Entwicklung und Validierung eines Multiple-Choice-Tests zur Erfassung prozessbezogener naturwissenschaftlicher Grundbildung

Inga Glug

Kiel, März 2009

Entwicklung und Validierung eines Multiple-Choice-Tests zur Erfassung prozessbezogener naturwissenschaftlicher Grundbildung

Dissertation
zur Erlangung des Doktorgrades
der Philosophischen Fakultät der
Christian-Albrechts-Universität zu Kiel

vorgelegt von
Inga Glug

Kiel
2009

Erstgutachter: Prof. Dr. M. Prenzel
Zweitgutachter: Prof. Dr. Claus H. Carstensen

Tag der mündlichen Prüfung: 22.06.2009

Durch den zweiten Prodekan, Prof. Dr. Rainer Zaiser,
zum Druck genehmigt am: 03.07.2009

Danksagung

Die Liste der Personen, denen ich im Rückblick auf die Arbeit an meiner Dissertation danken möchte, ist lang und es können an dieser Stelle leider nur einige namentlich erwähnt werden. Ich hoffe dennoch, dass sich alle, die mich auf meinem Weg zur Fertigstellung dieser Arbeit unterstützt haben, in diesen Zeilen wiederfinden.

Zunächst möchte ich Professor Dr. Manfred Prenzel und Professor Dr. Claus H. Carstensen für ihre konstruktive Unterstützung und ihren fachlichen Rat danken. Ich habe Ihre Betreuung zu jeder Zeit als freundlich, fördernd, unkompliziert und lehrreich empfunden.

Auch von meinen Kolleginnen und Kollegen am IPN habe ich viel Unterstützung erfahren. Mein besonderer Dank gilt Dr. Dirk Hillebrandt, Dr. Katrin Schöps und meinen unmittelbaren Mit-Doktorandinnen und -doktoranden Friederike Gienke, Christoph Pawek, Silke Vorst und Susanne Weßnigk, die mich mit fachlichem Rat und allerlei Aufmunterungen (auch zuckerhaltiger Art) begleitet haben. Die Gespräche mit euch und eure Sicht der Dinge haben mir sehr geholfen und ich hoffe, dass ich von dieser Unterstützung in Zukunft etwas zurückgeben kann. Für die Beratung und Unterstützung in methodischen Fragen möchte ich Dr. Oliver Walter danken.

Danken möchte ich auch allen Expertinnen und Experten, Lehrerinnen und Lehrern sowie Schülerinnen und Schülern, die meine Studie unterstützt haben, indem sie entweder als Testpersonen fungierten oder aber als Expertinnen und Experten wichtige Anregungen für den Reviewprozess der Testitems lieferten. Ebenso danken möchte ich unseren wissenschaftlichen Hilfskräften, deren unermüdlicher Einsatz mir eine schnelle Datenerhebung und -eingabe ermöglichte.

Speziellen Dank schulde ich meinen Freundinnen und Freunden, die es in letzter Zeit sicher nicht leicht mit mir hatten und dennoch immer für mich da waren und aufmunternde Worte fanden. Ihr habt nicht zugelassen, dass ich den Kontakt zur Außenwelt verliere und mir dadurch die nötige Balance gegeben. Ich bin froh und glücklich, euch als Freunde zu haben.

Zuletzt, aber eigentlich zuallererst möchte ich mich bei meiner Familie bedanken. Ohne eure bedingungslose Unterstützung und euren Glauben an mich wäre ich heute nicht da, wo ich bin. Mein besonderer Dank gilt dabei Thorben Hahn. Du hast die Höhen und Tiefen meiner Promotionszeit verständnisvoll und geduldig mit mir durchlebt und bist schon so lange Zeit mein ruhender Pol. Du gibst mir Kraft und ich freue mich auf unser weiteres gemeinsames Leben.

ZUSAMMENFASSUNG

Naturwissenschaftliche Grundbildung stellt eine notwendige Voraussetzung für die kritische Beurteilung von Entwicklungen in den Naturwissenschaften und die Auseinandersetzung mit einer durch Naturwissenschaften und Technik geprägten Kultur dar. Insbesondere der Prozess naturwissenschaftlicher Erkenntnisgewinnung wird dabei als bedeutsam für ein grundlegendes Verständnis von Wissenschaft angesehen. Nicht erst seit den Ergebnissen aus TIMSS und PISA gibt es daher Bemühungen, die naturwissenschaftliche Grundbildung von Schülerinnen und Schülern zu verbessern. Leider fehlte es bisher abseits der veröffentlichten PISA-Items an Instrumenten, die eine Überprüfung dieser Bemühungen in ökonomischer, objektiver und valider Weise erlauben und insbesondere prozessbezogene Anteile naturwissenschaftlicher Grundbildung erfassen.

Das Ziel dieser Arbeit bestand demnach in der Entwicklung eines Tests zur Erfassung der hier als *prozessbezogen* bezeichneten naturwissenschaftlichen Grundbildung, die anhand der Fertigkeiten *Identifizieren wissenschaftlicher Hypothesen*, *Planen einer wissenschaftlichen Untersuchung* und *Nutzen wissenschaftlicher Ergebnisse* operationalisiert wurde. Das als Gruppen-Screening angelegte Verfahren wurde nach der probabilistischen Testtheorie für Schülerinnen und Schüler der neunten Klasse aus Haupt-, Realschule und Gymnasium entwickelt. Auf Basis der in unterschiedlichen Testphasen erhobenen Daten wurde es überarbeitet und in eine abschließende Testform gebracht. Im Anschluss folgte die Validierung.

