise die interne Struktur eines Konstrukts aufgestellt ist, desto schwerer wird die valide Auslegung von Testwerten bei einer weiteren Aufspaltung des eingesetzten Itempools.

Dennoch wäre es wohl lohnenswert, über die Weiterentwicklung bereits etablierter Instrumente nachzudenken, um die Belastung von Proband*innen möglichst gering zu halten. Das betrifft besonders zeitaufwändige Leistungstests und all diejenigen Instrumente, die im Rahmen aktueller Forschung zur Erhebung von Kovariaten (wie Fachwissen) neben den eigentlich fokussierten Tests verwendet werden.

Literaturverzeichnis

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Al-Amri, S. (2007). *Computer-based vs. Paper-based Testing: Does the test administration mode matter?*, University of Essex. Zugriff am 23.01.2020. Verfügbar unter https://www.semanticscholar.org/paper/Computer-based-vs.-Paper-based-Testing%3A-Does-the-Alamri/86dfc22d8a8b5ff19325b6c991ab37e43ab9801d

Amelang, M. & Zielinski, W. (1997). *Psychologische Diagnostik und Intervention* (Springer-Lehrbuch, 2., korrigierte, aktualisierte und überarbeitete Auflage). Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-662-22370-3

Amt für Statistik Berlin-Brandenburg. (2019, Juli). *Studierende an Hochschulen im Land Berlin Wintersemester 2018/2019. Teil 2: Ausführliche Ergebnisse (Endgültige Angaben)*. Berlin. Zugriff am 29.06.2020.

Andersen, E. B. (1970). Asymptotic Properties of Conditional Maximum-Likelihood Estimators. *Journal of the Royal Statistical Society: Series B (Methodological)*, *32*(2), 283–301. https://doi.org/10.1111/j.2517-6161.1970.tb00842.x

Armstrong, R. D., Jones, D. H., Koppel, N. B. & Pashley, P. J. (2004). Computerized Adaptive Testing With Multiple-Form Structures. *Applied Psychological Measurement*, *28*(3), 147–164. https://doi.org/10.1177/0146621604263652

Army Marketing and Research Group; The United States Army. (2020, 14. Januar). *Understanding the ASVAB Test and Your Scores*, Army Marketing and Research Group; The United States Army. Zugriff am 21.01.2020. Verfügbar unter https://www.goarmy.com/learn/understanding-the-asvab.html

Ayala, R. J. de & Kenny, D. A. (2009). *The theory and practice of item response theory* (Methodology in the social sciences). New York: Guilford Press. Retrieved from http://site.ebrary.com/lib/alltitles/docDetail.action?docID=10468533

Baker, F. B. & Kim, S.-H. (2017). *The Basics of Item Response Theory Using R* (Statistics for Social and Behavioral Sciences). Cham: Springer. https://doi.org/10.1007/978-3-319-54205-8

Behnke, J. & Behnke, N. (Hrsg.). (2006). *Grundlagen der statistischen Datenanalyse. Eine Einführung für Politikwissenschaftler* (Grundwissen Politik, Bd. 41, 1. Aufl.). Wiesbaden: VS Verlag für Sozialwissenschaften | GWV Fachverlage GmbH Wiesbaden. https://doi.org/10.1007/978-3-531-90003-2

Bergstrom, B. A., Lunz, M. E. & Gershon, R. C. (1992). Altering the Level of Difficulty in Computer Adaptive Testing. *Applied Measurement in Education*, *5*(2), 137–149. https://doi.org/10.1207/s15324818ame0502_4

Betz, N. E. & Weiss, D. J. (1976, Juni). *Psychological Effects of Immediate Knowledge of Results and Adaptive Ability Testing. Research Report 76-4.* Minnesota University, Minneapolis Departement of Psychology. Zugriff am 03.02.2020.

Birnbaum, A. (2008). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord, M. R. Novick & A. Birnbaum (Eds.), *Statistical theories of mental test scores* (Behavioral science quantitative methods). Charlotte, NC: Information Age Publ.

Blömeke, S. (Ed.). (2008). *Professionelle Kompetenz angehender Lehrerinnen und Lehrer. Wissen, Überzeugungen und Lerngelegenheiten deutscher Mathematikstudierender und -referendare; erste Ergebnisse zur Wirksamkeit der Lehrerausbildung*. Münster: Waxmann.

Boake, C. (2002). From the Binet-Simon to the Wechsler-Bellevue: tracing the history of intelligence testing. *Journal of Clinical and Experimental Neuropsychology*, *24*(3), 383–405. https://doi.org/10.1076/jcen.24.3.383.981

Boone, W. J., Staver, J. R. & Yale, M. S. (2014). *Rasch analysis in the human sciences*. Dordrecht: Springer. https://doi.org/10.1007/978-94-007-6857-4

Bortz, J. & Döring, N. (2006). *Forschungsmethoden und Evaluation. Für Human- und Sozialwissenschaftler; mit 87 Tabellen* (Springer-Lehrbuch Bachelor, Master, 4., überarb. Aufl., [Nachdr.]. Heidelberg: Springer-Medizin-Verl. Zugriff am 26.08.2019.

Bosbach, E. (2007). *Neue Texte und Hilfestellungen zur Umsetzung der Ziele des Bologna-Prozesses and deutschen Hochschulen* (Beiträge zur Hochschulpolitik, 2007,5). Bonn: Hochschulrektorenkonferenz. Zugriff am 18.07.2020.

Brennan, R. L. (2010). Generalizability Theory and Classical Test Theory. *Applied Measurement in Education*, *24*(1), 1–21. https://doi.org/10.1080/08957347.2011.532417

Bühner, M. (2011). *Einführung in die Test- und Fragebogenkonstruktion* (Always learning, 3., aktualisierte und erweiterte Auflage). München: Pearson.

Bundesministerium für Bildung und Forschung. (2015). *Bericht der Bundesregierung Bericht über die Umsetzung des Bologna-Prozesses 2012 - 2015 in Deutschland*. Zugriff am 18.07.2020.

Bundesministerium für Bildung und Forschung. (2018). *Die Umsetzung der Ziele des Bologna-Prozesses 2015 – 2018*. Zugriff am 18.07.2020.

Burnham, K. P. & Anderson, D. R. (2010). *Model selection and multimodel inference. A practical information-theoretic approach* (2. ed.). New York, NY: Springer.

Chang, H.-H. (2015). Psychometrics behind Computerized Adaptive Testing. *Psychometrika*, *80*(1), 1–20. https://doi.org/10.1007/s11336-014-9401-5

Chang, H.-H. & Ying, Z. (1999). a-Stratified Multistage Computerized Adaptive Testing. *Applied Psychological Measurement*, *23*(3), 211–222. https://doi.org/10.1177/01466219922031338

Chang, S.-W. & Ansley, T. N. (2003). A Comparative Study of Item Exposure Control Methods in Computerized Adaptive Testing. *Journal of Educational Measurement*, *40*(1), 71–103. https://doi.org/10.1111/j.1745-3984.2003.tb01097.x

Cheng, P. E. & Liou, M. (2000). Estimation of Trait Level in Computerized Adaptive Testing. *Applied Psychological Measurement*, *24*(3), 257–265. https://doi.org/10.1177/01466210022031723

Christensen, K. B., Mesbah, M. & Kreiner, S. (2013). *Rasch models in health* (Applied mathematics series). London: ISTE Ltd. https://doi.org/10.1002/9781118574454

Döring, N. & Bortz, J. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften* (Springer-Lehrbuch, 5. vollständig überarbeitete, aktualisierte und erweiterte Auflage). Berlin: Springer. https://doi.org/10.1007/978-3-642-41089-5

Drasgow, F. (1989). An Evaluation of Marginal Maximum Likelihood Estimation for the Two-Parameter Logistic Model. *Applied Psychological Measurement*, *13*(1), 77–90. https://doi.org/10.1177/014662168901300108

Eggen, T. J. H. M. (2004). *Contributions to the theory and practice of computerized adaptive testing*.

Eggen, T. J. H. M. & Verschoor, A. J. (2006). Optimal Testing With Easy or Difficult Items in Computerized Adaptive Testing. *Applied Psychological Measurement*, *30*(5), 379–393. https://doi.org/10.1177/0146621606288890

Ehmke, T. (2006). Entwicklung von Testverfahren für die Bildungsstandards Mathematik. Rahmenkonzeption, Aufgabenentwicklung, Feld- und Haupttest. *Zeitschrift für Lernforschung*, *34*(3), 220–238. Zugriff am 18.07.2020.

Embretson, S. E. & Reise, S. P. (2000). *Item response theory for psychologists* (Multivariate applications book series). Mahwah, N.J.: L. Erlbaum Associates. Zugriff am 28.08.2019.

Europäische Kommission. (2019, 6. August). *The Bologna Process and the European Higher Education Area - Allgemeine und berufliche Bildung - European Commission*. Zugriff am 18.07.2020. Verfügbar unter https://ec.europa.eu/education/policies/higher-education/bologna-process-and-european-higher-education-area_de

Field, A. (2011). *Discovering statistics using SPSS. (and sex and drugs and rock 'n' roll)* (3. ed., reprinted.). Los Angeles, California: Sage. Accessed 26.08.2019.

Frey, A. (2012). Adaptives Testen. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (Springer-Lehrbuch, 2., aktualisierte und überarbeitete Auflage, S. 275–293). Berlin, Heidelberg: Springer-Verlag Berlin Heidelberg.

Frey, A. & Ehmke, T. (2008). Hypothetischer Einsatz adaptiven Testens bei der Überprüfung von Bildungsstandards. In M. Prenzel, I. Gogolin & H.-H. Krüger (Hrsg.), *Kompetenzdiagnostik* (Bd. 34, S. 169–184).

Wiesbaden: VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-90865-6_10

Frey, A., Hartig, J. & Moosbrugger, H. (2009). Effekte des adaptiven Testens auf die Motivation zur Testbearbeitung am Beispiel des Frankfurter Adaptiven Konzentrationsleistungs-Tests. *Diagnostica*, *55*(1), 20–28. https://doi.org/10.1026/0012-1924.55.1.20

Green, B. F. (2012). The Promise of Tailored Tests. *Principals of Modern Psychological Measurement: A Festschrift for Frederic M. Lord*, 69.

Haase, S., Kirstein, J. & Nordmeier, V. (2016). tet.folio: Neue Ansätze zur digitalen Unterstützung individualisierten Lernens. *PhyDid B - Didaktik der Physik - Beiträge zur DPG-Frühjahrstagung, 2016*. Zugriff am 06.07.2020.

Hambleton, R. K., Swaminathan, H. & Rogers, H. J. (1995). *Fundamentals of item response theory* (Measurement methods for the social sciences, vol. 2, 4. [ed.]. Newbury Park: Sage Publ.

Hartig, J. & Klieme, E. (2006). Kompetenz und Kompetenzdiagnostik. In K. Schweizer (Hrsg.), *Leistung und Leistungsdiagnostik. Mit 18 Tabellen* (S. 127–144). Berlin/Heidelberg: Springer-Verlag.

Hartmann, S., Mathesius, S., Stiller, J., Straube, P., Krüger, D. & Upmeier zu Belzen, A. (2015). Kompetenzen der naturwissenschaftlichen Erkenntnisgewinnung als Teil des Professionswissens zukünftiger Lehrkräfte: Das Projekt Ko-WADiS. In B. Koch-Priewe, A. Köker, J. Seifried & E. Wuttke (Hrsg.), *Kompetenzerwerb an Hochschulen: Modellierung und Messung. Zur Professionalisierung angehender Lehrerinnen und Lehrer sowie frühpädagogischer Fachkräfte* (S. 39–58). Bad Heilbrunn: Verlag Julius Klinkhardt. Zugriff am 10.10.2019.

Hartmann, S., Upmeier zu Belzen, A. & Krüger, D. (2015). *Ko-WADiS. Kompetenzmodellierung und -erfassung zum Wissenschaftsverständnis über naturwissenschaftliche Arbeits- und Denkweisen bei Studierenden (Lehramt) in den drei naturwissenschaftlichen Fächern Biologie, Chemie und Physik*. Schlussbericht zum Kooperationsprojekt innerhalb des vom BMBF geförderten wissenschaftlichen Transferprojekts „Kompetenzmodellierung und Kompetenzerfassung im Hochschulsektor“ (KoKoHs). Berlin: Humboldt-Universität Berlin. Zugriff am 10.10.2019.

Hatzinger, R. (2010, November). *Parameter Estimation in the Rasch Model. Psychometric Methods Part 3.* Psychometric Methods, Wien. Zugriff am 06.11.2019. Verfügbar unter http://statmath.wu-wien.ac.at/people/hatz/psychometrics/10w/RM_handouts_3.pdf

Hendrickson, A. (2007). An NCME Instructional Module on Multistage Testing. *Educational Measurement: Issues and Practice*, *26*(2), 44–52. https://doi.org/10.1111/j.1745-3992.2007.00093.x

Hetter, R. D. & Sympson, J. B. (1997). Item exposure control in CAT-ASVAB. In W. A. Sands, B. K. Waters & J. R. McBride (Eds.), *Computerized adaptive testing. From inquiry to operation* (1st ed., pp. 141–144). Washington, DC: American Psychological Association. https://doi.org/10.1037/10244-014

Higgins, J., Russell, M. & Hoffmann, T. (2005). Examining the Effect of Computer-Based Passage Presentation of Reading Test Performance. *The Journal of Technology, Learning and Assessment*, *3*(4). Verfügbar unter https://ejournals.bc.edu/index.php/jtla/article/view/1657

Johnson, M. S. (2007). Marginal Maximum Likelihood Estimation of Item Response Models in R. *Journal of Statistical Software*, *20*(10). https://doi.org/10.18637/jss.v020.i10

Khoshsima, H., Hashemi Toroujeni, S. M., Thompson, N. & Reza Ebrahimi, M. (2019). Computer-Based (CBT) vs. Paper-Based (PBT) Testing: Mode Effect, Relationship between Computer Familiarity, Attitudes, Aversion and Mode Preference with CBT Test Scores in an Asian Private EFL Context. *Teaching English with Technology*, *19*(1), 86–101. Zugriff am 13.12.2019.

Kim, H. & Plake, B. S. (1994). Monte Carlo simulation comparison of two-stage testing and computerized adaptive testing. *Dissertation Abstracts International Section A: Humanities & Social Sciences*, *54*(7-A), 2548.

Kingsbury, G. G. & Zara, A. R. (1989). Procedures for Selecting Items for Computerized Adaptive Tests. *Applied Measurement in Education*, *2*(4), 359–375. https://doi.org/10.1207/s15324818ame0204_6

Kingsbury, G. G. & Zara, A. R. (1991). A Comparison of Procedures for Content-Sensitive Item Selection in Computerized Adaptive Tests.

Applied Measurement in Education, *4*(3), 241–261. https://doi.org/10.1207/s15324818ame0403_4

Klieme, E. (Hrsg.). ((2010)). *Zeitschrift für Pädagogik*. Weinheim: Beltz Juventa. Zugriff am 18.07.2020.

Krüger, D., Upmeier zu Belzen, A. & Hartmann, S. (2016). ValiDiS – Kompetenzmodellierung und -erfassung: Validierungsstudie zum wissenschaftlichen Denken im naturwissenschaftlichen Studium. In O. Zlatkin-Troitschanskaia, H. A. Pant, C. Lautenbach & M. Toepper (Hrsg.), *Kompetenzmodelle und Instrumente der Kompetenzerfassung im Hochschulsektor – Validierungen und methodische Innovationen (KoKoHs): Übersicht der Forschungsprojekte* (S. 22–25). Johannes Gutenberg-Universität Mainz; Humboldt-Universität Berlin.

Krüger, D., Upmeier zu Belzen, A. & Hartmann, S. (2020). Scientific Thinking in Natural Sciences: Ko-WaDiS Test. In O. Zlatkin-Troitschanskaia, H. A. Pant, M.-T. Nagel, D. Molerov, C. Lautenbach & M. Toepper (Hrsg.), *Portfolio of KoKoHs Assessments. Test Instruments for Modeling and Measuring Domain-specific and Generic Competencies of Higher Education Students and Graduates* (S. 82–84). Mainz, Berlin.

Kullback, S. & Leibler, R. A. (1951). On Information and Sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79–86. https://doi.org/10.1214/aoms/1177729694

Kultusministerkonferenz. (2005). *Bildungsstandards der Kultusministerkonferenz. Erläuterungen zur Konzeption und Entwicklung* (Veröffentlichungen der Kultusministerkonferenz). Zugriff am 18.07.2020.

Laborda, J. G., Santiago, M. L., Juan, N. O. de & Álvarez, A. Á. (2014). Communicative Language Testing: Implications for Computer Based Language Testing in French for Specific Purposes. *Journal of Language Teaching and Research*, *5*(5). https://doi.org/10.4304/jltr.5.5.971-975

Leung, C.-K., Chang, H.-H. & Hau, K.-T. (2003). Computerized Adaptive Testing: A Comparison of Three Content Balancing Methods. *Journal of Technology, Learning, and Assessment*, *2*. Zugriff am 29.01.2020.

Linacre, J. M. (1994). Sample Size and Item Calibration or Person Measure Stability. *Rasch Measurement Transactions,* 4, S. 328. Zugriff am 26.08.2019. Verfügbar unter https://www.rasch.org/rmt/rmt74m.htm

Linacre, J. M. (2002). Expected A Posteriori (EAP) Measures. *Rasch Measurement Transactions*, *16*(3), 891. Zugriff am 25.10.2019. Verfügbar unter https://www.rasch.org/rmt/rmt163i.htm

Linacre, J. M. (2005). Rasch dichotomous model vs. One-parameter Logistic Model. *Rasch Measurement Transactions*, *19*(3), 1032. Zugriff am 30.01.2020.