Die Reliabilität des Testverfahrens nach Cronbachs Alpha beträgt $0,81$, die probabilistische WLE-Reliabilität $0,77$. Diese Werte sind für einen Gruppen-Leistungstest als gut zu bezeichnen. Die durchschnittliche Trennschärfe der Items liegt bei $0,45$ und die Itemschwierigkeiten variieren zwischen $-1,66$ und $0,88$. Die mittlere Itemschwierigkeit beträgt $-0,14$. Der Test kann damit für die vorliegende Stichprobe als angemessen bezeichnet werden, auch wenn er ein wenig zu leicht erscheint. Die interne Validierung ergab, dass die entwickelten Testitems als raschhomogen bezeichnet werden können. Die als eindimensional postulierte Kompetenz konnte nach der Durchführung von Modellgeltungstests und nach Überprüfung der Korrelationen der drei Fertigkeiten bestätigt werden.

Die Bemühungen um die Entwicklung eines *Tests zu Erfassung prozessbezogener naturwissenschaftlicher Grundbildung* können als erfolgreich bezeichnet werden. Das entstandene Verfahren ist ökonomisch, valide und reliabel und zeichnet sich durch gute statistische Kennwerte aus.

Inhaltsverzeichnis

1 Einleitung

> „Naturwissenschaftliche Kompetenz ist eine Voraussetzung für die Teilhabe an der Wissensgesellschaft und für eine lebenslange Auseinandersetzung mit einer sich verändernden Welt."
> *Deutsches PISA-Konsortium, 2001*

Dieses Zitat beschreibt eine Einsicht, zu der es in Deutschland vergleichsweise spät kam. Erst in den 1970er Jahren begann ein Umdenken bezüglich des Stellenwertes naturwissenschaftlicher Kompetenz. Es wurde erkannt, dass sie eine Ressource darstellt, die nicht nur für das Berufsleben der Menschen bedeutsam ist, sondern die darüber hinaus für den Umgang mit Technik im Alltag, für verantwortliches Handeln und für das effektive und begründete Treffen von Entscheidungen unabdingbar ist.

Interessant ist in diesem Zusammenhang, dass Diskussionen über den Stellenwert naturwissenschaftlicher Bildung in Deutschland in ähnlicher Form bereits zu Beginn des 19. Jahrhunderts stattfanden, als die Naturwissenschaften gegen das vorherrschende humanistische Bildungsideal um ihren Einzug in den Fächerkanon an deutschen Schulen kämpfen mussten. Mehr als 180 Jahre liegen zwischen der immer wieder umstrittenen und zunächst zaghaften Einführung der Naturwissenschaften in den Lehrplan der preußischen Gymnasien im Jahre 1812 und dem ungenügenden Abschneiden deutscher Schülerinnen und Schüler in den TIMSS- (Martin, Mullis, Gonzales & Chrostowski, 2004; Mullis et al., 2005) und PISA-Studien (Baumert et al., 2001; Prenzel et al., 2004; Prenzel, Artelt et al., 2007b). Die Argumentation und die Gründe für eine an die täglichen Herausforderungen der Menschen angepasste naturwissenschaftliche Bildung ähneln sich damals wie heute. Die Zeit der Industrialisierung führte mit ihren Veränderungen des Berufs- und Alltagslebens zu dem Bestreben, dass alle Menschen, egal, ob zukünftige Wissenschaftlerinnen und Wissenschaftler, Ingenieurinnen und Ingenieure oder Arbeiterinnen und Arbeiter, in der Schule naturwissenschaftlich gebildet werden sollten. Dies geschah mit dem Ziel, den Anforderungen, die neue Berufe und Maschinen im Berufs- und Alltagsleben an die Menschen stellten, gerecht werden zu können.

Heute gilt dieses Bestreben in noch höherem Maße. Die Geschwindigkeit, mit der

sich komplexe technische Entwicklungen und Neuerungen vollziehen, hat sich seit damals potenziert. Die Anforderungen an die Menschen, sich in ihrer Alltags- und Berufswelt zurechtzufinden, Neuerungen kritisch zu beurteilen und begründete Entscheidungen zu treffen, sind deutlich gestiegen. Anders als damals wird in Deutschland nicht mehr in Frage gestellt, *dass* naturwissenschaftliche Bildung notwendig ist. Dafür wird heute darüber gestritten, *welche Inhalte* eine angemessene naturwissenschaftliche Bildung enthalten sollte und *auf welche Weise* diese vermittelt werden sollten.

Die Diskussionen um Inhalte, Arten und Orte der Vermittlung naturwissenschaftlicher Bildung führten dabei in Deutschland, zusätzlich verstärkt durch die geringe Zahl und mangelnde Ausbildung des wissenschaftlichen und Ingenieurnachwuchses, in unterschiedliche Richtungen. Auf politischer Ebene mündeten sie, die Schulen betreffend, in der Festlegung naturwissenschaftlicher Bildungsstandards (KMK, 2005a, 2005c, 2005b). Außerschulisch führten sie ab Mitte der 90er Jahre an Universitäten, Forschungseinrichtungen und Unternehmen zur Einrichtung erster Science Center und Schülerlabore, von denen es heute in Deutschland über 200 gibt. Sie verschrieben sich dem Ziel, Schülerinnen und Schülern Einblicke in aktuelle naturwissenschaftliche Forschung zu gewähren und ihnen das Experimentieren mit authentischen Forschungsgegenständen zu ermöglichen. Auf diese Weise wollte man ihr Interesse an Naturwissenschaften wecken sowie naturwissenschaftliche Inhalte und Arbeitsweisen wissenschaftlicher Erkenntnisgewinnung vermitteln (Engeln & Rost, 2006).

Je nach inhaltlicher Ausrichtung bieten gerade Schülerlabore aufgrund ihrer authentischen Lernumgebung und der Zeit, die ihnen während der Schülerbesuche zur Verfügung steht, hervorragende Möglichkeiten, neben dem einfachen Wecken von Interesse unter anderem Fähigkeiten und Fertigkeiten zu vermitteln, die Merkmale und Ablauf experimenteller Erkenntnisgewinnung betreffen (Prenzel & Ringelband, 2001). Diese Fähigkeiten und Fertigkeiten sind gerade deshalb von besonderem Interesse, da sie zu einem umfangreicheren Verständnis wissenschaftlicher Sachverhalte (Songer & Linn, 1991) sowie zu einer adäquaten Einordnung und Bewertung wissenschaftlicher Erkenntnisse und zu fundierter wissenschaftlicher Argumentation führen können (Carey & Smith, 1993; Kuhn, Amsel & O´Loughlin, 1988). Schülerlabore, die sich durch authentisches, offenes und durch die Schülerinnen und Schüler gestaltbares experimentelles Arbeiten auszeichnen, besitzen also die Möglichkeit, auf diesen speziellen naturwissenschaftlichen Kompetenzbereich Einfluss zu nehmen.

Befunde darüber, ob sich dieses Potential der Einflussnahme außerschulischer Lernorte in einer Verbesserung der angesprochenen naturwissenschaftlichen Kompetenzen niederschlägt, stehen noch aus. Dieses Manko bildete einen ersten Ausgangspunkt dieser Arbeit. Sie entstand im Rahmen des Projektes *Lernort Labor* am Institut für die Pädagogik der Naturwissenschaften (IPN), das sich der Betreuung und Erforschung außerschulischer Lernorte verschrieben hat. Da dieser Forschungsbereich noch sehr jung ist, fehlen Instrumente zur Untersuchung der Einflüsse dieser Lernorte auf die Entwicklung von Kompetenzen und Interessen. So bestand das Ziel dieser Arbeit zunächst darin, ein Evaluationsinstrument zu entwickeln, das im Rahmen außerschulischer Lernorte auf ökonomische Weise Kompetenzen erfasst, die zur Planung und Durchführung der Prozesse experimenteller Erkenntnisgewinnung benötigt werden.

Bei den auf die Zielsetzung folgenden Recherchen wurde deutlich, dass das Fehlen von Instrumenten zur Feststellung naturwissenschaftlicher Kompetenz und die Bedeutsamkeit des Kompetenzbereichs eine weitreichendere und über das Feld außerschulischer Lernorte hinausgehende Betrachtung erforderlich macht. Auch den verstärkten Bestrebungen der Schulen, die naturwissenschaftliche Kompetenz ihrer Schülerinnen und Schüler zu verbessern, fehlen trotz der auch in Deutschland wachsenden Testindustrie die Instrumente, um die Ergebnisse dieser Bemühungen oder auch nur den aktuellen Stand bestimmter naturwissenschaftlicher Kompetenzen in objektiver und valider Form feststellen zu können. Abgesehen von den wenigen veröffentlichten Items der PISA-Studien liegen bisher nur wenige, frei zugängliche Verfahren zur Erfassung naturwissenschaftlicher Kompetenz vor. Ein Rückblick auf bestehende Verfahren erweist sich unter besonderer Betrachtung der Validität und Objektivität als ernüchternd. Es gibt zwar eine Vielzahl an validen Fragebögen und Interviews, die Wissen, Ansichten und Interessen der Testpersonen hinsichtlich der Naturwissenschaften erfassen. Diese erweisen sich jedoch meist als subjektiv und in Durchführung und Auswertung sehr aufwendig. Offene Einschätzungen und Aussagen von Testpersonen erfordern zur Auswertung eine Klassifizierung und lassen den Anwendern dieser Verfahren viel Auswertungs- und Interpretationsspielraum. Die wenigen Testverfahren, die sich durch hohe Objektivität und hohe Ökonomie auszeichnen, sind dagegen oft durch geringe Validität oder eine unzureichende Validitätsprüfung gekennzeichnet.

ZIELSTELLUNG DER ARBEIT

Vor diesem Hintergrund änderte sich die anfängliche Zielstellung der vorliegenden Arbeit in Richtung einer bereichsübergreifenden Entwicklung eines Tests, der valide Aussagen über die naturwissenschaftliche Kompetenz von Schülerinnen und Schülern erlaubt und sich durch eine objektive Durchführung, Auswertung und Interpretation auszeichnet. Eine weitere Anforderung ist die ökonomische, also zeitlich wenig aufwendige Anwendung und Auswertung. Aus diesem Grund wurde das Multiple-Choice-Format für die Testentwicklung gewählt. Die Testleistung einer jeden Person ist so aufgrund genau festgelegter Testwerte problemlos zu bestimmen.

Im Rahmen der Testentwicklung wurde der Fokus auf die Erfassung von Fähigkeiten und Fertigkeiten gelegt, die im Rahmen des Prozesses experimenteller Erkenntnisgewinnung von Bedeutung sind. Aufgrund dieses Rahmens werden die entsprechenden Fähigkeiten und Fertigkeiten in dieser Arbeit unter dem Begriff der *prozessbezogenen naturwissenschaftlichen Grundbildung* zusammengefasst. Es handelt sich dabei um einen speziellen Kompetenzbereich, der schulisch wie auch außerschulisch bedeutsam ist und der sich darüber hinaus auf das Lernen, die Einordnung und die Bewertung wissenschaftlicher Inhalte auswirkt. Kenntnisse darüber, wie wissenschaftliches Wissen entsteht und sich verändert, sind von großer Bedeutung, wenn es darum geht, mit einer sich ständig verändernden Umwelt umgehen und diese Veränderungen bewältigen zu können.