Linacre, J. M. (2009). The Efficacy of Warm's Weighted Mean Likelihood Estimate (WLE) Correction to Maximum Likelihood Estimate (MLE) Bias. *Rasch Measurement Transactions*, *23*(1), 1188–1189. Zugriff am 06.11.2019. Verfügbar unter https://www.rasch.org/rmt/rmt231d.htm

Linden, W. J. & Glas, G. A.W. (Eds.). (2002). *Computerized Adaptive Testing: Theory and Practice*. Dordrecht: Kluwer Academic Publishers. https://doi.org/10.1007/0-306-47531-6

Ling, G., Attali, Y., Finn, B. & Stone, E. A. (2017). Is a Computerized Adaptive Test More Motivating Than a Fixed-Item Test? *Applied Psychological Measurement*, *41*(7), 495–511. https://doi.org/10.1177/0146621617707556

Liu, Y. & Maydeu-Olivares, A. (2013). Local Dependence Diagnostics in IRT Modeling of Binary Data. *Educational and Psychological Measurement*, *73*(2), 254–274. https://doi.org/10.1177/0013164412453841

Lord, F. M. (2012). *Applications of Item Response Theory To Practical Testing Problems*: Routledge. https://doi.org/10.4324/9780203056615

Lord, F. M., Novick, M. R. & Birnbaum, A. (Eds.). (2008). *Statistical theories of mental test scores* (Behavioral science quantitative methods). Charlotte, NC: Information Age Publ.

Lüdtke, O. & Robitzsch, A. (2017). Eine Einführung in die Plausible-Values-Technik für die psychologische Forschung. *Diagnostica*, *63*(3), 193–205. https://doi.org/10.1026/0012-1924/a000175

Luecht, R. M. (2003). Exposure Control Using Adaptive Multi-Stage Item Bundles.

Luecht, R. M., Brumfield, T. & Breithaupt, K. (2006). A Testlet Assembly Design for Adaptive Multistage Tests. *Applied Measurement in Education*, *19*(3), 189–202. https://doi.org/10.1207/s15324818ame1903_2

Luecht, R. M. & Nungester, R. J. (1998). Some Practical Examples of Computer-Adaptive Sequential Testing. *Journal of Educational Measurement*, *35*(3), 229–249. https://doi.org/10.1111/j.1745-3984.1998.tb00537.x

Lunz, M. E., Bergstrom, B. A. & Gershon, R. C. (1994). Computer adaptive testing. *International Journal of Educational Research*, *21*(6), 623–634. https://doi.org/10.1016/0883-0355(94)90015-9

MacCallum, R. C., Browne, M. W. & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological methods*, *1*(2), 130–149. https://doi.org/10.1037/1082-989X.1.2.130

Magis, D., Yan, D. & Davier, A. A. von. (2017). *Computerized Adaptive and Multistage Testing with R*. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-69218-0

Mayer, J. (2007). Erkenntnisgewinnung als wissenschaftliches Problemlösen. In D. Krüger & H. Vogt (Hrsg.), *Theorien in der biologiedidaktischen Forschung. Ein Handbuch für Lehramtsstudenten und Doktoranden* (Springer-Lehrbuch, 1st ed., S. 177–187). Berlin, Heidelberg: Springer-Verlag Berlin Heidelberg. Zugriff am 10.10.2019.

McBride, J. R. & Martin, J. T. (1983). Reliability and validity of adaptive ability tests in a military setting. In D. J. Weiss (Ed.), *New Horizons in Testing. Latent Trait Test Theory and Computerized Adaptive Testing.* Burlington: Elsevier Science.

McDonald, R. P. (1999). *Test theory. A unified treatment*. Mahwah, NJ: L. Erlbaum Associates. Retrieved from http://site.ebrary.com/lib/alltitles/docDetail.action?docID=10333567

Microsoft Corporation & Weston, S. (2019a). *doParallel: Foreach Parallel Adaptor for the 'parallel' Package*. Verfügbar unter https://CRAN.R-project.org/package=doParallel

Microsoft Corporation & Weston, S. (2019b). *doSNOW: Foreach Parallel Adaptor for the 'snow' Package*. Verfügbar unter https://CRAN.R-project.org/package=doSNOW

Monseur, C. & Adams, R. (2009). Plausible Values: How to Deal with Their Limitations. *Journal of Applied Measurement*, *10*(3), 1–15. Zugriff am 26.08.2019.

Moosbrugger, H. (2012). Item-Response-Theory (IRT). In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (Springer-Lehrbuch, 2., aktualisierte und überarbeitete Auflage, S. 228–274). Berlin, Heidelberg: Springer-Verlag Berlin Heidelberg.

Moosbrugger, H. & Kelava, A. (Hrsg.). (2012). *Testtheorie und Fragebogenkonstruktion* (Springer-Lehrbuch, 2., aktualisierte und überarbeitete Auflage). Berlin, Heidelberg: Springer-Verlag Berlin Heidelberg. https://doi.org/10.1007/978-3-642-20072-4

Olsen, J. B. (1986). Comparison and equating of paper-administered, computer-administered and computerized adaptive tests of achievement. *Paper presented at the meeting of the American Educational Research Association, San Francisco, 1986*. Verfügbar unter https://ci.nii.ac.jp/naid/10020974946/en/

Öz, H. & Özturan, T. (2018). Computer-based and Paper-based Testing: Does the Test Administration Mode Influence the Reliability and Validity of Achievement Tests? *The journal of language and linguistic studies*, *14*(1), 67–85. Zugriff am 13.12.2019.

Paek, P. (2005). Recent Trends in Comparability Studies Using testing and assessment to promote learning. *Pearson Educational Measurement*.

Patsula, L. N. & Hambleton, R. K. (1999). *A comparative study of ability estimates from computer-adaptive testing and multi-stage testing.* Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal Canada.

Patsula, L. N. (1999). *A comparison of computerized adaptive testing and multi-stage testing.* Dissertation. University of Massachusetts Amherst. Zugriff am 31.01.2020.

R Core Team. (2020). *R: A Language and Environment for Statistical Computing,* R Foundation for Statistical Computing. Verfügbar unter https://www.R-project.org/

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago usw.: Univ. of Chicago Pr.

Reckase, M. D. (2009). *Multidimensional Item Response Theory* (Statistics for social and behavorial sciences, 1. Aufl.). s.l.: Springer-Verlag. Retrieved

from http://site.ebrary.com/lib/alltitles/docDetail.action?docID=10315546

Roberts, R. D., Goff, G. N., Anjoul, F., Kyllonen, P. C., Pallier, G. & Stankov, L. (2000). The Armed Services Vocational Aptitude Battery (ASVAB). *Learning and Individual Differences*, *12*(1), 81–103. https://doi.org/10.1016/S1041-6080(00)00035-2

Robitzsch, A., Kiefer, T. & Wu, M. (2020). *TAM: Test Analysis Modules*. Verfügbar unter https://CRAN.R-project.org/package=TAM

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

Rotou, O., Patsula, L. N., Steffen, M. & Rizavi, S. (2007). Comparison of Multistage Tests With Computerized Adaptive and Paper-and-Pencil Tests. *ETS Research Report Series*, *2007*(4). Zugriff am 14.10.2019.

Sari, H. I., Yahsi-Sari, H. & Huggins-Manley, A. C. (2016). Computer Adaptive Multistage Testing: Practical Issues, Challenges and Principles. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, 388. https://doi.org/10.21031/epod.280183

Schmidt-Atzert, L., Deter, B. & Jaeckel, S. (2004). Prädiktion von Ausbildungserfolg: Allgemeine Intelligenz (g) oder spezifische kognitive Fähigkeiten? *Zeitschrift für Personalpsychologie*, *3*(4), 147–158. https://doi.org/10.1026/1617-6391.3.4.147

Schnipke, D. L. & Reese, L. M. (1997). A Comparison of Testlet-Based Test Designs for Computerized Adaptive Testing. *Paper presented at the meeting of the American Educational Research Association, Chicago, 1997*. Zugriff am 03.02.2020.

Schulz, M. (2002). The Standardization of Mean-Squares. *Rasch Measurement Transactions*, *16*(2), 879. Verfügbar unter https://www.rasch.org/rmt/rmt162g.htm

Schweizer, K. (Hrsg.). (2006). *Leistung und Leistungsdiagnostik. Mit 18 Tabellen*. Berlin/Heidelberg: Springer-Verlag. https://doi.org/10.1007/3-540-33020-8

Segall, D. O. (2005). Computerized Adaptive Testing. In K. Kempf-Leonard (Ed.), *Encyclopedia of social measurement.* Amsterdam: Elsevier. Zugriff am 03.02.2020.

Smith, R. M., Schumacker, R. E. & Bush, M. J. (1998). Using item mean squares to evaluate fit to the Rasch model. *Journal of Outcome Measurement*, *2*(1), 66–78.

Stocking, M. L. & Lewis, C. (2002). Methods of Controlling the Exposure of Items in CAT. In W. J. Linden & G. A.W. Glas (Eds.), *Computerized Adaptive Testing: Theory and Practice* (pp. 163–182). Dordrecht: Kluwer Academic Publishers.

Straube, P. (2016). *Modellierung und Erfassung von Kompetenzen naturwissenschaftlicher Erkenntnisgewinnung bei (Lehramts-) Studierenden im Fach Physik*. Dissertation. Freie Universität Berlin, Berlin. Zugriff am 14.10.2019.

Tennant, A. & Pallant, J. F. (2012). The Root Mean Square Error of Approximation (RMSEA) as a supplementary statistic to determine fit to the Rasch model with large sample sizes. *Rasch Measurement Transactions*, *25*(4), 1348–1349. Verfügbar unter https://www.rasch.org/rmt/rmt254d.htm

Terzer, E., Hartig, J. & Upmeier zu Belzen, A. (2013). Systematische Konstruktion eines Tests zu Modellkompetenz im Biologieunterricht unter Berücksichtigung von Gütekriterien. *Zeitschrift der Didaktik der Naturwissenschaften*, *19*, 51–76. Zugriff am 09.08.2020.

Tierney, L., Rossini, A. T., Li, N. & Sevcikova, H. (2018). *snow: Simple Network of Workstations*. Verfügbar unter https://CRAN.R-project.org/package=snow

Upmeier zu Belzen, A. & Krüger, D. (2010). Modellkompetenz im Biologieunterricht. *Zeitschrift der Didaktik der Naturwissenschaften*, *16*, 41–57. Zugriff am 22.07.2020.

Urry, V. W. (1970). *Monte Carlo investigation of logistic test models*. Dissertation. Purdue University, West Lafayette, IN.

Van der Linden, W. J. (1999). Empirical Initialization of the Trait Estimator in Adaptive Testing. *Applied Psychological Measurement*, *23*(1), 21–29. https://doi.org/10.1177/01466219922031149

Van der Linden, W. J. & Glas, C. A.W. (Eds.). (2010). *Elements of Adaptive Testing* (Statistics for Social and Behavioral Sciences). New York, NY:

Springer Science+Business Media LLC. https://doi.org/10.1007/978-0-387-85461-8

Van der Linden, W. J. & Hambleton, R. K. (Eds.). (1997). *Handbook of Modern Item Response Theory*. New York: Springer. https://doi.org/10.1007/978-1-4757-2691-6

Vrieze, S. I. (2012). Model selection and psychological theory: a discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological Methods*, *17*(2), 228–243. https://doi.org/10.1037/a0027127

Wang, K. (2017). *A Fair Comparison of the Performance of Computerized Adaptive Testing and Multistage Adaptive Testing*. Dissertation. Michigan State University. Zugriff am 14.10.2019.

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

Weinert, F. E. (Hrsg.). (2002). *Leistungsmessungen in Schulen* (Beltz Pädagogik, 2., unveränd. Aufl., Dr. nach Typoskript). Weinheim: Beltz. Zugriff am 14.10.2019.

Weiss, D. J. (1982). Improving Measurement Quality and Efficiency with Adaptive Testing. *Applied Psychological Measurement*, *6*(4), 473–492. https://doi.org/10.1177/014662168200600408

Weiss, D. J. (Ed.). (1983). *New Horizons in Testing. Latent Trait Test Theory and Computerized Adaptive Testing*. Burlington: Elsevier Science. Accessed 14.10.2019.

Weissman, A., Belov, D. & Armstrong, R. (2007, Oktober). *Information-Based Versus Number-Correct Routing in Multistage Classification Tests* (LSAC Research Report Series). Law School Admision Council (LSAC).

Wilson, E. B. & Hilferty, M. M. (1931). The Distribution of Chi-Square. *Proceedings of the National Academy of Sciences of the United States of America*, *17*(12), 684–688. https://doi.org/10.1073/pnas.17.12.684

Wright, B. D. (1989). Dichotomous Rasch Model derived from Counting Right Answers: Raw Scores as Sufficient Statistics. *Rasch Measurement Transactions*, *3*(2), 62. Verfügbar unter https://www.rasch.org/rmt/rmt32e.htm

Wright, B. D. & Masters, G. N. (1982). *Rating Scale Analysis. Rasch measurement*. Chicago, Ill.: Mesa Pr. Zugriff am 30.01.2020.

Wu, M. (2004). Plausible Values. *Rasch Measurement Transactions*, *18*(2), 976–978. Zugriff am 30.01.2020.

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation*, *31*(2-3), 114–128. https://doi.org/10.1016/j.stueduc.2005.05.005

Yan, D., Davier, A. A. von & Lewis, C. (Eds.). (2014). *Computerized Multistage Testing. Theory and Applications* (Chapman & Hall / CRC Statistics in the Social and Behavioral Sciences). Hoboken: Taylor and Francis. Accessed 14.10.2019.

Zheng, Y. & Chang, H.-H. (2015). On-the-Fly Assembled Multistage Adaptive Testing. *Applied Psychological Measurement*, *39*(2), 104–118. https://doi.org/10.1177/0146621614544519

Zlatkin-Troitschanskaia, O., Pant, H. A., Nagel, M.-T., Molerov, D., Lautenbach, C. & Toepper, M. (Hrsg.). (2020). *Portfolio of KoKoHs Assessments. Test Instruments for Modeling and Measuring Domain-specific and Generic Competencies of Higher Education Students and Graduates*. Mainz, Berlin. Zugriff am 24.07.2020.

Anhang A – Itemparameter und Fitwerte 1pl

Itemname	Schwierig-keit	Itemname	Schwierig-keit
BA_Blattnektar_01	0.092	CH_Druckknoepfe_011	0.207
BA_Finger_011	0.49	CH_Kunststoffe_02	-0.606
BA_Magnetsinn_01	-0.219	CP_Fluessigkeit_01	0.161
BE_Museum_03	-0.17	CP_Reaktionen_01	0.046
BE_Urschildkroete_01	0.386	CT_dirigierende_021	0.495
BF_Ausserirdisch_01	0.905	CT_Kohlensaeure_06	0.803
BF_Biotueten_01	-0.242	CT_Kugel_Stab_Modell_01	1.191
BF_Sojabohne_01	0.577	CZ_Rutherford_04	0.49
BH_Fledermaus_01	0.111	CZ_Thomsonsches_01	1.381
BH_Im_Weltall_01	0.105	PA_Gammadetektor_01	0.063
BH_Vegetation_02	0.04	PA_Radioaktiv_01	0.395
BP_Eichhoernchen_02	0.556	PA_Wegzeit_01	-1.119
BP_Mais_01	-0.462	PE_Kondensator_03	0.19
BP_Phantom_01	-0.16	PF_Kompakt_01	-0.439
BP_Schimmelpilz_01	0.773	PF_Luftblasen_01	-0.273
BP_Verhalten_01	-0.689	PF_Radioaktiv_022	0.696
BT_Biomembran_01	0.588	PH_Mars_012	0.279
BT_Kuerbis_011	0.462	PH_Ohmsches_03	0.308
BZ_Essen_01	1.337	PP_Erdbeschleunigung_03	-0.896
BZ_Klima_01	1.527	PP_Materialeig_05	-0.08
BZ_Sprache_01	-0.509	PP_Reibungskraft_01	1.141
CA_Kohlensaeure_04	-1.235	PP_Schiefe_02	0.182
CA_Kohlensaeure_05	-0.807	PP_UV_Test_01	-0.338
CA_Loeslichkeit_041	0.61	PT_Magnetfeld_02	0.956
CE_Kohlensaeure_03	1.44	PT_Meeresspiegel_021	-0.412
CF_Meerwasser_01	-0.535	PT_Welten_01	-0.57
CF_Muell_011	-0.712	PZ_Meeresspiegel_01	-0.309
CF_Sonnenblu-menoel_01	0.014	PZ_Ohmsches_01	0.538
CH_Diamant_01	-0.025	PZ_Optik_01	1.323