Drei zentrale Elemente wurden als Repräsentanten des Prozesses experimenteller Erkenntnisgewinnung definiert. Diese sind das *Identifizieren wissenschaftlicher Hypothesen*, das *Planen einer wissenschaftlichen Untersuchung* und das *Nutzen wissenschaftlicher Ergebnisse*. Sie stellen Testinhalte dar, die im Hinblick auf die Bildungsstandards nicht nur schulisch und im Hinblick auf die Möglichkeiten ihrer Vermittlung nicht nur außerschulisch eine Rolle spielen. Sie sind darüber hinaus auch übergreifend für die Naturwissenschaften Physik, Chemie und Biologie von Bedeutung.

Die drei genannten Elemente bilden die Grundlage der *Entwicklung und Validierung eines Multiple-Choice-Tests zur Erfassung prozessbezogener naturwissenschaftlicher Grundbildung* anhand einer Stichprobe, die sich aus Schülerinnen und Schülern der neunten Klasse (Haupt-, Realschule und Gymnasium) zusammensetzt. Da naturwissenschaftliche Grundbildung für alle Schulniveaus wichtig ist und gleichermaßen messbar sein soll, muss die zur Testentwicklung in diesem Bereich herangezogene Stichprobe dementsprechend umfassend gestaltet sein.

Gliederung der Arbeit

Um das Ziel der Entwicklung eines *Tests zur Erfassung prozessbezogener naturwissenschaftlicher Grundbildung* zu erreichen, muss zunächst der eigentliche Testgegenstand definiert werden. Dies geschieht in Kapitel 2 ausgehend von einer allgemeinen geschichtlichen Betrachtung naturwissenschaftlicher Bildung und Grundbildung, welche die Entwicklung und die Gründe ihrer zunehmenden Bedeutung verdeutlicht. Um den umfangreichen Bereich naturwissenschaftlicher Grundbildung so weit eingrenzen zu können, dass die Entwicklung von Aufgaben möglich wird, werden die Betrachtungen zunehmend spezieller. Auf der Grundlage bestehender Konzeptionen naturwissenschaftlicher Grundbildung und mit Hilfe nationaler sowie internationaler Bildungsstandards wird die sogenannte *prozessbezogene naturwissenschaftliche Grundbildung* definiert und auf drei zentrale Fertigkeiten eingegrenzt. Diese drei zu messenden Fertigkeiten werden hinsichtlich ihrer kognitiven Anforderungen betrachtet, so dass eine Abschätzung einer für die Testentwicklung günstigen Altersstufe erfolgen kann. Nach einer ersten strukturellen Darstellung der zu messenden Kompetenz wird in weiteren Unterabschnitten des Kapitels 2 ein Rückblick auf bisherige Versuche und Grenzen der Erfassung prozessbezogener naturwissenschaftlicher Grundbildung geworfen, der neben der Verdeutlichung des Mangels an adäquaten Testverfahren auch der Begründung des Multiple-Choice-Formats als Testformat dieser Arbeit dient. Zur Vorbereitung des späteren Operationalisierungskapitels folgen im Anschluss daran zum einen theoretische Betrachtungen, auf welche Weise Informationen in Testaufgaben dargestellt sein sollten, um eine gute Verständlichkeit der Aufgabeninhalte zu erreichen. Zum anderen folgen theoretische Überlegungen zu Kriterien, die der späteren Testvalidierung dienen werden. Den Abschluss des zweiten Kapitels bildet das Rahmenkonzept des Tests, das die wichtigsten Grundlagen des Theoriekapitels hinsichtlich der Testentwicklung noch einmal zusammenfasst.

Kapitel 3 gibt einen Überblick über die testtheoretischen Grundlagen, die für eine probabilistische Testentwicklung notwendig sind. Neben einigen Grundzügen der probabilistischen Testtheorie und dem zu Grunde liegenden Testmodell werden hier vor allem Informationen über Item- und Personenkennwerte vermittelt, die für die Überarbeitung der Testitems und die Zusammenstellung des Tests von großer Bedeutung sind. Im Rahmen der verfahrenstheoretischen Grundlagen wird noch einmal auf das spezielle Testformat und die hinsichtlich der Aufgabenzusammenstellung zu beachtenden Punkte sowie auf den Umgang mit fehlenden Werten eingegangen. Den Abschluss des Kapitels bildet die Darstellung eines typischen Testentwicklungspro-

zesses, die zwei der Itementwicklung und -überarbeitung sehr dienliche Methoden beinhaltet: das Expertenpanel und Cognitive Lab Interviews.

Nach diesen grundlegenden Kapiteln, die ein Verständnis der Beweggründe, der Inhalte und des methodischen Vorgehens der Testentwicklung vermitteln, folgt mit Kapitel 4 die Darstellung der Operationalisierung. Hier wird Schritt für Schritt beschrieben, wie die theoretischen und definitorischen Grundlagen der ersten Kapitel in Testaufgaben umgesetzt wurden, anhand welcher Stichproben die Itementwicklung vorgenommen wurde und auf welche Weise die Prüfung, Weiterentwicklung und Validierung der Items erfolgte.

Kapitel 5 stellt die Ergebnisse der Testentwicklung nach Entwicklungsstufen, angefangen von der Pilotierung bis hin zum Haupttest, dar. Im Rahmen der Validierung, die dieses Kapitel abschließt, zeigt sich, inwiefern der entwickelte Test das misst, was er zu messen beansprucht.

Im Rahmen von Kapitel 6 werden die Ergebnisse der Testenwicklung diskutiert und theoretisch eingeordnet. Den Abschluss der Arbeit bildet mit Kapitel 7 der Ausblick, in dessen Rahmen unter anderem der weitere Umgang und weitere Einsatzmöglichkeiten des Testverfahrens erörtert werden.

2 Inhaltliche und theoretische Grundlagen

Die inhaltlichen und theoretischen Grundlagen werden, ausgehend von einer geschichtlichen und definitorischen Betrachtung, auf die Bedeutsamkeit naturwissenschaftlicher Grundbildung eingehen und zeigen, welchen besonderen Einfluss außerschulische Lernorte auf diese Grundbildung nehmen können. Auf der Grundlage eines an Weinert (Weinert, 1999) und Klieme et al. (2001) angelehnten Kompetenzbegriffs wird anhand unterschiedlicher Quellen verdeutlicht, welche Komponenten die spezielle *prozessbezogene naturwissenschaftliche Grundbildung* ausmachen und wie diese Kompetenz strukturell aufgebaut ist. In diesem Zusammenhang wird vertiefend darauf eingegangen, ab welchem Alter Testpersonen die notwendigen kognitiven Grundlagen besitzen, um die Fertigkeiten erlernen bzw. besitzen zu können, die der *Test zur Erfassung prozessbezogener naturwissenschaftlicher Grundbildung* messen soll. Diese Überlegungen sind der Ausgangspunkt für die Auswahl der Stichprobe, anhand derer die Testentwicklung realisiert wird.

Um zeigen zu können, welche Anforderungen ein Test in diesem speziellen Bereich naturwissenschaftlicher Grundbildung zu erfüllen hat, werden Vor- und Nachteile unterschiedlicher Verfahren zur Kompetenzmessung beschrieben. Diese Betrachtungen sind verbunden mit einer kurzen Rückschau auf bisherige Versuche und Grenzen der Erfassung naturwissenschaftlicher Grundbildung und einer Abgrenzung des im Rahmen dieser Arbeit entwickelten Verfahrens von bisherigen Verfahren.

Gegen Ende des Kapitels werden praktische Aspekte der Aufgabenentwicklung betrachtet. Es wird darauf eingegangen, auf welche Weise Informationen in Tests möglichst verständlich dargestellt werden können, welche Art von Graphiken von Schülerinnen und Schülern besonders leicht verstanden werden und was bei der Darstellung von Bildmaterial zu beachten ist.

Bevor das Kapitel durch das Rahmenkonzept abgeschlossen wird, welches der Testentwicklung zu Grunde liegen soll, werden Kriterien theoretisch begründet, anhand derer eine erste Testvalidierung erfolgen kann.

2.1 NATURWISSENSCHAFTLICHE GRUNDBILDUNG

Wie insbesondere die internationalen Vergleichsstudien TIMSS[1] (Baumert et al., 1997; Martin et al., 2004; Mullis et al., 2005) und die ersten beiden PISA[2]-Studien gezeigt haben (Baumert et al., 2001; Prenzel et al., 2004), bedurften die Kompetenzen deutscher Schülerinnen und Schüler im Bereich der *naturwissenschaftlichen Grundbildung* (*Scientific Literacy*) einer Verbesserung. Zwar hat sich dieses Bild aufgrund der PISA-Ergebnisse 2006 (OECD, 2007) relativiert, doch die Notwendigkeit, die naturwissenschaftliche Grundbildung weiter zu verbessern, besteht weiterhin und umso mehr, da diese Art der Grundbildung weitreichende Folgen für das Erlernen und die Anwendung naturwissenschaftlicher Inhalte hat (Songer & Linn, 1991). Der folgende Abschnitt wird zunächst den Begriff der naturwissenschaftlichen Grundbildung geschichtlich und definitorisch betrachten und auf die Bedeutung naturwissenschaftlicher Grundbildung eingehen. Den Abschluss bildet eine Beschreibung des speziellen Nutzens außerschulischer Lernorte für diese Art der Bildung.

2.1.1 GESCHICHTE, BEGRIFF UND ALLGEMEINE BEDEUTUNG NATURWISSENSCHAFTLICHER GRUNDBILDUNG

Die aktuellen Diskussionen über die Bedeutung naturwissenschaftlicher Bildung sind nicht neu und gehen zum Teil bis in die Zeit des alten Griechenlands zurück. Für den Rahmen dieser Arbeit hat es sich als sinnvoll und zielführend erwiesen, das 19. Jahrhundert als Ausgangspunkt der geschichtlichen Betrachtungen zu wählen. Zu dieser Zeit lag die Bedeutung von Wissenschaft - begründet durch die industrielle Revolution - in der Entwicklung praktischer technischer Anwendungen, wodurch der Wissenschaft und den Wissenschaftlerinnen und Wissenschaftlern auch eine gewisse soziale Verantwortung zukam. Die raschen Entwicklungen der Wissenschaft hatten zunehmend Auswirkungen auf die Erziehung. Man befasste sich mit der Frage, ob das Studium der Naturwissenschaften in eine allgemeine, freie, kulturelle und humanistische Bildung eingeschlossen werden könne. Einer Befürwortung stand hauptsächlich das Argument entgegen, dass Wissenschaft zu sehr auf den Nutzen und auf wirtschaftliche Ziele ausgerichtet sei, als dass sie zu den Werten beitragen könne, die Menschen zu einem zufriedenstellenden Leben verhelfen. Darüber hinaus war die praktische Anwendbarkeit der wissenschaftlichen Produkte zu sehr mit der allgemeinen Öffentlichkeit und der Arbeiterschaft verbunden und stand damit den in-