Itemname	Outfit	Outfit_t	Outfit_p
BA_Blattnektar_01	0.994	-0.225	0.822
BA_Finger_011	1.017	0.566	0.572
BA_Magnetsinn_01	1.004	0.094	0.925
BE_Museum_03	1.022	0.955	0.34
BE_Urschildkroete_01	0.986	-0.597	0.551
BF_Ausserirdisch_01	1.009	0.239	0.811
BF_Biotueten_01	1.01	0.399	0.69
BF_Sojabohne_01	1.03	0.953	0.34
BH_Fledermaus_01	0.99	-0.441	0.659
BH_Im_Weltall_01	1.002	0.102	0.919
BH_Vegetation_02	1.044	2.11	0.035
BP_Eichhoernchen_02	0.974	-0.9	0.368
BP_Mais_01	0.99	-0.336	0.737
BP_Phantom_01	0.972	-0.648	0.517
BP_Schimmelpilz_01	1.003	0.069	0.945
BP_Verhalten_01	0.997	-0.055	0.956
BT_Biomembran_01	0.971	-0.885	0.376
BT_Kuerbis_011	0.985	-0.472	0.637
BZ_Essen_01	1.028	0.46	0.645
BZ_Klima_01	1.059	0.77	0.441
BZ_Sprache_01	1.015	0.541	0.589
CA_Kohlensaeure_04	0.985	-0.127	0.899
CA_Kohlensaeure_05	0.976	-0.661	0.508
CA_Loeslichkeit_041	1.032	0.813	0.416
CE_Kohlensaeure_03	1.021	0.299	0.765
CF_Meerwasser_01	0.993	-0.179	0.858
CF_Muell_011	1.018	0.313	0.754
CF_Sonnenblumenoel_01	1.002	0.073	0.942
CH_Diamant_01	1.029	1.28	0.2
CH_Druckknoepfe_011	1.007	0.353	0.724
CH_Kunststoffe_02	1.008	0.225	0.822
CP_Fluessigkeit_01	1.015	0.42	0.674
CP_Reaktionen_01	0.967	-1.442	0.149
CT_dirigierende_021	1.033	1.001	0.317
CT_Kohlensaeure_06	1.028	0.659	0.51
CT_Kugel_Stab_Modell_01	0.957	-0.391	0.696
CZ_Rutherford_04	1.03	1.092	0.275
CZ_Thomsonsches_01	0.966	-0.222	0.824
PA_Gammadetektor_01	0.99	-0.377	0.706
PA_Radioaktiv_01	0.977	-0.848	0.396
PA_Wegzeit_01	1.005	0.112	0.911
PE_Kondensator_03	1.005	0.228	0.819
PF_Kompakt_01	1	0.026	0.979

Itemname	Outfit	Outfit_t	Outfit_p
PF_Luftblasen_01	0.991	-0.339	0.735
PF_Radioaktiv_022	0.97	-0.705	0.481
PH_Mars_012	1.003	0.115	0.908
PH_Ohmsches_03	0.966	-1.431	0.152
PP_Erdbeschleunigung_03	0.969	-0.347	0.729
PP_Materialeig_05	1.022	0.927	0.354
PP_Reibungskraft_01	1.012	0.235	0.814
PP_Schiefe_02	0.971	-0.648	0.517
PP_UV_Test_01	0.961	-1.515	0.13
PT_Magnetfeld_02	1.006	0.13	0.897
PT_Meeresspiegel_021	1.003	0.111	0.911
PT_Welten_01	0.962	-1.151	0.25
PZ_Meeresspiegel_01	0.972	-1.191	0.234
PZ_Ohmsches_01	1.02	0.657	0.511
PZ_Optik_01	1.002	0.056	0.955

Itemname	Infit	Infit_t	Infit_p
BA_Blattnektar_01	0.995	-0.213	0.831
BA_Finger_011	1.013	0.53	0.596
BA_Magnetsinn_01	1.004	0.116	0.908
BE_Museum_03	1.021	1.075	0.283
BE_Urschildkroete_01	0.988	-0.574	0.566
BF_Ausserirdisch_01	1.007	0.217	0.828
BF_Biotueten_01	1.007	0.335	0.738
BF_Sojabohne_01	1.022	0.87	0.384
BH_Fledermaus_01	0.991	-0.466	0.641
BH_Im_Weltall_01	1.002	0.112	0.911
BH_Vegetation_02	1.038	2.083	0.037
BP_Eichhoernchen_02	0.977	-0.946	0.344
BP_Mais_01	0.988	-0.501	0.616
BP_Phantom_01	0.976	-0.643	0.52
BP_Schimmelpilz_01	1.003	0.075	0.94
BP_Verhalten_01	0.999	-0.032	0.974
BT_Biomembran_01	0.975	-0.898	0.369
BT_Kuerbis_011	0.985	-0.559	0.576
BZ_Essen_01	1.014	0.293	0.769
BZ_Klima_01	1.02	0.333	0.739
BZ_Sprache_01	1.01	0.446	0.656
CA_Kohlensaeure_04	0.993	-0.059	0.953
CA_Kohlensaeure_05	0.984	-0.521	0.602

Itemname	Infit	Infit_t	Infit_p
CA_Loeslichkeit_041	1.019	0.587	0.557
CE_Kohlensaeure_03	1.011	0.203	0.839
CF_Meerwasser_01	0.995	-0.164	0.87
CF_Muell_011	1.005	0.123	0.902
CF_Sonnenblumenoel_01	1.002	0.07	0.945
CH_Diamant_01	1.025	1.275	0.202
CH_Druckknoepfe_011	1.005	0.304	0.761
CH_Kunststoffe_02	1.007	0.229	0.819
CP_Fluessigkeit_01	1.014	0.446	0.656
CP_Reaktionen_01	0.971	-1.466	0.143
CT_dirigierende_021	1.03	1.093	0.274
CT_Kohlensaeure_06	1.021	0.614	0.539
CT_Kugel_Stab_Modell_01	0.986	-0.146	0.884
CZ_Rutherford_04	1.024	1.067	0.286
CZ_Thomsonsches_01	0.987	-0.091	0.928
PA_Gammadetektor_01	0.992	-0.338	0.735
PA_Radioaktiv_01	0.981	-0.844	0.398
PA_Wegzeit_01	1.004	0.103	0.918
PE_Kondensator_03	1.004	0.222	0.824
PF_Kompakt_01	1.002	0.093	0.926
PF_Luftblasen_01	0.994	-0.277	0.782
PF_Radioaktiv_022	0.974	-0.728	0.466
PH_Mars_012	1.001	0.042	0.966
PH_Ohmsches_03	0.971	-1.422	0.155
PP_Erdbeschleunigung_03	0.981	-0.253	0.8
PP_Materialeig_05	1.02	0.966	0.334
PP_Reibungskraft_01	1	0.011	0.991
PP_Schiefe_02	0.973	-0.686	0.492
PP_UV_Test_01	0.967	-1.472	0.141
PT_Magnetfeld_02	1.004	0.116	0.908
PT_Meeresspiegel_021	1.004	0.207	0.836
PT_Welten_01	0.971	-1.085	0.278
PZ_Meeresspiegel_01	0.977	-1.137	0.255
PZ_Ohmsches_01	1.016	0.627	0.53
PZ_Optik_01	0.995	-0.08	0.936

Itemname	RMSD	Itemname	RMSD
BA_Blattnektar_01	0.006	CH_Druckknoepfe_011	0.009
BA_Finger_011	0.012	CH_Kunststoffe_02	0.02
BA_Magnetsinn_01	0.014	CP_Fluessigkeit_01	0.029
BE_Museum_03	0.027	CP_Reaktionen_01	0.033
BE_Urschildkroete_01	0.011	CT_dirigierende_021	0.025
BF_Ausserirdisch_01	0.01	CT_Kohlensaeure_06	0.022
BF_Biotueten_01	0.01	CT_Kugel_Stab_Modell_01	0.023
BF_Sojabohne_01	0.024	CZ_Rutherford_04	0.013
BH_Fledermaus_01	0.015	CZ_Thomsonsches_01	0.009
BH_Im_Weltall_01	0.01	PA_Gammadetektor_01	0.017
BH_Vegetation_02	0.036	PA_Radioaktiv_01	0.007
BP_Eichhoernchen_02	0.023	PA_Wegzeit_01	0.007
BP_Mais_01	0.019	PE_Kondensator_03	0.009
BP_Phantom_01	0.023	PF_Kompakt_01	0.011
BP_Schimmelpilz_01	0.008	PF_Luftblasen_01	0.029
BP_Verhalten_01	0.005	PF_Radioaktiv_022	0.011
BT_Biomembran_01	0.024	PH_Mars_012	0.028
BT_Kuerbis_011	0.019	PH_Ohmsches_03	0.017
BZ_Essen_01	0.013	PP_Erdbeschleunigung_03	0.021
BZ_Klima_01	0.022	PP_Materialeig_05	0.014
BZ_Sprache_01	0.012	PP_Reibungskraft_01	0.027
CA_Kohlensaeure_04	0.008	PP_Schiefe_02	0.034
CA_Kohlensaeure_05	0.015	PP_UV_Test_01	0.006
CA_Loeslichkeit_041	0.024	PT_Magnetfeld_02	0.01
CE_Kohlensaeure_03	0.011	PT_Meeresspiegel_021	0.027
CF_Meerwasser_01	0.005	PT_Welten_01	0.021
CF_Muell_011	0.019	PZ_Meeresspiegel_01	0.016
CF_Sonnenblumenoel_01	0.008	PZ_Ohmsches_01	0.017
CH_Diamant_01	0.025	PZ_Optik_01	0.009

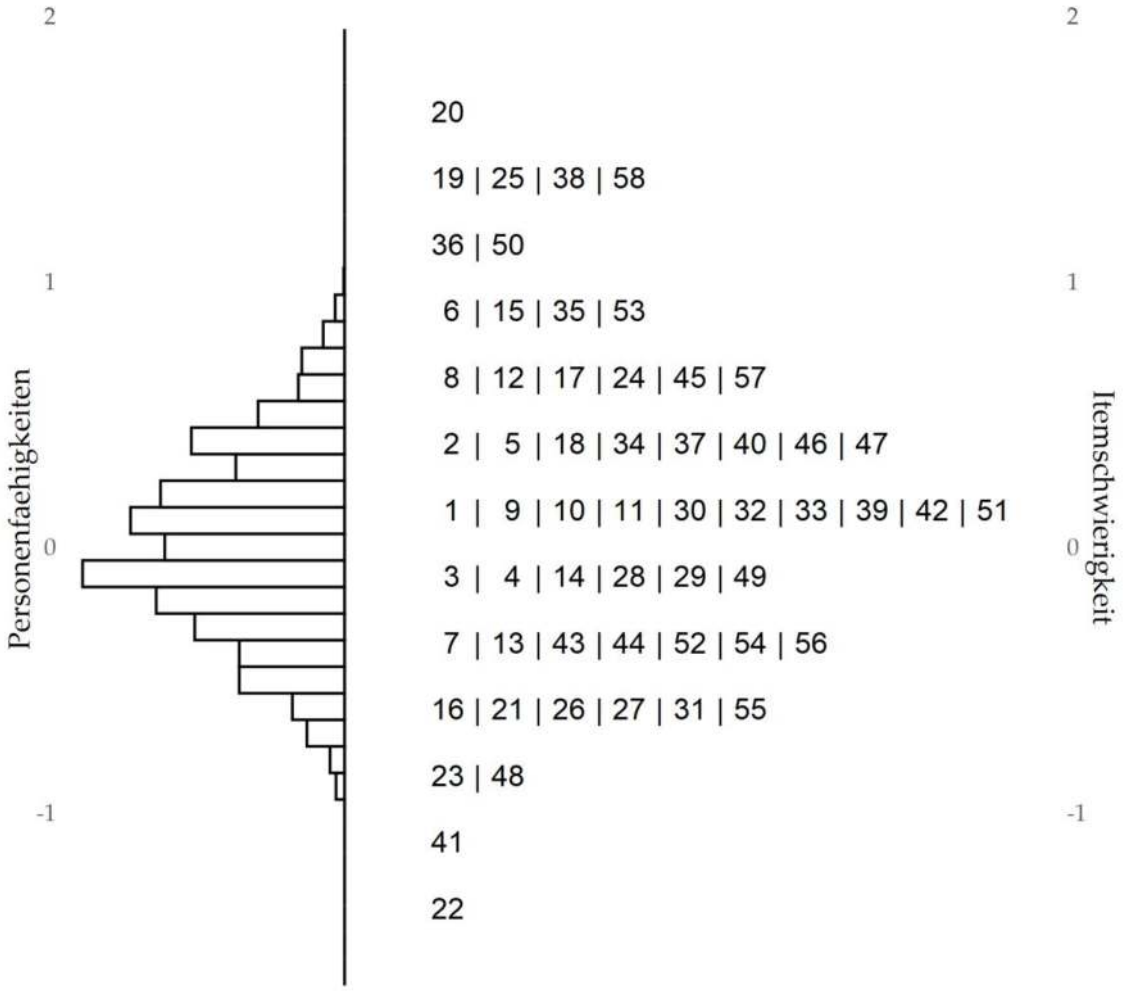

Abbildung 27: Wrightmap des 1pl-Modells (siehe Abschnitt 6.4.1). Die Verteilung der Personenfähigkeiten im Datensatz (links) sind den Itemschwierigkeiten (rechts) in einer gemeinsamen Skala gegenübergestellt. Die Items sind ihrer alphabetischen Reihenfolge nach durchnummeriert.

Anhang B – Itemparameter und Fitwerte 2pl

Itemname	Schwierig-keit	Trennschärfe	RMSD
BA_Blattnektar_01	0.03	0.572	0.01
BA_Finger_011	0.469	0.401	0.012
BA_Kokainsucht_01	-1.45	1.191	0.01
BE_Art_01	1.216	0.926	0.011
BE_Museum_03	-0.278	0.374	0.02
BE_Urschildkroete_01	0.403	0.583	0.011
BF_Ausserirdisch_01	0.876	0.468	0.013
BF_Biotueten_01	-0.194	0.44	0.006
BF_Sojabohne_01	0.582	0.375	0.006
BH_Fledermaus_01	0.058	0.558	0.016
BH_Im_Weltall_01	0.088	0.468	0.013
BP_Eichhoernchen_02	0.568	0.763	0.009
BP_Kohlweissling_02	-1.144	0.407	0.027
BP_Mais_01	-0.448	0.775	0.013
BP_Phantom_01	-0.241	0.957	0.005
BP_Verhalten_01	-0.722	0.666	0.01
BT_Biomembran_01	0.642	0.701	0.008
BT_Kuerbis_011	0.418	0.734	0.013
BT_Kuerbis_012	-0.518	1.622	0.009
BT_Museum_02	-0.119	1.056	0.01
BT_Population_01	0.993	0.902	0.006
BT_Sprache_02	0.026	0.48	0.034
BZ_Kartoffel_01	-0.852	0.519	0.023
BZ_Sprache_01	-0.553	0.564	0.01
CA_Kohlensaeure_05	-0.846	0.806	0.003
CA_Loeslichkeit_041	0.566	0.32	0.016
CE_Kohlensaeure_03	1.384	0.485	0.012
CF_Meerwasser_01	-0.607	0.469	0.007
CH_Druckknoepfe_011	0.321	0.316	0.017
CH_Kunststoffe_02	-0.648	0.512	0.005
CP_Ionennachweise_01	-0.781	0.894	0.02
CP_Ionennachweise_02	-0.308	1.066	0.008
CP_Loeslichkeit_043	-0.403	1.028	0.01
CP_Reaktionen_01	-0.001	0.911	0.004
CT_Strukturm_04	0.966	1.436	0.005
CZ_dirigierende_011	-0.53	1.032	0.002
CZ_Thomsonsches_01	1.472	1.697	0.008
PA_Gammadetektor_01	0.071	0.769	0.003

Itemname	Schwierig-keit	Trennschärfe	RMSD
PA_Radioaktiv_01	0.42	0.681	0.01
PA_Radioaktiv_05	0.24	0.794	0.018
PA_Szintillation_01	0.368	0.939	0.004
PE_Dusche_03	-1.272	0.441	0.02
PE_Kondensator_03	0.182	0.524	0.015
PE_Standardmodell_01	-0.646	0.999	0.005
PF_Kompakt_01	-0.426	0.413	0.013
PF_Luftblasen_01	-0.251	0.716	0.01
PF_Radioaktiv_022	0.726	0.737	0.014
PH_Erdbeschleunigung_02	-0.078	0.64	0.026
PH_Mars_012	0.293	0.437	0.008
PH_Materialeig_012	-1.539	1.238	0.011
PH_Ohmsches_03	0.357	0.742	0.003
PP_Materialeig_02	-0.963	1.265	0.013
PP_UV_Test_01	-0.462	0.909	0.004
PT_Magnetfeld_02	0.956	0.503	0.011
PT_Massepunkt_02	0.219	0.751	0.005
PT_Meeresspiegel_021	-0.43	0.468	0.007
PT_Planck_01	1.142	0.679	0.01
PT_Welten_01	-0.569	0.885	0.003
PZ_Meeresspiegel_01	-0.273	0.758	0.009
PZ_Ohmsches_01	0.567	0.545	0.009
PZ_Optik_01	1.31	0.508	0.015
PZ_Wellenmodell_01	-0.643	0.731	0.012

Itemname	Outfit	Outfit_t	Outfit_p
BA_Blattnektar_01	0.999	-0.042	0.967
BA_Finger_011	1.001	0.071	0.943
BA_Kokainsucht_01	0.971	-0.287	0.774
BE_Art_01	0.992	-0.015	0.988
BE_Museum_03	0.999	-0.048	0.962
BE_Urschildkroete_01	0.999	-0.032	0.974
BF_Ausserirdisch_01	1.003	0.09	0.928
BF_Biotueten_01	1	0.025	0.98
BF_Sojabohne_01	1	0.01	0.992
BH_Fledermaus_01	0.999	-0.022	0.983
BH_Im_Weltall_01	1	-0.003	0.997
BP_Eichhoernchen_02	1	0.021	0.984
BP_Kohlweissling_02	1.006	0.089	0.929
BP_Mais_01	1.008	0.201	0.841

Itemname	Outfit	Outfit_t	Outfit_p
BP_Phantom_01	1.003	0.068	0.946
BP_Verhalten_01	0.999	-0.02	0.984
BT_Biomembran_01	1.005	0.138	0.89
BT_Kuerbis_011	1.004	0.116	0.908
BT_Kuerbis_012	1.015	0.158	0.875
BT_Museum_02	0.995	0	1
BT_Population_01	0.999	-0.001	0.999
BT_Sprache_02	1.002	0.053	0.958
BZ_Kartoffel_01	0.999	0.025	0.98
BZ_Sprache_01	1.003	0.115	0.909
CA_Kohlensaeure_05	0.998	-0.019	0.985
CA_Loeslichkeit_041	1.001	0.056	0.955
CE_Kohlensaeure_03	1.007	0.127	0.899
CF_Meerwasser_01	1.001	0.03	0.976
CH_Druckknoepfe_011	1.001	0.05	0.96
CH_Kunststoffe_02	1.001	0.031	0.975
CP_Ionennachweise_01	0.986	-0.108	0.914
CP_Ionennachweise_02	1	0.016	0.987
CP_Loeslichkeit_043	0.995	-0.087	0.93
CP_Reaktionen_01	1	0.011	0.992
CT_Strukturm_04	0.992	0.023	0.982
CZ_dirigierende_011	1.001	0.041	0.967
CZ_Thomsonsches_01	0.955	0.007	0.994
PA_Gammadetektor_01	1	0.005	0.996
PA_Radioaktiv_01	1.001	0.051	0.959
PA_Radioaktiv_05	1.002	0.055	0.956
PA_Szintillation_01	1	0.046	0.963
PE_Dusche_03	0.992	-0.005	0.996
PE_Kondensator_03	1.001	0.066	0.948
PE_Standardmodell_01	0.996	-0.06	0.952
PF_Kompakt_01	0.999	-0.035	0.972
PF_Luftblasen_01	0.999	-0.04	0.968
PF_Radioaktiv_022	1.01	0.231	0.818
PH_Erdbeschleunigung_02	0.998	-0.03	0.976
PH_Mars_012	0.999	-0.017	0.986
PH_Materialeig_012	1.038	0.223	0.824
PH_Ohmsches_03	0.999	-0.036	0.971
PP_Materialeig_02	1.032	0.211	0.833
PP_UV_Test_01	0.997	-0.055	0.956
PT_Magnetfeld_02	0.996	-0.065	0.948
PT_Massepunkt_02	1.001	0.038	0.97
PT_Meeresspiegel_021	1	0.004	0.997
PT_Planck_01	1.007	0.102	0.919

Itemname	Outfit	Outfit_t	Outfit_p
PT_Welten_01	0.999	-0.006	0.995
PZ_Meeresspiegel_01	0.997	-0.067	0.947
PZ_Ohmsches_01	0.997	-0.068	0.946
PZ_Optik_01	1.008	0.154	0.877
PZ_Wellenmodell_01	0.995	-0.035	0.972

Itemname	Infit	Infit_t	Infit_p
BA_Blattnektar_01	1	0.029	0.977
BA_Finger_011	0.999	-0.021	0.983
BA_Kokainsucht_01	1.006	0.138	0.891
BE_Art_01	1.003	0.059	0.953
BE_Museum_03	1	0.03	0.976
BE_Urschildkroete_01	1	0.024	0.981
BF_Ausserirdisch_01	0.999	-0.024	0.981
BF_Biotueten_01	1	-0.003	0.998
BF_Sojabohne_01	1	0.008	0.994
BH_Fledermaus_01	1	0.021	0.983
BH_Im_Weltall_01	1	0.009	0.993
BP_Eichhoernchen_02	1	0.002	0.998
BP_Kohlweissling_02	0.998	0.006	0.995
BP_Mais_01	0.998	-0.061	0.951
BP_Phantom_01	0.999	0.015	0.988
BP_Verhalten_01	1.001	0.032	0.975
BT_Biomembran_01	0.999	-0.033	0.974
BT_Kuerbis_011	0.999	-0.036	0.971
BT_Kuerbis_012	0.998	-0.02	0.984
BT_Museum_02	1.001	0.041	0.968
BT_Population_01	1	0.025	0.98
BT_Sprache_02	0.999	-0.007	0.995
BZ_Kartoffel_01	1	0.032	0.975
BZ_Sprache_01	0.999	-0.038	0.97
CA_Kohlensaeure_05	1	0.022	0.982
CA_Loeslichkeit_041	1	-0.014	0.989
CE_Kohlensaeure_03	0.998	-0.019	0.985
CF_Meerwasser_01	1	0.004	0.997
CH_Druckknoepfe_011	1	-0.015	0.988
CH_Kunststoffe_02	1	0.004	0.997
CP_Ionennachweise_01	1.004	0.081	0.935
CP_Ionennachweise_02	1	0.019	0.985
CP_Loeslichkeit_043	1.001	0.056	0.955
CP_Reaktionen_01	1	0.009	0.993
CT_Strukturm_04	1.002	0.05	0.96

Itemname	Infit	Infit_t	Infit_p
CZ_dirigierende_011	1	0.001	0.999
CZ_Thomsonsches_01	1.006	0.091	0.928
PA_Gammadetektor_01	1	0.015	0.988
PA_Radioaktiv_01	0.999	-0.012	0.99
PA_Radioaktiv_05	0.999	0.011	0.992
PA_Szintillation_01	1	0.026	0.979
PE_Dusche_03	1.002	0.062	0.95
PE_Kondensator_03	1	-0.02	0.984
PE_Standardmodell_01	1.001	0.044	0.965
PF_Kompakt_01	1	0.028	0.978
PF_Luftblasen_01	1.001	0.031	0.975
PF_Radioaktiv_022	0.997	-0.073	0.942
PH_Erdbeschleunigung_02	1.001	0.034	0.973
PH_Mars_012	1	0.017	0.987
PH_Materialeig_012	0.993	-0.001	0.999
PH_Ohmsches_03	1	0.024	0.981
PP_Materialeig_02	0.994	-0.005	0.996
PP_UV_Test_01	1.001	0.033	0.973
PT_Magnetfeld_02	1.001	0.046	0.963
PT_Massepunkt_02	1	0.009	0.993
PT_Meeresspiegel_021	1	0.011	0.991
PT_Planck_01	0.998	0.003	0.997
PT_Welten_01	1	0.016	0.987
PZ_Meeresspiegel_01	1.001	0.042	0.967
PZ_Ohmsches_01	1.001	0.04	0.968
PZ_Optik_01	0.998	-0.033	0.973
PZ_Wellenmodell_01	1.002	0.046	0.963

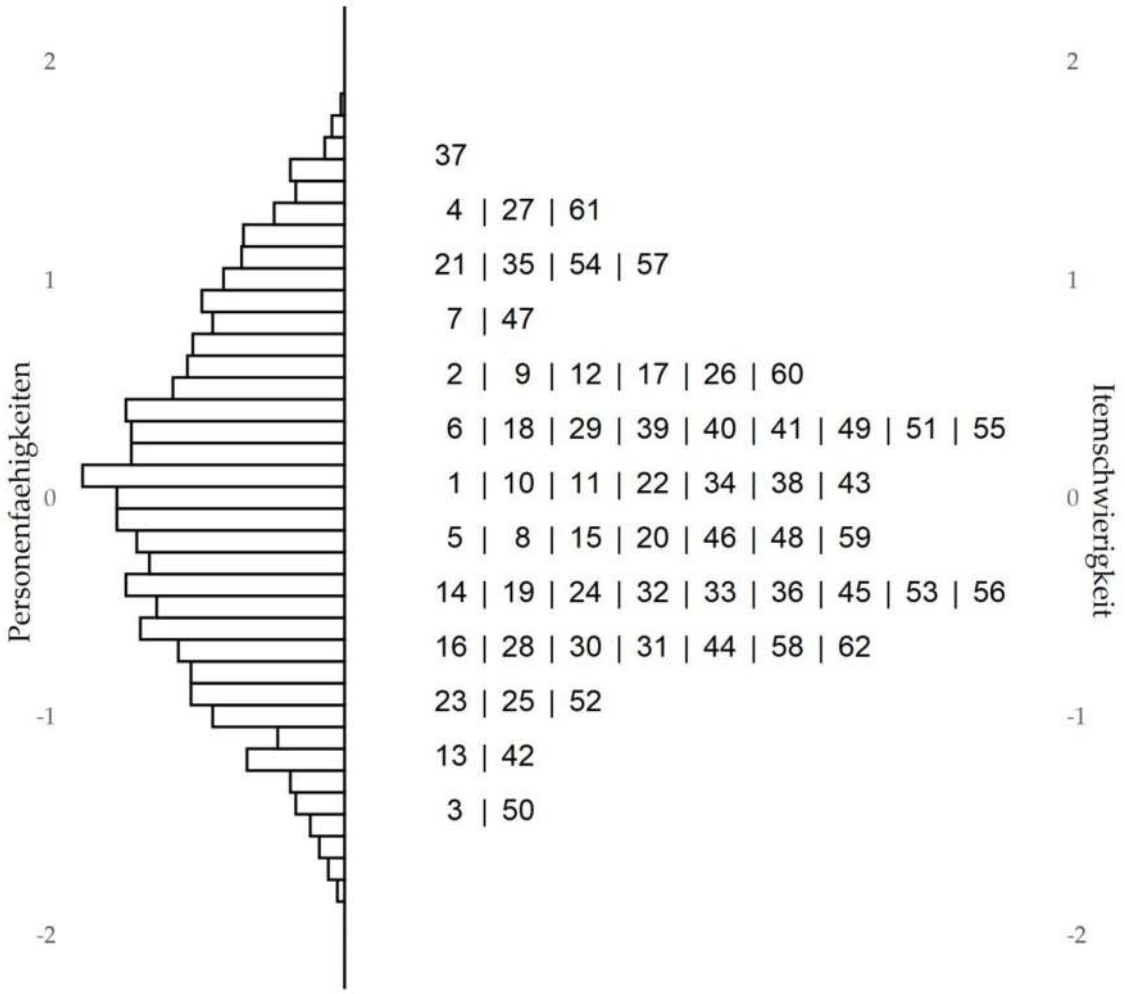

Abbildung 28: Wrightmap des 2pl-Modells (siehe Abschnitt 6.4.2). Die Verteilung der Personenfähigkeiten im Datensatz (links) sind den Itemschwierigkeiten (rechts) in einer gemeinsamen Skala gegenübergestellt. Die Items sind ihrer alphabetischen Reihenfolge nach durchnummeriert.

Danksagung

An dieser Stelle will ich all jenen danken, die mich auf meinem Weg zur Promotion begleitet und unterstützt haben.

Als erstes danke ich meinem Doktorvater, Herrn Prof. Dr. Volkhard Nordmeier. Ohne Ihr Vertrauen in meine Fähigkeiten und die Freiheiten, die Sie mir bei der Gestaltung meiner Arbeit ließen, hätte ich die Promotion weder beginnen noch beenden können.

Weiterhin danke ich Herrn Prof. Dr. Thomas Trefzger für das Interesse an meiner Arbeit und die Bereitschaft, als zweiter Gutachter an meinem Abschluss mitzuwirken.

Meinen Vorgängern und Mitdoktorierenden Philipp Straube, René Dohrmann, Nikola Schild, Julia Milster, Dorothee Ermel, Novid Ghassemi und Christine Meißner danke ich einerseits für die produktiven Diskussionen und andererseits für die moralische Unterstützung.

Besonderen Dank möchte ich an dieser Stelle Daniel Rehfeldt aussprechen. Du hast mich nicht nur ebenso tatkräftig unterstützt, durch deine Betreuung in der Masterarbeit und die anschließende Empfehlung habe ich diesen Weg überhaupt erst beschritten.

Sebastian Haase möchte ich dafür danken, dass er mir bei der Umsetzung des Tests in tet.folio jede noch so kleine Frage begeistert erklärt hat. Ich weiß nicht, wie ich das ohne dich hätte machen sollen.

Den Kolleginnen und Kollegen aus ValiDiS gilt mein Dank nicht zuletzt wegen der Gelegenheit zur Promotion, die mir durch die Aufnahme in das Projekt geboten wurde. Ausdrücklich danke ich Stefan Hartmann, der mir häufig Rat bei statistischen Fragen gab und immer gerne zu einem Gedankenaustausch bereit war.

Bei der gesamten AG Physikdidaktik bedanke ich mich für den herzlichen Empfang, der mir gemacht wurde. Mit euch Arbeiten zu können hat mir viel Freude bereitet. Ebenso danke ich den vielen Mitgliedern aus K2teach, mit denen ich regelmäßig viel Spaß in Fortbildungen hatte.

Schließlich danke ich meiner Familie, meinen Eltern und meinen beiden Brüdern. Ihr habt mich stets auf meinem Weg unterstützt und mir die nötige Motivation und einen Ort zum Ausruhen gegeben. Euch widme ich diese Arbeit.

Bisher erschienene Bände der Reihe *„Studien zum Physik- und Chemielernen“*

ISSN 1614-8967 (vormals *Studien zum Physiklernen* ISSN 1435-5280)

1 Helmut Fischler, Jochen Peuckert (Hrsg.): Concept Mapping in fachdidaktischen Forschungsprojekten der Physik und Chemie
ISBN 978-3-89722-256-4 40.50 EUR

2 Anja Schoster: Bedeutungsentwicklungsprozesse beim Lösen algorithmischer Physikaufgaben. *Eine Fallstudie zu Lernprozessen von Schülern im Physiknachhilfeunterricht während der Bearbeitung algorithmischer Physikaufgaben*
ISBN 978-3-89722-045-4 40.50 EUR

3 Claudia von Aufschnaiter: Bedeutungsentwicklungen, Interaktionen und situatives Erleben beim Bearbeiten physikalischer Aufgaben
ISBN 978-3-89722-143-7 40.50 EUR

4 Susanne Haeberlen: Lernprozesse im Unterricht mit Wasserstromkreisen. *Eine Fallstudie in der Sekundarstufe I*
ISBN 978-3-89722-172-7 40.50 EUR

5 Kerstin Haller: Über den Zusammenhang von Handlungen und Zielen. *Eine empirische Untersuchung zu Lernprozessen im physikalischen Praktikum*
ISBN 978-3-89722-242-7 40.50 EUR

6 Michaela Horstendahl: Motivationale Orientierungen im Physikunterricht
ISBN 978-3-89722-227-4 50.00 EUR

7 Stefan Deylitz: Lernergebnisse in der Quanten-Atomphysik. *Evaluation des Bremer Unterrichtskonzepts*
ISBN 978-3-89722-291-5 40.50 EUR

8 Lorenz Hucke: Handlungsregulation und Wissenserwerb in traditionellen und computergestützten Experimenten des physikalischen Praktikums
ISBN 978-3-89722-316-5 50.00 EUR

9 Heike Theyßen: Ein Physikpraktikum für Studierende der Medizin. *Darstellung der Entwicklung und Evaluation eines adressatenspezifischen Praktikums nach dem Modell der Didaktischen Rekonstruktion*
ISBN 978-3-89722-334-9 40.50 EUR

10 Annette Schick: Der Einfluß von Interesse und anderen selbstbezogenen Kognitionen auf Handlungen im Physikunterricht. *Fallstudien zu Interessenhandlungen im Physikunterricht*
ISBN 978-3-89722-380-6 40.50 EUR

11 Roland Berger: Moderne bildgebende Verfahren der medizinischen Diagnostik. *Ein Weg zu interessanterem Physikunterricht*
ISBN 978-3-89722-445-2 40.50 EUR

12 Johannes Werner: Vom Licht zum Atom. *Ein Unterrichtskonzept zur Quantenphysik unter Nutzung des Zeigermodells*
ISBN 978-3-89722-471-1 40.50 EUR

13 Florian Sander: Verbindung von Theorie und Experiment im physikalischen Praktikum. *Eine empirische Untersuchung zum handlungsbezogenen Vorverständnis und dem Einsatz grafikorientierter Modellbildung im Praktikum*
ISBN 978-3-89722-482-7 40.50 EUR

14 Jörn Gerdes: Der Begriff der physikalischen Kompetenz. *Zur Validierung eines Konstruktes*
ISBN 978-3-89722-510-7 40.50 EUR

15 Malte Meyer-Arndt: Interaktionen im Physikpraktikum zwischen Studierenden und Betreuern. *Feldstudie zu Bedeutungsentwicklungsprozessen im physikalischen Praktikum*
ISBN 978-3-89722-541-1 40.50 EUR

16 Dietmar Höttecke: Die Natur der Naturwissenschaften historisch verstehen. *Fachdidaktische und wissenschaftshistorische Untersuchungen*
ISBN 978-3-89722-607-4 40.50 EUR

17 Gil Gabriel Mavanga: Entwicklung und Evaluation eines experimentell- und phänomenorientierten Optikcurriculums. *Untersuchung zu Schülervorstellungen in der Sekundarstufe I in Mosambik und Deutschland*
ISBN 978-3-89722-721-7 40.50 EUR

18 Meike Ute Zastrow: Interaktive Experimentieranleitungen. *Entwicklung und Evaluation eines Konzeptes zur Vorbereitung auf das Experimentieren mit Messgeräten im Physikalischen Praktikum*
ISBN 978-3-89722-802-3 40.50 EUR

19 Gunnar Friege: Wissen und Problemlösen. *Eine empirische Untersuchung des wissenszentrierten Problemlösens im Gebiet der Elektrizitätslehre auf der Grundlage des Experten-Novizen-Vergleichs*
ISBN 978-3-89722-809-2 40.50 EUR

20 Erich Starauschek: Physikunterricht nach dem Karlsruher Physikkurs. *Ergebnisse einer Evaluationsstudie*
ISBN 978-3-89722-823-8 40.50 EUR

21 Roland Paatz: Charakteristika analogiebasierten Denkens. *Vergleich von Lernprozessen in Basis- und Zielbereich*
ISBN 978-3-89722-944-0 40.50 EUR