1 Third International Mathematics and Science Study
2 Programme for International Student Assessment

tellektuellen Ansprüchen der elitären Oberklasse gegenüber, die gefüllt waren mit ideellen und moralisch erhabenen Zielen (Daum, 2002). Obwohl wissenschaftliche Studien und die dahinter stehenden Vorstellungen einer von gegebenen Autoritäten unabhängigen und damit demokratisch offenen Forschung im Grunde begrüßt wurden, reichte dies nicht als Rechtfertigung für einen Platz im Curriculum. Diese Betrachtungsweise wandelte sich jedoch. Charles DeGarmo, ein Befürworter der Ideen Herbarts[3] in den Vereinigten Staaten, vertrat eine Strömung, die sich gegen eine elitäre Bildung richtete. Bildung sollte einer breiteren Öffentlichkeit zugänglich sein und Inhalte einschließen, die von zeitgenössischer Relevanz sind (DeGarmo, 1895/2007). 1893 versuchte das *National Education Association´s Committee of Ten* eine Verbindung zu schaffen zwischen alten Disziplinen wie der Mathematik und den klassischen Sprachen und neuen Disziplinen wie moderner Geschichte, modernen Sprachen und den Naturwissenschaften (W. Gräber & Bolte, 1997). Dieses Bestreben sollte der intellektuellen Entwicklung aller Schülerinnen und Schülern gerecht werden und alle gleichermaßen auf das Leben vorbereiten, egal ob sie später studieren oder direkt einen Beruf ergreifen würden. Damit sollte Schülerinnen und Schülern die Möglichkeit gegeben werden, für sich selbst zu denken und nicht nur Fakten über die natürliche Welt auswendig zu lernen. Eine Veränderung der Lehrpläne in diese Richtung verfolgte insbesondere Dewey (1916), Philosoph, Pädagoge, Psychologe und Reformer des amerikanischen Schulwesens. Er vertrat die Meinung, dass man die Naturwissenschaften unbedingt in den Kanon der zu unterrichtenden Fächer aufnehmen sollte, da sie aus den Alltagserfahrungen der Schülerinnen und Schülern ohnehin nicht wegzudenken seien: *„Die Berufsarbeit ihrer Eltern ist angewandte Naturwissenschaft, und die Vorgänge im Haushalt, die Verhütung von Krankheiten, die Eindrücke der Straße enthalten naturwissenschaftliche Gedanken und erregen das Interesse an den verwandten wissenschaftlichen Prinzipien.“* (Dewey, 1916/1993, S.374). Würde man die Naturwissenschaften nicht unterrichten, so würden die einheitlichen Erfahrungen der Schülerinnen und Schüler zerrissen, die sich im Alltag *„naturwissenschaftlichen Tatsachen und Prinzipien in Verflechtungen mit den verschiedensten Formen menschlicher Betätigung“* gegenübersehen (Dewey, 1916/1993, S. 372).

In Deutschland verlief die Integration der Naturwissenschaften in ähnlich schwieriger Weise und unterlag einem langwierigen Prozess, der zwischen Anerkennung

3 Johann Friedrich Herbart: deutscher Philosoph, Psychologe und Pädagoge, der über den deutschen Sprachraum hinaus als Klassiker der Pädagogik gilt. Er begründete den Herbartianismus und gilt als Pionier in der Entwicklung einer auf der Psychologie basierenden systematischen Theorie zum Lernen und Lehren. Seine Intention war es, Schülerinnen und Schülern durch Anstoß und Unterstützung zur Selbstbildung zu verhelfen.

und Beschneidung in den Lehrplänen der Schulen schwankte. In Preußen beispielsweise führte Süvern 1812/1816 die Naturwissenschaften mit je zwei Stunden pro Klasse in den Gymnasien ein. 1891 wurde Naturkunde an den Gymnasien Bayerns zum Pflichtfach und erst 1925 beschloss das preußische Kultusministerium, dass Biologie fester Bestandteil der Lehrpläne sein sollte. Zwischen der Einführung und der vollständigen Anerkennung der Naturwissenschaften lagen demnach mehr als einhundert Jahre, die geprägt waren von Auseinandersetzungen zwischen Humanisten und Anhängern einer modernen naturwissenschaftlichen Bildung sowie dem Darwinismusstreit. Am Ende setzte sich die Entwicklung eines Lehrplanes, der die Naturwissenschaften enthielt, nicht zuletzt auch aufgrund des massiven Drucks der Öffentlichkeit, insbesondere der Techniker und Ingenieure, der wirtschaftlichen Administratoren, der Universitätsprofessoren naturwissenschaftlicher Fächer und von Teilen des Militärs, durch. (Daum, 2002).

Von Beginn an wurden sowohl Inhalt als auch Methode als zentrale Komponenten einer naturwissenschaftlichen Bildung zum Zwecke einer allgemeinen Bildung betrachtet. Diese allgemeine naturwissenschaftliche Bildung wurde einerseits als Teil menschlicher Kultur gesehen. Andererseits sollte sie Menschen in die Lage versetzen, als mündige Bürger Entwicklungen in den Naturwissenschaften beurteilen zu können. Diese Ziele änderten sich in der Zeit nach dem Zweiten Weltkrieg und im Rahmen der Geschehnisse des Kalten Krieges. Zwar spielten die ursprünglichen Ziele noch immer eine Rolle, aber es entstanden weitere, die durch nationale Sicherheitsfragen sowie durch den wissenschaftlichen und militärischen Wettstreit im Rahmen des Kalten Krieges begründet waren. Naturwissenschaftliche Bildung sollte militärische und wirtschaftliche Vorteile verschaffen (DeBoer, 1997).