22 Silke Mikelskis-Seifert: Die Entwicklung von Metakonzepten zur Teilchenvorstellung bei Schülern. *Untersuchung eines Unterrichts über Modelle mithilfe eines Systems multipler Repräsentationsebenen*
ISBN 978-3-8325-0013-9 40.50 EUR

23 Brunhild Landwehr: Distanzen von Lehrkräften und Studierenden des Sachunterrichts zur Physik. *Eine qualitativ-empirische Studie zu den Ursachen*
ISBN 978-3-8325-0044-3 40.50 EUR

24 Lydia Murmann: Physiklernen zu Licht, Schatten und Sehen. *Eine phänomenografische Untersuchung in der Primarstufe*
ISBN 978-3-8325-0060-3 40.50 EUR

25 Thorsten Bell: Strukturprinzipien der Selbstregulation. *Komplexe Systeme, Elementarisierungen und Lernprozessstudien für den Unterricht der Sekundarstufe II*
ISBN 978-3-8325-0134-1 40.50 EUR

26 Rainer Müller: Quantenphysik in der Schule
ISBN 978-3-8325-0186-0 40.50 EUR

27 Jutta Roth: Bedeutungsentwicklungsprozesse von Physikerinnen und Physikern in den Dimensionen Komplexität, Zeit und Inhalt
ISBN 978-3-8325-0183-9 40.50 EUR

28 Andreas Saniter: Spezifika der Verhaltensmuster fortgeschrittener Studierender der Physik
ISBN 978-3-8325-0292-8 40.50 EUR

29 Thomas Weber: Kumulatives Lernen im Physikunterricht. *Eine vergleichende Untersuchung in Unterrichtsgängen zur geometrischen Optik*
ISBN 978-3-8325-0316-1 40.50 EUR

30 Markus Rehm: Über die Chancen und Grenzen moralischer Erziehung im naturwissenschaftlichen Unterricht
ISBN 978-3-8325-0368-0 40.50 EUR

31 Marion Budde: Lernwirkungen in der Quanten-Atom-Physik. *Fallstudien über Resonanzen zwischen Lernangeboten und SchülerInnen-Vorstellungen*
ISBN 978-3-8325-0483-0 40.50 EUR

32 Thomas Reyer: Oberflächenmerkmale und Tiefenstrukturen im Unterricht. *Exemplarische Analysen im Physikunterricht der gymnasialen Sekundarstufe*
ISBN 978-3-8325-0488-5 40.50 EUR

33 Christoph Thomas Müller: Subjektive Theorien und handlungsleitende Kognitionen von Lehrern als Determinanten schulischer Lehr-Lern-Prozesse im Physikunterricht
ISBN 978-3-8325-0543-1 40.50 EUR

34 Gabriela Jonas-Ahrend: Physiklehrervorstellungen zum Experiment im Physikunterricht
ISBN 978-3-8325-0576-9 40.50 EUR

35 Dimitrios Stavrou: Das Zusammenspiel von Zufall und Gesetzmäßigkeiten in der nichtlinearen Dynamik. *Didaktische Analyse und Lernprozesse*
ISBN 978-3-8325-0609-4 40.50 EUR

36 Katrin Engeln: Schülerlabors: authentische, aktivierende Lernumgebungen als Möglichkeit, Interesse an Naturwissenschaften und Technik zu wecken
ISBN 978-3-8325-0689-6 40.50 EUR

37 Susann Hartmann: Erklärungsvielfalt
ISBN 978-3-8325-0730-5 40.50 EUR

38 Knut Neumann: Didaktische Rekonstruktion eines physikalischen Praktikums für Physiker
ISBN 978-3-8325-0762-6 40.50 EUR

39 Michael Späth: Kontextbedingungen für Physikunterricht an der Hauptschule. *Möglichkeiten und Ansatzpunkte für einen fachübergreifenden, handlungsorientierten und berufsorientierten Unterricht*
ISBN 978-3-8325-0827-2 40.50 EUR

40 Jörg Hirsch: Interesse, Handlungen und situatives Erleben von Schülerinnen und Schülern beim Bearbeiten physikalischer Aufgaben
ISBN 978-3-8325-0875-3 40.50 EUR

41 Monika Hüther: Evaluation einer hypermedialen Lernumgebung zum Thema Gasgesetze. *Eine Studie im Rahmen des Physikpraktikums für Studierende der Medizin*
ISBN 978-3-8325-0911-8 40.50 EUR

42 Maike Tesch: Das Experiment im Physikunterricht. *Didaktische Konzepte und Ergebnisse einer Videostudie*
ISBN 978-3-8325-0975-0 40.50 EUR

43 Nina Nicolai: Skriptgeleitete Eltern-Kind-Interaktion bei Chemiehausaufgaben. *Eine Evaluationsstudie im Themenbereich Säure-Base*
ISBN 978-3-8325-1013-8 40.50 EUR

44 Antje Leisner: Entwicklung von Modellkompetenz im Physikunterricht
ISBN 978-3-8325-1020-6 40.50 EUR

45 Stefan Rumann: Evaluation einer Interventionsstudie zur Säure-Base-Thematik
ISBN 978-3-8325-1027-5 40.50 EUR

46 Thomas Wilhelm: Konzeption und Evaluation eines Kinematik/Dynamik-Lehrgangs zur Veränderung von Schülervorstellungen mit Hilfe dynamisch ikonischer Repräsentationen und graphischer Modellbildung – mit CD-ROM
ISBN 978-3-8325-1046-6 45.50 EUR

47 Andrea Maier-Richter: Computerunterstütztes Lernen mit Lösungsbeispielen in der Chemie. *Eine Evaluationsstudie im Themenbereich Löslichkeit*
ISBN 978-3-8325-1046-6 40.50 EUR

48 Jochen Peuckert: Stabilität und Ausprägung kognitiver Strukturen zum Atombegriff
ISBN 978-3-8325-1104-3 40.50 EUR

49 Maik Walpuski: Optimierung von experimenteller Kleingruppenarbeit durch Strukturierungshilfen und Feedback
ISBN 978-3-8325-1184-5 40.50 EUR

50 Helmut Fischler, Christiane S. Reiners (Hrsg.): Die Teilchenstruktur der Materie im Physik- und Chemieunterricht
ISBN 978-3-8325-1225-5 34.90 EUR

51 Claudia Eysel: Interdisziplinäres Lehren und Lernen in der Lehrerbildung. *Eine empirische Studie zum Kompetenzerwerb in einer komplexen Lernumgebung*
ISBN 978-3-8325-1238-5 40.50 EUR

52 Johannes Günther: Lehrerfortbildung über die Natur der Naturwissenschaften. *Studien über das Wissenschaftsverständnis von Grundschullehrkräften*
ISBN 978-3-8325-1287-3 40.50 EUR

53 Christoph Neugebauer: Lernen mit Simulationen und der Einfluss auf das Problemlösen in der Physik
ISBN 978-3-8325-1300-9 40.50 EUR

54 Andreas Schnirch: Gendergerechte Interessen- und Motivationsförderung im Kontext naturwissenschaftlicher Grundbildung. *Konzeption, Entwicklung und Evaluation einer multimedial unterstützten Lernumgebung*
ISBN 978-3-8325-1334-4 40.50 EUR

55 Hilde Köster: Freies Explorieren und Experimentieren. *Eine Untersuchung zur selbstbestimmten Gewinnung von Erfahrungen mit physikalischen Phänomenen im Sachunterricht*
ISBN 978-3-8325-1348-1 40.50 EUR

56 Eva Heran-Dörr: Entwicklung und Evaluation einer Lehrerfortbildung zur Förderung der physikdidaktischen Kompetenz von Sachunterrichtslehrkräften
ISBN 978-3-8325-1377-1 40.50 EUR

57 Agnes Szabone Varnai: Unterstützung des Problemlösens in Physik durch den Einsatz von Simulationen und die Vorgabe eines strukturierten Kooperationsformats
ISBN 978-3-8325-1403-7 40.50 EUR

58 Johannes Rethfeld: Aufgabenbasierte Lernprozesse in selbstorganisationsoffenem Unterricht der Sekundarstufe I zum Themengebiet ELEKTROSTATIK. *Eine Feldstudie in vier 10. Klassen zu einer kartenbasierten Lernumgebung mit Aufgaben aus der Elektrostatik*
ISBN 978-3-8325-1416-7 40.50 EUR

59 Christian Henke: Experimentell-naturwissenschaftliche Arbeitsweisen in der Oberstufe. *Untersuchung am Beispiel des HIGHSEA-Projekts in Bremerhaven*
ISBN 978-3-8325-1515-7 40.50 EUR

60 Lutz Kasper: Diskursiv-narrative Elemente für den Physikunterricht. *Entwicklung und Evaluation einer multimedialen Lernumgebung zum Erdmagnetismus*
ISBN 978-3-8325-1537-9 40.50 EUR

61 Thorid Rabe: Textgestaltung und Aufforderung zu Selbsterklärungen beim Physiklernen mit Multimedia
ISBN 978-3-8325-1539-3 40.50 EUR

62 Ina Glemnitz: Vertikale Vernetzung im Chemieunterricht. *Ein Vergleich von traditionellem Unterricht mit Unterricht nach Chemie im Kontext*
ISBN 978-3-8325-1628-4 40.50 EUR

63 Erik Einhaus: Schülerkompetenzen im Bereich Wärmelehre. *Entwicklung eines Testinstruments zur Überprüfung und Weiterentwicklung eines normativen Modells fachbezogener Kompetenzen*
ISBN 978-3-8325-1630-7 40.50 EUR

64 Jasmin Neuroth: Concept Mapping als Lernstrategie. *Eine Interventionsstudie zum Chemielernen aus Texten*
ISBN 978-3-8325-1659-8 40.50 EUR

65 Hans Gerd Hegeler-Burkhart: Zur Kommunikation von Hauptschülerinnen und Hauptschülern in einem handlungsorientierten und fächerübergreifenden Unterricht mit physikalischen und technischen Inhalten
ISBN 978-3-8325-1667-3 40.50 EUR

66 Karsten Rincke: Sprachentwicklung und Fachlernen im Mechanikunterricht. *Sprache und Kommunikation bei der Einführung in den Kraftbegriff*
ISBN 978-3-8325-1699-4 40.50 EUR

67 Nina Strehle: Das Ion im Chemieunterricht. *Alternative Schülervorstellungen und curriculare Konsequenzen*
ISBN 978-3-8325-1710-6 40.50 EUR

68 Martin Hopf: Problemorientierte Schülerexperimente
ISBN 978-3-8325-1711-3 40.50 EUR

69 Anne Beerenwinkel: Fostering conceptual change in chemistry classes using expository texts
ISBN 978-3-8325-1721-2 40.50 EUR

70 Roland Berger: Das Gruppenpuzzle im Physikunterricht der Sekundarstufe II. *Eine empirische Untersuchung auf der Grundlage der Selbstbestimmungstheorie der Motivation*
ISBN 978-3-8325-1732-8 40.50 EUR

71 Giuseppe Colicchia: Physikunterricht im Kontext von Medizin und Biologie. *Entwicklung und Erprobung von Unterrichtseinheiten*
ISBN 978-3-8325-1746-5 40.50 EUR

72 Sandra Winheller: Geschlechtsspezifische Auswirkungen der Lehrer-Schüler-Interaktion im Chemieanfangsunterricht
ISBN 978-3-8325-1757-1 40.50 EUR

73 Isabel Wahser: Training von naturwissenschaftlichen Arbeitsweisen zur Unterstützung experimenteller Kleingruppenarbeit im Fach Chemie
ISBN 978-3-8325-1815-8 40.50 EUR

74 Claus Brell: Lernmedien und Lernerfolg - reale und virtuelle Materialien im Physikunterricht. *Empirische Untersuchungen in achten Klassen an Gymnasien (Laborstudie) zum Computereinsatz mit Simulation und IBE*
ISBN 978-3-8325-1829-5 40.50 EUR

75 Rainer Wackermann: Überprüfung der Wirksamkeit eines Basismodell-Trainings für Physiklehrer
ISBN 978-3-8325-1882-0 40.50 EUR

76 Oliver Tepner: Effektivität von Aufgaben im Chemieunterricht der Sekundarstufe I
ISBN 978-3-8325-1919-3 40.50 EUR

77 Claudia Geyer: Museums- und Science-Center-Besuche im naturwissenschaftlichen Unterricht aus einer motivationalen Perspektive. *Die Sicht von Lehrkräften und Schülerinnen und Schülern*
ISBN 978-3-8325-1922-3 40.50 EUR

78 Tobias Leonhard: Professionalisierung in der Lehrerbildung. *Eine explorative Studie zur Entwicklung professioneller Kompetenzen in der Lehrererstausbildung*
ISBN 978-3-8325-1924-7 40.50 EUR

79 Alexander Kauertz: Schwierigkeitserzeugende Merkmale physikalischer Leistungstestaufgaben
ISBN 978-3-8325-1925-4 40.50 EUR

80 Regina Hübinger: Schüler auf Weltreise. *Entwicklung und Evaluation von Lehr-/ Lernmaterialien zur Förderung experimentell-naturwissenschaftlicher Kompetenzen für die Jahrgangsstufen 5 und 6*
ISBN 978-3-8325-1932-2 40.50 EUR

81 Christine Waltner: Physik lernen im Deutschen Museum
ISBN 978-3-8325-1933-9 40.50 EUR

82 Torsten Fischer: Handlungsmuster von Physiklehrkräften beim Einsatz neuer Medien. *Fallstudien zur Unterrichtspraxis*
ISBN 978-3-8325-1948-3 42.00 EUR

83 Corinna Kieren: Chemiehausaufgaben in der Sekundarstufe I des Gymnasiums. *Fragebogenerhebung zur gegenwärtigen Praxis und Entwicklung eines optimierten Hausaufgabendesigns im Themenbereich Säure-Base*
978-3-8325-1975-9 37.00 EUR

84 Marco Thiele: Modelle der Thermohalinen Zirkulation im Unterricht. *Eine empirische Studie zur Förderung des Modellverständnisses*
ISBN 978-3-8325-1982-7 40.50 EUR

85 Bernd Zinn: Physik lernen, um Physik zu lehren. *Eine Möglichkeit für interessanteren Physikunterricht*
ISBN 978-3-8325-1995-7 39.50 EUR

86 Esther Klaes: Außerschulische Lernorte im naturwissenschaftlichen Unterricht. *Die Perspektive der Lehrkraft*
ISBN 978-3-8325-2006-9 43.00 EUR

87 Marita Schmidt: Kompetenzmodellierung und -diagnostik im Themengebiet Energie der Sekundarstufe I. *Entwicklung und Erprobung eines Testinventars*
ISBN 978-3-8325-2024-3 37.00 EUR

88 Gudrun Franke-Braun: Aufgaben mit gestuften Lernhilfen. *Ein Aufgabenformat zur Förderung der sachbezogenen Kommunikation und Lernleistung für den naturwissenschaftlichen Unterricht*
ISBN 978-3-8325-2026-7 38.00 EUR

89 Silke Klos: Kompetenzförderung im naturwissenschaftlichen Anfangsunterricht. *Der Einfluss eines integrierten Unterrichtskonzepts*
ISBN 978-3-8325-2133-2 37.00 EUR

90 Ulrike Elisabeth Burkard: Quantenphysik in der Schule. *Bestandsaufnahme, Perspektiven und Weiterentwicklungsmöglichkeiten durch die Implementation eines Medienservers*
ISBN 978-3-8325-2215-5 43.00 EUR

91 Ulrike Gromadecki: Argumente in physikalischen Kontexten. *Welche Geltungsgründe halten Physikanfänger für überzeugend?*
ISBN 978-3-8325-2250-6 41.50 EUR

92 Jürgen Bruns: Auf dem Weg zur Förderung naturwissenschaftsspezifischer Vorstellungen von zukünftigen Chemie-Lehrenden
ISBN 978-3-8325-2257-5 43.50 EUR

93 Cornelius Marsch: Räumliche Atomvorstellung. *Entwicklung und Erprobung eines Unterrichtskonzeptes mit Hilfe des Computers*
ISBN 978-3-8325-2293-3 82.50 EUR

94 Maja Brückmann: Sachstrukturen im Physikunterricht. *Ergebnisse einer Videostudie*
ISBN 978-3-8325-2272-8 39.50 EUR

95 Sabine Fechner: Effects of Context-oriented Learning on Student Interest and Achievement in Chemistry Education
ISBN 978-3-8325-2343-5 36.50 EUR

96 Clemens Nagel: eLearning im Physikalischen Anfängerpraktikum
ISBN 978-3-8325-2355-8 39.50 EUR

97 Josef Riese: Professionelles Wissen und professionelle Handlungskompetenz von (angehenden) Physiklehrkräften
ISBN 978-3-8325-2376-3 39.00 EUR

98 Sascha Bernholt: Kompetenzmodellierung in der Chemie. *Theoretische und empirische Reflexion am Beispiel des Modells hierarchischer Komplexität*
ISBN 978-3-8325-2447-0 40.00 EUR

99 Holger Christoph Stawitz: Auswirkung unterschiedlicher Aufgabenprofile auf die Schülerleistung. *Vergleich von Naturwissenschafts- und Problemlöseaufgaben der PISA 2003-Studie*
ISBN 978-3-8325-2451-7 37.50 EUR

100 Hans Ernst Fischer, Elke Sumfleth (Hrsg.): nwu-essen – 10 Jahre Essener Forschung zum naturwissenschaftlichen Unterricht
ISBN 978-3-8325-3331-1 40.00 EUR

101 Hendrik Härtig: Sachstrukturen von Physikschulbüchern als Grundlage zur Bestimmung der Inhaltsvalidität eines Tests
ISBN 978-3-8325-2512-5 34.00 EUR

102 Thomas Grüß-Niehaus: Zum Verständnis des Löslichkeitskonzeptes im Chemieunterricht. *Der Effekt von Methoden progressiver und kollaborativer Reflexion*
ISBN 978-3-8325-2537-8 40.50 EUR

103 Patrick Bronner: Quantenoptische Experimente als Grundlage eines Curriculums zur Quantenphysik des Photons
ISBN 978-3-8325-2540-8 36.00 EUR

104 Adrian Voßkühler: Blickbewegungsmessung an Versuchsaufbauten. *Studien zur Wahrnehmung, Verarbeitung und Usability von physikbezogenen Experimenten am Bildschirm und in der Realität*
ISBN 978-3-8325-2548-4 47.50 EUR

105 Verena Tobias: Newton'sche Mechanik im Anfangsunterricht. *Die Wirksamkeit einer Einführung über die zweidimensionale Dynamik auf das Lehren und Lernen*
ISBN 978-3-8325-2558-3 54.00 EUR

106 Christian Rogge: Entwicklung physikalischer Konzepte in aufgabenbasierten Lernumgebungen
ISBN 978-3-8325-2574-3 45.00 EUR

107 Mathias Ropohl: Modellierung von Schülerkompetenzen im Basiskonzept Chemische Reaktion. *Entwicklung und Analyse von Testaufgaben*
ISBN 978-3-8325-2609-2 36.50 EUR

108 Christoph Kulgemeyer: Physikalische Kommunikationskompetenz. *Modellierung und Diagnostik*
ISBN 978-3-8325-2674-0 44.50 EUR

109 Jennifer Olszewski: The Impact of Physics Teachers' Pedagogical Content Knowledge on Teacher Actions and Student Outcomes
ISBN 978-3-8325-2680-1 33.50 EUR

110 Annika Ohle: Primary School Teachers' Content Knowledge in Physics and its Impact on Teaching and Students' Achievement
ISBN 978-3-8325-2684-9 36.50 EUR

111 Susanne Mannel: Assessing scientific inquiry. *Development and evaluation of a test for the low-performing stage*
ISBN 978-3-8325-2761-7 40.00 EUR

112 Michael Plomer: Physik physiologisch passend praktiziert. *Eine Studie zur Lernwirksamkeit von traditionellen und adressatenspezifischen Physikpraktika für die Physiologie*
ISBN 978-3-8325-2804-1 34.50 EUR

113 Alexandra Schulz: Experimentierspezifische Qualitätsmerkmale im Chemieunterricht. *Eine Videostudie*
ISBN 978-3-8325-2817-1 40.00 EUR

114 Franz Boczianowski: Eine empirische Untersuchung zu Vektoren im Physikunterricht der Mittelstufe
ISBN 978-3-8325-2843-0 39.50 EUR

115 Maria Ploog: Internetbasiertes Lernen durch Textproduktion im Fach Physik
ISBN 978-3-8325-2853-9 39.50 EUR

116 Anja Dhein: Lernen in Explorier- und Experimentiersituationen. *Eine explorative Studie zu Bedeutungsentwicklungsprozessen bei Kindern im Alter zwischen 4 und 6 Jahren*
ISBN 978-3-8325-2859-1 45.50 EUR

117 Irene Neumann: Beyond Physics Content Knowledge. *Modeling Competence Regarding Nature of Scientific Inquiry and Nature of Scientific Knowledge*
ISBN 978-3-8325-2880-5 37.00 EUR

118 Markus Emden: Prozessorientierte Leistungsmessung des naturwissenschaftlich-experimentellen Arbeitens. *Eine vergleichende Studie zu Diagnoseinstrumenten zu Beginn der Sekundarstufe I*
ISBN 978-3-8325-2867-6 38.00 EUR

119 Birgit Hofmann: Analyse von Blickbewegungen von Schülern beim Lesen von physikbezogenen Texten mit Bildern. *Eye Tracking als Methodenwerkzeug in der physikdidaktischen Forschung*
ISBN 978-3-8325-2925-3 59.00 EUR

120 Rebecca Knobloch: Analyse der fachinhaltlichen Qualität von Schüleräußerungen und deren Einfluss auf den Lernerfolg. *Eine Videostudie zu kooperativer Kleingruppenarbeit*
ISBN 978-3-8325-3006-8 36.50 EUR

121 Julia Hostenbach: Entwicklung und Prüfung eines Modells zur Beschreibung der Bewertungskompetenz im Chemieunterricht
ISBN 978-3-8325-3013-6 38.00 EUR

122 Anna Windt: Naturwissenschaftliches Experimentieren im Elementarbereich. *Evaluation verschiedener Lernsituationen*
ISBN 978-3-8325-3020-4 43.50 EUR

123 Eva Kölbach: Kontexteinflüsse beim Lernen mit Lösungsbeispielen
ISBN 978-3-8325-3025-9 38.50 EUR

124 Anna Lau: Passung und vertikale Vernetzung im Chemie- und Physikunterricht
ISBN 978-3-8325-3021-1 36.00 EUR

125 Jan Lamprecht: Ausbildungswege und Komponenten professioneller Handlungskompetenz. *Vergleich von Quereinsteigern mit Lehramtsabsolventen für Gymnasien im Fach Physik*
ISBN 978-3-8325-3035-8 38.50 EUR

126 Ulrike Böhm: Förderung von Verstehensprozessen unter Einsatz von Modellen
ISBN 978-3-8325-3042-6 41.00 EUR

127 Sabrina Dollny: Entwicklung und Evaluation eines Testinstruments zur Erfassung des fachspezifischen Professionswissens von Chemielehrkräften
ISBN 978-3-8325-3046-4 37.00 EUR

128 Monika Zimmermann: Naturwissenschaftliche Bildung im Kindergarten. *Eine integrative Längsschnittstudie zur Kompetenzentwicklung von Erzieherinnen*
ISBN 978-3-8325-3053-2 54.00 EUR

129 Ulf Saballus: Über das Schlussfolgern von Schülerinnen und Schülern zu öffentlichen Kontroversen mit naturwissenschaftlichem Hintergrund. *Eine Fallstudie*
ISBN 978-3-8325-3086-0 39.50 EUR

130 Olaf Krey: Zur Rolle der Mathematik in der Physik. *Wissenschaftstheoretische Aspekte und Vorstellungen Physiklernender*
ISBN 978-3-8325-3101-0 46.00 EUR

131 Angelika Wolf: Zusammenhänge zwischen der Eigenständigkeit im Physikunterricht, der Motivation, den Grundbedürfnissen und dem Lernerfolg von Schülern
ISBN 978-3-8325-3161-4 45.00 EUR

132 Johannes Börlin: Das Experiment als Lerngelegenheit. *Vom interkulturellen Vergleich des Physikunterrichts zu Merkmalen seiner Qualität*
ISBN 978-3-8325-3170-6 45.00 EUR

133 Olaf Uhden: Mathematisches Denken im Physikunterricht. *Theorieentwicklung und Problemanalyse*
ISBN 978-3-8325-3170-6 45.00 EUR

134 Christoph Gut: Modellierung und Messung experimenteller Kompetenz. *Analyse eines large-scale Experimentiertests*
ISBN 978-3-8325-3213-0 40.00 EUR

135 Antonio Rueda: Lernen mit ExploMultimedial in kolumbianischen Schulen. *Analyse von kurzzeitigen Lernprozessen und der Motivation beim länderübergreifenden Einsatz einer deutschen computergestützten multimedialen Lernumgebung für den naturwissenschaftlichen Unterricht*
ISBN 978-3-8325-3218-5 45.50 EUR

136 Krisztina Berger: Bilder, Animationen und Notizen. *Empirische Untersuchung zur Wirkung einfacher visueller Repräsentationen und Notizen auf den Wissenserwerb in der Optik*
ISBN 978-3-8325-3238-3 41.50 EUR

137 Antony Crossley: Untersuchung des Einflusses unterschiedlicher physikalischer Konzepte auf den Wissenserwerb in der Thermodynamik der Sekundarstufe I
ISBN 978-3-8325-3275-8 40.00 EUR

138 Tobias Viering: Entwicklung physikalischer Kompetenz in der Sekundarstufe I. *Validierung eines Kompetenzentwicklungsmodells für das Energiekonzept im Bereich Fachwissen*
ISBN 978-3-8325-3277-2 37.00 EUR

139 Nico Schreiber: Diagnostik experimenteller Kompetenz. *Validierung technologiegestützter Testverfahren im Rahmen eines Kompetenzstrukturmodells*
ISBN 978-3-8325-3284-0 39.00 EUR

140 Sarah Hundertmark: Einblicke in kollaborative Lernprozesse. *Eine Fallstudie zur reflektierenden Zusammenarbeit unterstützt durch die Methoden Concept Mapping und Lernbegleitbogen*
ISBN 978-3-8325-3251-2 43.00 EUR

141 Ronny Scherer: Analyse der Struktur, Messinvarianz und Ausprägung komplexer Problemlösekompetenz im Fach Chemie. *Eine Querschnittstudie in der Sekundarstufe I und am Übergang zur Sekundarstufe II*
ISBN 978-3-8325-3312-0 43.00 EUR

142 Patricia Heitmann: Bewertungskompetenz im Rahmen naturwissenschaftlicher Problemlöseprozesse. *Modellierung und Diagnose der Kompetenzen Bewertung und analytisches Problemlösen für das Fach Chemie*
ISBN 978-3-8325-3314-4 37.00 EUR

143 Jan Fleischhauer: Wissenschaftliches Argumentieren und Entwicklung von Konzepten beim Lernen von Physik
ISBN 978-3-8325-3325-0 35.00 EUR

144 Nermin Özcan: Zum Einfluss der Fachsprache auf die Leistung im Fach Chemie. *Eine Förderstudie zur Fachsprache im Chemieunterricht*
ISBN 978-3-8325-3328-1 36.50 EUR

145 Helena van Vorst: Kontextmerkmale und ihr Einfluss auf das Schülerinteresse im Fach Chemie
ISBN 978-3-8325-3321-2 38.50 EUR

146 Janine Cappell: Fachspezifische Diagnosekompetenz angehender Physiklehrkräfte in der ersten Ausbildungsphase
ISBN 978-3-8325-3356-4 38.50 EUR

147 Susanne Bley: Förderung von Transferprozessen im Chemieunterricht
ISBN 978-3-8325-3407-3 40.50 EUR

148 Cathrin Blaes: Die übungsgestützte Lehrerpräsentation im Chemieunterricht der Sekundarstufe I. *Evaluation der Effektivität*
ISBN 978-3-8325-3409-7 43.50 EUR

149 Julia Suckut: Die Wirksamkeit von piko-OWL als Lehrerfortbildung. Eine Evaluation zum Projekt *Physik im Kontext* in Fallstudien
ISBN 978-3-8325-3440-0 45.00 EUR

150 Alexandra Dorschu: Die Wirkung von Kontexten in Physikkompetenztestaufgaben
ISBN 978-3-8325-3446-2 37.00 EUR

151 Jochen Scheid: Multiple Repräsentationen, Verständnis physikalischer Experimente und kognitive Aktivierung: *Ein Beitrag zur Entwicklung der Aufgabenkultur*
ISBN 978-3-8325-3449-3 49.00 EUR

152 Tim Plasa: Die Wahrnehmung von Schülerlaboren und Schülerforschungszentren
ISBN 978-3-8325-3483-7 35.50 EUR

153 Felix Schoppmeier: Physikkompetenz in der gymnasialen Oberstufe.*Entwicklung und Validierung eines Kompetenzstrukturmodells für den Kompetenzbereich Umgang mit Fachwissen*
ISBN 978-3-8325-3502-5 36.00 EUR

154 Katharina Groß: Experimente alternativ dokumentieren. *Eine qualitative Studie zur Förderung der Diagnose- und Differenzierungskompetenz in der Chemielehrerbildung*
ISBN 978-3-8325-3508-7 43.50 EUR

155 Barbara Hank: Konzeptwandelprozesse im Anfangsunterricht Chemie. *Eine quasiexperimentelle Längsschnittstudie*
ISBN 978-3-8325-3519-3 38.50 EUR

156 Katja Freyer: Zum Einfluss von Studieneingangsvoraussetzungen auf den Studienerfolg Erstsemesterstudierender im Fach Chemie
ISBN 978-3-8325-3544-5 38.00 EUR

157 Alexander Rachel: Auswirkungen instruktionaler Hilfen bei der Einführung des (Ferro-)Magnetismus. *Eine Vergleichsstudie in der Primar- und Sekundarstufe*
ISBN 978-3-8325-3548-3 43.50 EUR

158 Sebastian Ritter: Einfluss des Lerninhalts Nanogrößeneffekte auf Teilchen- und Teilchenmodellvorstellungen von Schülerinnen und Schülern
ISBN 978-3-8325-3558-2 36.00 EUR

159 Andrea Harbach: Problemorientierung und Vernetzung in kontextbasierten Lernaufgaben
ISBN 978-3-8325-3564-3 39.00 EUR

160 David Obst: Interaktive Tafeln im Physikunterricht. *Entwicklung und Evaluation einer Lehrerfortbildung*
ISBN 978-3-8325-3582-7 40.50 EUR

161 Sophie Kirschner: Modellierung und Analyse des Professionswissens von Physiklehrkräften
ISBN 978-3-8325-3601-5 35.00 EUR

162 Katja Stief: Selbstregulationsprozesse und Hausaufgabenmotivation im Chemieunterricht
ISBN 978-3-8325-3631-2 34.00 EUR

163 Nicola Meschede: Professionelle Wahrnehmung der inhaltlichen Strukturierung im naturwissenschaftlichen Grundschulunterricht. *Theoretische Beschreibung und empirische Erfassung*
ISBN 978-3-8325-3668-8 37.00 EUR

164 Johannes Maximilian Barth: Experimentieren im Physikunterricht der gymnasialen Oberstufe. *Eine Rekonstruktion übergeordneter Einbettungsstrategien*
ISBN 978-3-8325-3681-7 39.00 EUR

165 Sandra Lein: Das Betriebspraktikum in der Lehrerbildung. *Eine Untersuchung zur Förderung der Wissenschafts- und Technikbildung im allgemeinbildenden Unterricht*
ISBN 978-3-8325-3698-5 40.00 EUR

166 Veranika Maiseyenka: Modellbasiertes Experimentieren im Unterricht. *Praxistauglichkeit und Lernwirkungen*
ISBN 978-3-8325-3708-1 38.00 EUR

167 Christoph Stolzenberger: Der Einfluss der didaktischen Lernumgebung auf das Erreichen geforderter Bildungsziele am Beispiel der W- und P-Seminare im Fach Physik
ISBN 978-3-8325-3708-1 38.00 EUR

168 Pia Altenburger: Mehrebenenregressionsanalysen zum Physiklernen im Sachunterricht der Primarstufe. *Ergebnisse einer Evaluationsstudie.*
ISBN 978-3-8325-3717-3 37.50 EUR

169 Nora Ferber: Entwicklung und Validierung eines Testinstruments zur Erfassung von Kompetenzentwicklung im Fach Chemie in der Sekundarstufe I
ISBN 978-3-8325-3727-2 39.50 EUR

170 Anita Stender: Unterrichtsplanung: Vom Wissen zum Handeln.
Theoretische Entwicklung und empirische Überprüfung des Transformationsmodells der Unterrichtsplanung
ISBN 978-3-8325-3750-0 41.50 EUR

171 Jenna Koenen: Entwicklung und Evaluation von experimentunterstützten Lösungsbeispielen zur Förderung naturwissenschaftlich-experimenteller Arbeitsweisen
ISBN 978-3-8325-3785-2 43.00 EUR

172 Teresa Henning: Empirische Untersuchung kontextorientierter Lernumgebungen in der Hochschuldidaktik. *Entwicklung und Evaluation kontextorientierter Aufgaben in der Studieneingangsphase für Fach- und Nebenfachstudierende der Physik*
ISBN 978-3-8325-3801-9 43.00 EUR

173 Alexander Pusch: Fachspezifische Instrumente zur Diagnose und individuellen Förderung von Lehramtsstudierenden der Physik
ISBN 978-3-8325-3829-3 38.00 EUR

174 Christoph Vogelsang: Validierung eines Instruments zur Erfassung der professionellen Handlungskompetenz von (angehenden) Physiklehrkräften. *Zusammenhangsanalysen zwischen Lehrerkompetenz und Lehrerperformanz*
ISBN 978-3-8325-3846-0 50.50 EUR

175 Ingo Brebeck: Selbstreguliertes Lernen in der Studieneingangsphase im Fach Chemie
ISBN 978-3-8325-3859-0 37.00 EUR

176 Axel Eghtessad: Merkmale und Strukturen von Professionalisierungsprozessen in der ersten und zweiten Phase der Chemielehrerbildung. *Eine empirisch-qualitative Studie mit niedersächsischen Fachleiter_innen der Sekundarstufenlehrämter*
ISBN 978-3-8325-3861-3 45.00 EUR

177 Andreas Nehring: Wissenschaftliche Denk- und Arbeitsweisen im Fach Chemie. Eine kompetenzorientierte Modell- und Testentwicklung für den Bereich der Erkenntnisgewinnung
ISBN 978-3-8325-3872-9 39.50 EUR

178 Maike Schmidt: Professionswissen von Sachunterrichtslehrkräften. Zusammenhangsanalyse zur Wirkung von Ausbildungshintergrund und Unterrichtserfahrung auf das fachspezifische Professionswissen im Unterrichtsinhalt „Verbrennung“
ISBN 978-3-8325-3907-8 38.50 EUR

179 Jan Winkelmann: Auswirkungen auf den Fachwissenszuwachs und auf affektive Schülermerkmale durch Schüler- und Demonstrationsexperimente im Physikunterricht
ISBN 978-3-8325-3915-3 41.00 EUR