Im Laufe der Veränderung bzw. Erweiterung der Ziele naturwissenschaftlicher Bildung wurde der Begriff der *Scientific Literacy* geprägt und definitorisch über die Jahrzehnte immer weiter ausgeformt. Im Folgenden werden einige der Konzeptionen dieses Begriffs vorgestellt.

Begriffsdefinition

Obwohl es in der Geschichte schon über einen langen Zeitraum hinweg Überlegungen dazu gegeben hatte, wie naturwissenschaftliche Bildung zu definieren sei, wurden die Inhalte dieser Bildung erst nach dem Zweiten Weltkrieg unter dem Begriff der *Scientific Literacy* zusammengefasst. Zum ersten Mal tauchte der Begriff bei Watson und Cohen (1952) auf. Paul deHart Hurd nahm ihn in das Lexikon der Naturwis-

senschaftsdidaktik auf, als er ihn in einem Artikel für *Educational Leadership* (1958) gebrauchte. *Scientific Literacy* wurde über mittlerweile fünf Jahrzehnte hinweg sehr umfassend definiert, wobei sich jedoch die Definitionsschwerpunkte über die Zeit verändert haben. Noch in den 60er Jahren standen die sozialhistorische Entwicklung, Ethos, soziale und kulturelle Beziehung sowie soziale Verantwortung der Naturwissenschaften im Vordergrund der Begriffsdefinition. Bis heute geht die Entwicklung weiter in Richtung einer Definition, die naturwissenschaftliches Wissen (naturwissenschaftliche Konzepte), Prozesse der Naturwissenschaften und die mit Naturwissenschaften verbundenen Fertigkeiten beinhaltet (R. Bybee, 2002). Da sich der Begriff der *Scientific Literacy* einer wörtlichen Übersetzung entzieht, wurde der Begriff in Deutschland mit dem Ausdruck *naturwissenschaftliche Grundbildung* übersetzt. *Naturwissenschaftliche Grundbildung* wird im Rahmen des PISA-Projekts definiert als

> *„... Fähigkeit, naturwissenschaftliches Wissen anzuwenden, naturwissenschaftliche Fragen zu erkennen und aus Belegen Schlussfolgerungen zu ziehen, um Entscheidungen zu verstehen und zu treffen, welche die natürliche Welt und die durch menschliches Handeln an ihr vorgenommenen Veränderungen betrifft."* (OECD, 1999, S.60).

Sie sollte dabei Wissensstrukturen, die Methoden der Wissensproduktion und die Verbindungen zwischen Entdeckung und Anwendung, die Wissenschaft, ihre Methoden und deren kritische Reflexion beinhalten (Oelkers, 1997). Weiterhin drückt sie die Maßgabe aus, dass alle Bürgerinnen und Bürger einer fortschrittlichen, technologischen Industriegesellschaft über eine fundierte naturwissenschaftliche Bildung verfügen sollten. Diese sollte Kompetenzen umfassen, die die Aneignung neuer Wissensbestände (Anschlussfähigkeit des Wissens), die Nutzung von Wissen und Fertigkeiten sowie die Auseinandersetzung mit Naturwissenschaften über die Lebensspanne hinweg gewährleisten.

Der Begriff der *Bildung* sagt aus, dass es sich hier um Erlernbares und um Lernprozesse handelt. So kann sich die naturwissenschaftliche Grundbildung bei entsprechender Anstrengung des Lerners über die Lebensspanne entwickeln und sich immer weiter ausdifferenzieren. Diese Idee der Entwicklung naturwissenschaftlicher Grundbildung beschreibt Bybee (1997) anhand seines Stufenmodells:

- *Nominale naturwissenschaftliche Grundbildung*: Ein Ausdruck, eine Frage oder ein Thema wird als wissenschaftlich erkannt, aber das Verständnis einer besonderen Situation ist im Wesentlichen auf die Ebene der naiven Theorien (alternative oder Alltagskonzeptionen) beschränkt.

- *Funktionale naturwissenschaftliche Grundbildung*: Individuen können wissenschaftliches Vokabular benutzen, aber der Gebrauch ist oft auf eine bestimmte Aktivität oder ein bestimmtes Bedürfnis beschränkt.
- *Konzeptuelle und prozedurale naturwissenschaftliche Grundbildung*: Es ist ein Verständnis darüber vorhanden, in welchem Verhältnis konzeptuelle Teile einer Disziplin mit der Disziplin als Ganzes stehen. Darin sind auch prozedurales Wissen und Fertigkeiten enthalten, die sich auf den Prozess wissenschaftlicher Erkenntnisgewinnung beziehen.
- *Multidimensionale naturwissenschaftliche Grundbildung*: Stellt die höchste Ebene dar und umfasst ein Verständnis des Wesens der Naturwissenschaften, der Geschichte ihrer Ideen und ihrer Rolle in Kultur und Gesellschaft.

Zusammenfassend zeigt sich, dass nationale wie auch internationale Konzeptionen naturwissenschaftlicher Grundbildung relativ übereinstimmend folgende Bereiche enthalten (Duit, Häußler & Prenzel, 2001):

- Naturwissenschaftliche Begriffe und Prinzipien (Konzepte),
- Naturwissenschaftliche Untersuchungsmethoden und Denkweisen (Prozesse),
- Vorstellungen zur Natur der Naturwissenschaften (Nature of Science - NOS) [4],
- Vorstellungen und Einstellungen zur Relevanz der Naturwissenschaften in Gesellschaft und Technik.