180 Iwen Kobow: Entwicklung und Validierung eines Testinstrumentes zur Erfassung der Kommunikationskompetenz im Fach Chemie
ISBN 978-3-8325-3927-6 34.50 EUR

181 Yvonne Gramzow: Fachdidaktisches Wissen von Lehramtsstudierenden im Fach Physik. Modellierung und Testkonstruktion
ISBN 978-3-8325-3931-3 42.50 EUR

182 Evelin Schröter: Entwicklung der Kompetenzerwartung durch Lösen physikalischer Aufgaben einer multimedialen Lernumgebung
ISBN 978-3-8325-3975-7 54.50 EUR

183 Inga Kallweit: Effektivität des Einsatzes von Selbsteinschätzungsbögen im Chemieunterricht der Sekundarstufe I. *Individuelle Förderung durch selbstreguliertes Lernen*
ISBN 978-3-8325-3965-8 44.00 EUR

184 Andrea Schumacher: Paving the way towards authentic chemistry teaching. *A contribution to teachers' professional development*
ISBN 978-3-8325-3976-4 48.50 EUR

185 David Woitkowski: Fachliches Wissen Physik in der Hochschulausbildung. *Konzeptualisierung, Messung, Niveaubildung*
ISBN 978-3-8325-3988-7 53.00 EUR

186 Marianne Korner: Cross-Age Peer Tutoring in Physik. *Evaluation einer Unterrichtsmethode*
ISBN 978-3-8325-3979-5 38.50 EUR

187 Simone Nakoinz: Untersuchung zur Verknüpfung submikroskopischer und makroskopischer Konzepte im Fach Chemie
ISBN 978-3-8325-4057-9 38.50 EUR

188 Sandra Anus: Evaluation individueller Förderung im Chemieunterricht.*Adaptivität von Lerninhalten an das Vorwissen von Lernenden am Beispiel des Basiskonzeptes Chemische Reaktion*
ISBN 978-3-8325-4059-3 43.50 EUR

189 Thomas Roßbegalle: Fachdidaktische Entwicklungsforschung zum besseren Verständnis atmosphärischer Phänomene. *Treibhauseffekt, saurer Regen und stratosphärischer Ozonabbau als Kontexte zur Vermittlung von Basiskonzepten der Chemie*
ISBN 978-3-8325-4059-3 45.50 EUR

190 Kathrin Steckenmesser-Sander: Gemeinsamkeiten und Unterschiede physikbezogener Handlungs-, Denk- und Lernprozesse von Mädchen und Jungen
ISBN 978-3-8325-4066-1 38.50 EUR

191 Cornelia Geller: Lernprozessorientierte Sequenzierung des Physikunterrichts im Zusammenhang mit Fachwissenserwerb. *Eine Videostudie in Finnland, Deutschland und der Schweiz*
ISBN 978-3-8325-4082-1 35.50 EUR

192 Jan Hofmann: Untersuchung des Kompetenzaufbaus von Physiklehrkräften während einer Fortbildungsmaßnahme
ISBN 978-3-8325-4104-0 38.50 EUR

193 Andreas Dickhäuser: Chemiespezifischer Humor. *Theoriebildung, Materialentwicklung, Evaluation*
ISBN 978-3-8325-4108-8 37.00 EUR

194 Stefan Korte: Die Grenzen der Naturwissenschaft als Thema des Physikunterrichts
ISBN 978-3-8325-4112-5 57.50 EUR

195 Carolin Hülsmann: Kurswahlmotive im Fach Chemie. Eine Studie zum Wahlverhalten und Erfolg von Schülerinnen und Schülern in der gymnasialen Oberstufe
ISBN 978-3-8325-4144-6 49.00 EUR

196 Caroline Körbs: Mindeststandards im Fach Chemie am Ende der Pflichtschulzeit
ISBN 978-3-8325-4148-4 34.00 EUR

197 Andreas Vorholzer: Wie lassen sich Kompetenzen des experimentellen Denkens und Arbeitens fördern? *Eine empirische Untersuchung der Wirkung eines expliziten und eines impliziten Instruktionsansatzes*
ISBN 978-3-8325-4194-1 37.50 EUR

198 Anna Katharina Schmitt: Entwicklung und Evaluation einer Chemielehrerfortbildung zum Kompetenzbereich Erkenntnisgewinnung
ISBN 978-3-8325-4228-3 39.50 EUR

199 Christian Maurer: Strukturierung von Lehr-Lern-Sequenzen
ISBN 978-3-8325-4247-4 36.50 EUR

200 Helmut Fischler, Elke Sumfleth (Hrsg.): Professionelle Kompetenz von Lehrkräften der Chemie und Physik
ISBN 978-3-8325-4523-9 34.00 EUR

201 Simon Zander: Lehrerfortbildung zu Basismodellen und Zusammenhänge zum Fachwissen
ISBN 978-3-8325-4248-1 35.00 EUR

202 Kerstin Arndt: Experimentierkompetenz erfassen.
Analyse von Prozessen und Mustern am Beispiel von Lehramtsstudierenden der Chemie
ISBN 978-3-8325-4266-5 45.00 EUR

203 Christian Lang: Kompetenzorientierung im Rahmen experimentalchemischer Praktika
ISBN 978-3-8325-4268-9 42.50 EUR

204 Eva Cauet: Testen wir relevantes Wissen? *Zusammenhang zwischen dem Professionswissen von Physiklehrkräften und gutem und erfolgreichem Unterrichten*
ISBN 978-3-8325-4276-4 39.50 EUR

205 Patrick Löffler: Modellanwendung in Problemlöseaufgaben. *Wie wirkt Kontext?*
ISBN 978-3-8325-4303-7 35.00 EUR

206 Carina Gehlen: Kompetenzstruktur naturwissenschaftlicher Erkenntnisgewinnung im Fach Chemie
ISBN 978-3-8325-4318-1 43.00 EUR

207 Lars Oettinghaus: Lehrerüberzeugungen und physikbezogenes Professionswissen. *Vergleich von Absolventinnen und Absolventen verschiedener Ausbildungswege im Physikreferendariat*
ISBN 978-3-8325-4319-8 38.50 EUR

208 Jennifer Petersen: Zum Einfluss des Merkmals Humor auf die Gesundheitsförderung im Chemieunterricht der Sekundarstufe I.
Eine Interventionsstudie zum Thema Sonnenschutz
ISBN 978-3-8325-4348-8 40.00 EUR

209 Philipp Straube: Modellierung und Erfassung von Kompetenzen naturwissenschaftlicher Erkenntnisgewinnung bei (Lehramts-) Studierenden im Fach Physik
ISBN 978-3-8325-4351-8 35.50 EUR

210 Martin Dickmann: Messung von Experimentierfähigkeiten. *Validierungsstudien zur Qualität eines computerbasierten Testverfahrens*
ISBN 978-3-8325-4356-3 41.00 EUR

211 Markus Bohlmann: Science Education. Empirie, Kulturen und Mechanismen der Didaktik der Naturwissenschaften
ISBN 978-3-8325-4377-8 44.00 EUR

212 Martin Draude: Die Kompetenz von Physiklehrkräften, Schwierigkeiten von Schülerinnen und Schülern beim eigenständigen Experimentieren zu diagnostizieren
ISBN 978-3-8325-4382-2 37.50 EUR

213 Henning Rode: Prototypen evidenzbasierten Physikunterrichts. *Zwei empirische Studien zum Einsatz von Feedback und Blackboxes in der Sekundarstufe*
ISBN 978-3-8325-4389-1 42.00 EUR

214 Jan-Henrik Kechel: Schülerschwierigkeiten beim eigenständigen Experimentieren. *Eine qualitative Studie am Beispiel einer Experimentieraufgabe zum Hooke'schen Gesetz*
ISBN 978-3-8325-4392-1 55.00 EUR

215 Katharina Fricke: Classroom Management and its Impact on Lesson Outcomes in Physics. *A multi-perspective comparison of teaching practices in primary and secondary schools*
ISBN 978-3-8325-4394-5 40.00 EUR

216 Hannes Sander: Orientierungen von Jugendlichen beim Urteilen und Entscheiden in Kontexten nachhaltiger Entwicklung. *Eine rekonstruktive Perspektive auf Bewertungskompetenz in der Didaktik der Naturwissenschaft*
ISBN 978-3-8325-4434-8 46.00 EUR

217 Inka Haak: Maßnahmen zur Unterstützung kognitiver und metakognitiver Prozesse in der Studieneingangsphase. *Eine Design-Based-Research-Studie zum universitären Lernzentrum Physiktreff*
ISBN 978-3-8325-4437-9 46.50 EUR

218 Martina Brandenburger: Was beeinflusst den Erfolg beim Problemlösen in der Physik? *Eine Untersuchung mit Studierenden*
ISBN 978-3-8325-4409-6 42.50 EUR

219 Corinna Helms: Entwicklung und Evaluation eines Trainings zur Verbesserung der Erklärqualität von Schülerinnen und Schülern im Gruppenpuzzle
ISBN 978-3-8325-4454-6 42.50 EUR

220 Viktoria Rath: Diagnostische Kompetenz von angehenden Physiklehrkräften. *Modellierung, Testinstrumentenentwicklung und Erhebung der Performanz bei der Diagnose von Schülervorstellungen in der Mechanik*
ISBN 978-3-8325-4456-0 42.50 EUR

221 Janne Krüger: Schülerperspektiven auf die zeitliche Entwicklung der Naturwissenschaften
ISBN 978-3-8325-4457-7 45.50 EUR

222 Stefan Mutke: Das Professionswissen von Chemiereferendarinnen und -referendaren in Nordrhein-Westfalen. *Eine Längsschnittstudie*
ISBN 978-3-8325-4458-4 37.50 EUR

223 Sebastian Habig: Systematisch variierte Kontextaufgaben und ihr Einfluss auf kognitive und affektive Schülerfaktoren
ISBN 978-3-8325-4467-6 40.50 EUR

224 Sven Liepertz: Zusammenhang zwischen dem Professionswissen von Physiklehrkräften, dem sachstrukturellen Angebot des Unterrichts und der Schülerleistung
ISBN 978-3-8325-4480-5 34.00 EUR

225 Elina Platova: Optimierung eines Laborpraktikums durch kognitive Aktivierung
ISBN 978-3-8325-4481-2 39.00 EUR

226 Tim Reschke: Lesegeschichten im Chemieunterricht der Sekundarstufe I zur Unterstützung von situationalem Interesse und Lernerfolg
ISBN 978-3-8325-4487-4 41.00 EUR

227 Lena Mareike Walper: Entwicklung der physikbezogenen Interessen und selbstbezogenen Kognitionen von Schülerinnen und Schülern in der Übergangsphase von der Primar- in die Sekundarstufe. *Eine Längsschnittanalyse vom vierten bis zum siebten Schuljahr*
ISBN 978-3-8325-4495-9 43.00 EUR

228 Stefan Anthofer: Förderung des fachspezifischen Professionswissens von Chemielehramtsstudierenden
ISBN 978-3-8325-4498-0 39.50 EUR

229 Marcel Bullinger: Handlungsorientiertes Physiklernen mit instruierten Selbsterklärungen in der Primarstufe. *Eine experimentelle Laborstudie*
ISBN 978-3-8325-4504-8 44.00 EUR

230 Thomas Amenda: Bedeutung fachlicher Elementarisierungen für das Verständnis der Kinematik
ISBN 978-3-8325-4531-4 43.50 EUR

231 Sabrina Milke: Beeinflusst *Priming* das Physiklernen?
Eine empirische Studie zum Dritten Newtonschen Axiom
ISBN 978-3-8325-4549-4 42.00 EUR

232 Corinna Erfmann: Ein anschaulicher Weg zum Verständnis der elektromagnetischen Induktion. *Evaluation eines Unterrichtsvorschlags und Validierung eines Leistungsdiagnoseinstruments*
ISBN 978-3-8325-4550-5 49.50 EUR

233 Hanne Rautenstrauch: Erhebung des (Fach-)Sprachstandes bei Lehramtsstudierenden im Kontext des Faches Chemie
ISBN 978-3-8325-4556-7 40.50 EUR

234 Tobias Klug: Wirkung kontextorientierter physikalischer Praktikumsversuche auf Lernprozesse von Studierenden der Medizin
ISBN 978-3-8325-4558-1 37.00 EUR

235 Mareike Bohrmann: Zur Förderung des Verständnisses der Variablenkontrolle im naturwissenschaftlichen Sachunterricht
ISBN 978-3-8325-4559-8 52.00 EUR

236 Anja Schödl: FALKO-Physik – Fachspezifische Lehrerkompetenzen im Fach Physik. *Entwicklung und Validierung eines Testinstruments zur Erfassung des fachspezifischen Professionswissens von Physiklehrkräften*
ISBN 978-3-8325-4553-6 40.50 EUR

237 Hilda Scheuermann: Entwicklung und Evaluation von Unterstützungsmaßnahmen zur Förderung der Variablenkontrollstrategie beim Planen von Experimenten
ISBN 978-3-8325-4568-0 39.00 EUR

238 Christian G. Strippel: Naturwissenschaftliche Erkenntnisgewinnung an chemischen Inhalten vermitteln. *Konzeption und empirische Untersuchung einer Ausstellung mit Experimentierstation*
ISBN 978-3-8325-4577-2 41.50 EUR

239 Sarah Rau: Durchführung von Sachunterricht im Vorbereitungsdienst. *Eine längsschnittliche, videobasierte Unterrichtsanalyse*
ISBN 978-3-8325-4579-6 46.00 EUR

240 Thomas Plotz: Lernprozesse zu nicht-sichtbarer Strahlung. *Empirische Untersuchungen in der Sekundarstufe 2*
ISBN 978-3-8325-4624-3 39.50 EUR

241 Wolfgang Aschauer: Elektrische und magnetische Felder. *Eine empirische Studie zu Lernprozessen in der Sekundarstufe II*
ISBN 978-3-8325-4625-0 50.00 EUR

242 Anna Donhauser: Didaktisch rekonstruierte Materialwissenschaft. *Aufbau und Konzeption eines Schülerlabors für den Exzellenzcluster Engineering of Advanced Materials*
ISBN 978-3-8325-4636-6 39.00 EUR

243 Katrin Schüßler: Lernen mit Lösungsbeispielen im Chemieunterricht. *Einflüsse auf Lernerfolg, kognitive Belastung und Motivation*
ISBN 978-3-8325-4640-3 42.50 EUR

244 Timo Fleischer: Untersuchung der chemischen Fachsprache unter besonderer Berücksichtigung chemischer Repräsentationen
ISBN 978-3-8325-4642-7 46.50 EUR

245 Rosina Steininger: Concept Cartoons als Stimuli für Kleingruppendiskussionen im Chemieunterricht. *Beschreibung und Analyse einer komplexen Lerngelegenheit*
ISBN 978-3-8325-4647-2 39.00 EUR

246 Daniel Rehfeldt: Erfassung der Lehrqualität naturwissenschaftlicher Experimentalpraktika
ISBN 978-3-8325-4590-1 40.00 EUR

247 Sandra Puddu: Implementing Inquiry-based Learning in a Diverse Classroom: Investigating Strategies of Scaffolding and Students' Views of Scientific Inquiry
ISBN 978-3-8325-4591-8 35.50 EUR

248 Markus Bliersbach: Kreativität in der Chemie. *Erhebung und Förderung der Vorstellungen von Chemielehramtsstudierenden*
ISBN 978-3-8325-4593-2 44.00 EUR

249 Lennart Kimpel: Aufgaben in der Allgemeinen Chemie. *Zum Zusammenspiel von chemischem Verständnis und Rechenfähigkeit*
ISBN 978-3-8325-4618-2 36.00 EUR

250 Louise Bindel: Effects of integrated learning: explicating a mathematical concept in inquiry-based science camps
ISBN 978-3-8325-4655-7 37.50 EUR

251 Michael Wenzel: Computereinsatz in Schule und Schülerlabor. *Einstellung von Physiklehrkräften zu Neuen Medien*
ISBN 978-3-8325-4659-5 38.50 EUR

252 Laura Muth: Einfluss der Auswertephase von Experimenten im Physikunterricht. *Ergebnisse einer Interventionsstudie zum Zuwachs von Fachwissen und experimenteller Kompetenz von Schülerinnen und Schülern*
ISBN 978-3-8325-4675-5 36.50 EUR

253 Annika Fricke: Interaktive Skripte im Physikalischen Praktikum. *Entwicklung und Evaluation von Hypermedien für die Nebenfachausbildung*
ISBN 978-3-8325-4676-2 41.00 EUR

254 Julia Haase: Selbstbestimmtes Lernen im naturwissenschaftlichen Sachunterricht. *Eine empirische Interventionsstudie mit Fokus auf Feedback und Kompetenzerleben*
ISBN 978-3-8325-4685-4 38.50 EUR

255 Antje J. Heine: Was ist Theoretische Physik? *Eine wissenschaftstheoretische Betrachtung und Rekonstruktion von Vorstellungen von Studierenden und Dozenten über das Wesen der Theoretischen Physik*
ISBN 978-3-8325-4691-5 46.50 EUR

256 Claudia Meinhardt: Entwicklung und Validierung eines Testinstruments zu Selbstwirksamkeitserwartungen von (angehenden) Physiklehrkräften in physikdidaktischen Handlungsfeldern
ISBN 978-3-8325-4712-7 47.00 EUR

257 Ann-Kathrin Schlüter: Professionalisierung angehender Chemielehrkräfte für einen Gemeinsamen Unterricht
ISBN 978-3-8325-4713-4 53.50 EUR

258 Stefan Richtberg: Elektronenbahnen in Feldern. Konzeption und Evaluation einer webbasierten Lernumgebung
ISBN 978-3-8325-4723-3 49.00 EUR

259 Jan-Philipp Burde: Konzeption und Evaluation eines Unterrichtskonzepts zu einfachen Stromkreisen auf Basis des Elektronengasmodells
ISBN 978-3-8325-4726-4 57.50 EUR

260 Frank Finkenberg: Flipped Classroom im Physikunterricht
ISBN 978-3-8325-4737-4 42.50 EUR

261 Florian Treisch: Die Entwicklung der Professionellen Unterrichtswahrnehmung im Lehr-Lern-Labor Seminar
ISBN 978-3-8325-4741-4 41.50 EUR

262 Desiree Mayr: Strukturiertheit des experimentellen naturwissenschaftlichen Problemlöseprozesses
ISBN 978-3-8325-4757-8 37.00 EUR

263 Katrin Weber: Entwicklung und Validierung einer Learning Progression für das Konzept der chemischen Reaktion in der Sekundarstufe I
ISBN 978-3-8325-4762-2 48.50 EUR

264 Hauke Bartels: Entwicklung und Bewertung eines performanznahen Videovignettentests zur Messung der Erklärfähigkeit von Physiklehrkräften
ISBN 978-3-8325-4804-9 37.00 EUR

265 Karl Marniok: Zum Wesen von Theorien und Gesetzen in der Chemie. *Begriffsanalyse und Förderung der Vorstellungen von Lehramtsstudierenden*
ISBN 978-3-8325-4805-6 42.00 EUR

266 Marisa Holzapfel: Fachspezifischer Humor als Methode in der Gesundheitsbildung im Übergang von der Primarstufe zur Sekundarstufe I
ISBN 978-3-8325-4808-7 50.00 EUR

267 Anna Stolz: Die Auswirkungen von Experimentiersituationen mit unterschiedlichem Öffnungsgrad auf Leistung und Motivation der Schülerinnen und Schüler
ISBN 978-3-8325-4781-3 38.00 EUR

268 Nina Ulrich: Interaktive Lernaufgaben in dem digitalen Schulbuch eChemBook. *Einfluss des Interaktivitätsgrads der Lernaufgaben und des Vorwissens der Lernenden auf den Lernerfolg*
ISBN 978-3-8325-4814-8 43.50 EUR

269 Kim-Alessandro Weber: Quantenoptik in der Lehrerfortbildung. *Ein bedarfsgeprägtes Fortbildungskonzept zum Quantenobjekt Photon mit Realexperimenten*
ISBN 978-3-8325-4792-9 55.00 EUR

270 Nina Skorsetz: Empathisierer und Systematisierer im Vorschulalter. *Eine Fragebogen- und Videostudie zur Motivation, sich mit Naturphänomenen zu beschäftigen*
ISBN 978-3-8325-4825-4 43.50 EUR

271 Franziska Kehne: Analyse des Transfers von kontextualisiert erworbenem Wissen im Fach Chemie
ISBN 978-3-8325-4846-9 45.00 EUR

272 Markus Elsholz: Das akademische Selbstkonzept angehender Physiklehrkräfte als Teil ihrer professionellen Identität. *Dimensionalität und Veränderung während einer zentralen Praxisphase*
ISBN 978-3-8325-4857-5 37.50 EUR

273 Joachim Müller: Studienerfolg in der Physik. *Zusammenhang zwischen Modellierungskompetenz und Studienerfolg*
ISBN 978-3-8325-4859-9 35.00 EUR

274 Jennifer Dörschelln: Organische Leuchtdioden. *Implementation eines innovativen Themas in den Chemieunterricht*
ISBN 978-3-8325-4865-0 59.00 EUR

275 Stephanie Strelow: Beliefs von Studienanfängern des Kombi-Bachelors Physik über die Natur der Naturwissenschaften
ISBN 978-3-8325-4881-0 40.50 EUR

276 Dennis Jaeger: Kognitive Belastung und aufgabenspezifische sowie personenspezifische Einflussfaktoren beim Lösen von Physikaufgaben
ISBN 978-3-8325-4928-2 50.50 EUR

277 Vanessa Fischer: Der Einfluss von Interesse und Motivation auf die Messung von Fach- und Bewertungskompetenz im Fach Chemie
ISBN 978-3-8325-4933-6 39.00 EUR

278 René Dohrmann: Professionsbezogene Wirkungen einer Lehr-Lern-Labor-Veranstaltung. *Eine multimethodische Studie zu den professionsbezogenen Wirkungen einer Lehr-Lern-Labor-Blockveranstaltung auf Studierende der Bachelorstudiengänge Lehramt Physik und Grundschulpädagogik (Sachunterricht)*
ISBN 978-3-8325-4958-9 40.00 EUR

279 Meike Bergs: Can We Make Them Use These Strategies? *Fostering Inquiry-Based Science Learning Skills with Physical and Virtual Experimentation Environments*
ISBN 978-3-8325-4962-6 39.50 EUR

280 Marie-Therese Hauerstein: Untersuchung zur Effektivität von Strukturierung und Binnendifferenzierung im Chemieunterricht der Sekundarstufe I. *Evaluation der Strukturierungshilfe Lernleiter*
ISBN 978-3-8325-4982-4 42.50 EUR

281 Verena Zucker: Erkennen und Beschreiben von formativem Assessment im naturwissenschaftlichen Grundschulunterricht. *Entwicklung eines Instruments zur Erfassung von Teilfähigkeiten der professionellen Wahrnehmung von Lehramtsstudierenden*
ISBN 978-3-8325-4991-6 38.00 EUR

282 Victoria Telser: Erfassung und Förderung experimenteller Kompetenz von Lehrkräften im Fach Chemie
ISBN 978-3-8325-4996-1 50.50 EUR

283 Kristine Tschirschky: Entwicklung und Evaluation eines gedächtnisorientierten Aufgabendesigns für Physikaufgaben
ISBN 978-3-8325-5002-8 42.50 EUR

284 Thomas Elert: Course Success in the Undergraduate General Chemistry Lab
ISBN 978-3-8325-5004-2 41.50 EUR

285 Britta Kalthoff: Explizit oder implizit? *Untersuchung der Lernwirksamkeit verschiedener fachmethodischer Instruktionen im Hinblick auf fachmethodische und fachinhaltliche Fähigkeiten von Sachunterrichtsstudierenden*
ISBN 978-3-8325-5013-4 37.50 EUR

286 Thomas Dickmann: Visuelles Modellverständnis und Studienerfolg in der Chemie. *Zwei Seiten einer Medaille*
ISBN 978-3-8325-5016-5 44.00 EUR

287 Markus Sebastian Feser: Physiklehrkräfte korrigieren Schülertexte. *Eine Explorationsstudie zur fachlich-konzeptuellen und sprachlichen Leistungsfeststellung und -beurteilung im Physikunterricht*
ISBN 978-3-8325-5020-2 49.00 EUR

288 Matylda Dudzinska: Lernen mit Beispielaufgaben und Feedback im Physikunterricht der Sekundarstufe 1. *Energieerhaltung zur Lösung von Aufgaben nutzen*
ISBN 978-3-8325-5025-7 47.00 EUR

289 Ines Sonnenschein: Naturwissenschaftliche Denk- und Arbeitsprozesse Studierender im Labor
ISBN 978-3-8325-5033-2 52.00 EUR

290 Florian Simon: Der Einfluss von Betreuung und Betreuenden auf die Wirksamkeit von Schülerlaborbesuchen. *Eine Zusammenhangsanalyse von Betreuungsqualität, Betreuermerkmalen und Schülerlaborzielen sowie Replikationsstudie zur Wirksamkeit von Schülerlaborbesuchen*
ISBN 978-3-8325-5036-3 49.50 EUR

291 Marie-Annette Geyer: Physikalisch-mathematische Darstellungswechsel funktionaler Zusammenhänge. *Das Vorgehen von SchülerInnen der Sekundarstufe 1 und ihre Schwierigkeiten*
ISBN 978-3-8325-5047-9 46.50 EUR

292 Susanne Digel: Messung von Modellierungskompetenz in Physik. *Theoretische Herleitung und empirische Prüfung eines Kompetenzmodells physikspezifischer Modellierungskompetenz*
ISBN 978-3-8325-5055-4 41.00 EUR

293 Sönke Janssen: Angebots-Nutzungs-Prozesse eines Schülerlabors analysieren und gestalten. *Ein design-based research Projekt*
ISBN 978-3-8325-5065-3 57.50 EUR

294 Knut Wille: Der Productive Failure Ansatz als Beitrag zur Weiterentwicklung der Aufgabenkultur
ISBN 978-3-8325-5074-5 49.00 EUR

295 Lisanne Kraeva: Problemlösestrategien von Schülerinnen und Schülern diagnostizieren
ISBN 978-3-8325-5110-0 59.50 EUR

296 Jenny Lorentzen: Entwicklung und Evaluation eines Lernangebots im Lehramtsstudium Chemie zur Förderung von Vernetzungen innerhalb des fachbezogenen Professionswissens
ISBN 978-3-8325-5120-9 39.50 EUR

297 Micha Winkelmann: Lernprozesse in einem Schülerlabor unter Berücksichtigung individueller naturwissenschaftlicher Interessenstrukturen
ISBN 978-3-8325-5147-6 48.50 EUR

298 Carina Wöhlke: Entwicklung und Validierung eines Instruments zur Erfassung der professionellen Unterrichtswahrnehmung angehender Physiklehrkräfte
ISBN 978-3-8325-5149-0 43.00 EUR

299 Thomas Schubatzky: Das Amalgam Anfangs-Elektrizitätslehreunterricht. *Eine multiperspektivische Betrachtung in Deutschland und Österreich*
ISBN 978-3-8325-5159-9 50.50 EUR

300 Amany Annaggar: A Design Framework for Video Game-Based Gamification Elements to Assess Problem-solving Competence in Chemistry Education
ISBN 978-3-8325-5150-6 52.00 EUR

301 Alexander Engl: Chemie Pur – Unterrichten in der Natur: *Entwicklung und Evaluation eines kontextorientierten Unterrichtskonzepts im Bereich Outdoor Education zur nderung der Einstellung zu „Chemie und Natur“*
ISBN 978-3-8325-5174-2 59.00 EUR

302 Christin Marie Sajons: Kognitive und motivationale Dynamik in Schülerlaboren. *Kontextualisierung, Problemorientierung und Autonomieunterstützung der didaktischen Struktur analysieren und weiterentwickeln*
ISBN 978-3-8325-5155-1 56.00 EUR

303 Philipp Bitzenbauer: Quantenoptik an Schulen. *Studie im Mixed-Methods Design zur Evaluation des Erlanger Unterrichtskonzepts zur Quantenoptik*
ISBN 978-3-8325-5123-0 59.00 EUR

304 Malte S. Ubben: Typisierung des Verständnisses mentaler Modelle mittels empirischer Datenerhebung am Beispiel der Quantenphysik
ISBN 978-3-8325-5181-0 43.50 EUR

305 Wiebke Kuske-Janßen: Sprachlicher Umgang mit Formeln von LehrerInnen im Physikunterricht am Beispiel des elektrischen Widerstandes in Klassenstufe 8
ISBN 978-3-8325-5183-4 47.50 EUR

306 Kai Bliesmer: Physik der Küste für außerschulische Lernorte. *Eine Didaktische Rekonstruktion*
ISBN 978-3-8325-5190-2 58.00 EUR

307 Nikola Schild: Eignung von domänenspezifischen Studieneingangsvariablen als Prädiktoren für Studienerfolg im Fach und Lehramt Physik
ISBN 978-3-8325-5226-8 42.00 EUR

308 Daniel Averbeck: Zum Studienerfolg in der Studieneingangsphase des Chemiestudiums. *Der Einfluss kognitiver und affektiv-motivationaler Variablen*
ISBN 978-3-8325-5227-5 51.00 EUR

309 Martina Strübe: Modelle und Experimente im Chemieunterricht. *Eine Videostudie zum fachspezifischen Lehrerwissen und -handeln*
ISBN 978-3-8325-5245-9 45.50 EUR

310 Wolfgang Becker: Auswirkungen unterschiedlicher experimenteller Repräsentationen auf den Kenntnisstand bei Grundschulkindern
ISBN 978-3-8325-5255-8 50.00 EUR

311 Marvin Rost: Modelle als Mittel der Erkenntnisgewinnung im Chemieunterricht der Sekundarstufe I. *Entwicklung und quantitative Dimensionalitätsanalyse eines Testinstruments aus epistemologischer Perspektive*
ISBN 978-3-8325-5256-5 44.00 EUR

312 Christina Kobl: Förderung und Erfassung der Reflexionskompetenz im Fach Chemie
ISBN 978-3-8325-5259-6 41.00 EUR

313 Ann-Kathrin Beretz: Diagnostische Prozesse von Studierenden des Lehramts – *eine Videostudie in den Fächern Physik und Mathematik*
ISBN 978-3-8325-5288-6 45.00 EUR

314 Judith Breuer: Implementierung fachdidaktischer Innovationen durch das Angebot materialgestützter Unterrichtskonzeptionen. *Fallanalysen zum Nutzungsverhalten von Lehrkräften am Beispiel des Münchener Lehrgangs zur Quantenmechanik*
ISBN 978-3-8325-5293-0 50.50 EUR

315 Michaela Oettle: Modellierung des Fachwissens von Lehrkräften in der Teilchenphysik. *Eine Delphi-Studie*
ISBN 978-3-8325-5305-0 57.50 EUR

316 Volker Brüggemann: Entwicklung und Pilotierung eines adaptiven Multistage-Tests zur Kompetenzerfassung im Bereich naturwissenschaftlichen Denkens
ISBN 978-3-8325-5331-9 40.00 EUR

317 Stefan Müller: Die Vorläufigkeit und soziokulturelle Eingebundenheit naturwissenschaftlicher Erkenntnisse. *Kritische Reflexion, empirische Befunde und fachdidaktische Konsequenzen für die Chemielehrer*innenbildung*
ISBN 978-3-8325-5343-2 63.00 EUR

318 Laurence Müller: Alltagsentscheidungen für den Chemieunterricht erkennen und Entscheidungsprozesse explorativ begleiten
ISBN 978-3-8325-5379-1 59.00 EUR

Studien zum Physik- und Chemielernen

Herausgegeben von Martin Hopf, Hans Niedderer, Mathias Ropohl und Elke Sumfleth

Die Reihe umfasst inzwischen eine große Zahl von wissenschaftlichen Arbeiten aus vielen Arbeitsgruppen der Physik- und Chemiedidaktik und zeichnet damit ein gültiges Bild der empirischen physik- und chemiedidaktischen Forschung im deutschsprachigen Raum.
Die Herausgeber laden daher Interessenten zu neuen Beiträgen ein und bitten sie, sich im Bedarfsfall an den Logos-Verlag oder an ein Mitglied des Herausgeberteams zu wenden.

Kontaktadressen:

Univ.-Prof. Dr. Martin Hopf
Universität Wien,
Österreichisches Kompetenzzentrum
für Didaktik der Physik,
Porzellangasse 4, Stiege 2,
1090 Wien, Österreich,
Tel. +43-1-4277-60330,
e-mail: martin.hopf@univie.ac.at

Prof. Dr. Hans Niedderer
Inst. f. Didaktik der Naturwissenschaften,
Abt. Physikdidaktik,
FB Physik/ Elektrotechnik,
Universität Bremen,
Postfach 33 04 40, 28334 Bremen
Tel. 0421-218 4695 (Sekretariat),
e-mail: niedderer@physik.uni-bremen.de

Prof. Dr. Mathias Ropohl
Didaktik der Chemie,
Fakultät für Chemie,
Universität Duisburg-Essen,
Schützenbahn 70, 45127 Essen,
Tel. 0201-183 2704,
e-mail: mathias.ropohl@uni-due.de

Prof. Dr. Elke Sumfleth
Didaktik der Chemie,
Fakultät für Chemie,
Universität Duisburg-Essen,
Schützenbahn 70, 45127 Essen
Tel. 0201-183 3757/3761,
e-mail: elke.sumfleth@uni-due.de

In der aktuellen Kompetenzforschung werden diverse komplexe Konstrukte erfasst. Die verwendeten Leistungstests resultieren oft in langwierigen Befragungen von Studierenden, welche so einer hohen Testbelastung ausgesetzt werden. Damit gehen Nachteile, wie beispielsweise eine sinkende Teilnahmebereitschaft, einher. Auffallend ist daher der noch seltene Einsatz adaptiver Testformate. Diese weisen eine höhere Effizienz auf als lineare Formate, womit sie die Belastung senken und die Teilnahmebereitschaft erhöhen könnten.

In dieser Arbeit wird die Anpassung eines bereits bestehenden Leistungstests von einem linearen in ein adaptives Format beschrieben. Grundlage des Vorhabens ist der Ko-WADiS-Test zur Erfassung der Kompetenz naturwissenschaftlichen Denkens bei Studierenden der Fächer Biologie, Chemie und Physik.

Den Kern der Arbeit bilden die messtheoretische Rahmung adaptiver Testformate und die Konzeption des neuen Testinstruments. Aus den Items des linearen Tests wurden mehrere adaptive Teststrukturen entwickelt und in simulierten Befragungen verglichen. Ein Multistage-Test mit drei Stufen und jeweils zwei Schwierigkeitsniveaus wies die höchste Messeffizienz auf und wurde praktisch implementiert. Zum Abschluss der Testentwicklung wurden mit dem neuen Instrument Lehramtsstudierende befragt und Messgenauigkeit sowie Messdauer in der Stichprobe mit denen des linearen Ko-WADiS-Tests verglichen. Durch das neue Testformat konnte eine signifikante Steigerung der Effizienz um 53% (Messdauer -30%, Messgenauigkeit +8%) erreicht werden.

Logos Verlag Berlin

ISBN 978-3-8325-5331-9