Das Ideal, dass Bürgerinnen und Bürger im Besitz dieser naturwissenschaftlicher Grundbildung sein sollten, um in mündiger Weise an einer durch Naturwissenschaften und Technik geprägten Kultur teilhaben und sich in ihr zurechtfinden zu können, fand international wie auch national Niederschlag in der Formulierung konkreter Bildungsziele und Bildungsstandards. Anhand dieser Standards werden die Vorstellungen über Kompetenzen naturwissenschaftlicher Grundbildung konkretisiert und für die Entwicklung von Curricula umsetzbar gemacht. In den USA wurden diese Anforderungen in Form der *National Science Education Standards* manifestiert (NRC, 1996). In Deutschland verabschiedete die Kultusministerkonferenz (KMK) im Dezember 2004 entsprechende Bildungsstandards für den mittleren Bildungsabschluss

4 Die Natur der Naturwissenschaften (*engl.: Nature of Science - NOS*) bezieht sich auf die Werte und zu Grunde liegenden Annahmen, die für wissenschaftliches Wissen wesentlich sind. Dies schließt alle Einflüsse und Begrenzungen ein, die aus Wissenschaft als menschlichem Bemühen resultieren. Eine Aufstellung der NOS-Aspekte findet sich bei McComas (1998, 2005).

der Fächer Biologie (KMK, 2005a), Physik (KMK, 2005c) und Chemie (KMK, 2005b). Sie geben Auskunft darüber, wie weit die naturwissenschaftliche Bildung der Schülerinnen und Schüler bis zum Ende der Pflichtschulzeit fortgeschritten sein sollte.

BEDEUTUNG NATURWISSENSCHAFTLICHER GRUNDBILDUNG

Die Bedeutung naturwissenschaftlicher Grundbildung kann grob in zwei Sichtweisen, eine Makro- und eine Mikroebene, eingeteilt werden (Laugksch, 2000).

Auf der Makroebene wird die Verbindung zwischen naturwissenschaftlicher Grundbildung und Gesellschaft, Wirtschaft und Staat beschrieben. Der Reichtum eines Landes hängt unter anderem davon ab, wie konkurrenzfähig es sich auf dem internationalen Markt zeigt, wie gut es in der Lage ist, im Bereich neuer Technologien und Entwicklungen auf dem Laufenden zu bleiben und eventuell selbst eine Vorreiterrolle zu übernehmen. Für eine solche Entwicklung sind gut ausgebildete Personen notwendig, die sie tragen und vorantreiben können. Darüber hinaus bedürfen neue Entwicklungen einer Unterstützung durch die Bürgerinnen und Bürger eines Landes, die wissenschaftliche Entwicklungen nur beurteilen können, wenn sie eine Vorstellung davon haben, was die Wissenschaftlerinnen und Wissenschaftler mit welchem Ziel tun. Dabei geht es hier gerade nicht um blinde Zustimmung, sondern um die Möglichkeit einer sachlich richtigen Beurteilung. Außerdem soll das Interesse an Naturwissenschaften und Technik aufrechterhalten oder geschaffen werden. Dies wiederum ist wichtig für die Sicherung des wissenschaftlichen Nachwuchses und damit die Zukunft eines Landes (Laugksch, 2000).

Die Mikroebene beschreibt die Bedeutung der Grundbildung für das Individuum. Naturwissenschaftliche Grundbildung gibt den Bürgerinnen und Bürgern eines Landes ein komfortables und selbstsicheres Gefühl im Umgang mit wissenschafts- und technologiebezogenen Sachverhalten. Sie können zwischen pseudowissenschaftlichen und fundierten Informationen unterscheiden und kompetente Entscheidungen treffen (Laugksch, 2000). Dieser Punkt wird auch im Rahmen der PISA-Untersuchung hervorgehoben: Naturwissenschaftliche Grundbildung wird als unbedingt notwendig erachtet, um sich in der heutigen Informationsflut immer neuer wissenschaftlicher Erkenntnisse der unterschiedlichsten Bereiche orientieren zu können. Sie ermöglicht es, sich als mündiger Bürger zu erweisen, Entwicklungen in den Naturwissenschaften kritisch zu beurteilen (Field & Powell, 2001) und sich lebenslang mit den Veränderungen der Welt auseinandersetzen zu können (Duit et al., 2001). Auf diese Weise gebildete Menschen sind darüber hinaus in einer komfortableren Situation in

Bezug auf ihre Flexibilität in der Berufswahl. Sie können das Angebot wissenschafts- und technologiebezogener Berufe besser nutzen und sind auch auf neue Entwicklungen in ihrem Beruf gut vorbereitet (Thomas & Durant, 1987). Dabei sind hier keineswegs nur Berufe auf der Führungsebene von Wissenschaft, Wirtschaft und Politik angesprochen. Der Anteil an Arbeitsplätzen, die naturwissenschaftliches Grundverständnis und den Umgang mit immer hochtechnisierteren Geräten verlangen, steigt immer weiter und nicht zuletzt stellt auch das normale Alltagsleben hohe technische Anforderungen an jeden einzelnen (Baumert et al., 2001).

2.1.2 BEDEUTUNG AUSSERSCHULISCHER LERNORTE FÜR DEN ERWERB NATURWISSENSCHAFTLICHER GRUNDBILDUNG

Schulische Lernumgebungen müssen sich der Herausforderung stellen, die naturwissenschaftliche Grundbildung von Schülerinnen und Schüler effektiver als in der Vergangenheit zu fördern. Die Umgebung einer Schule kann den Schülerinnen und Schülern jedoch im Rahmen des naturwissenschaftlichen Unterrichts meist nicht die wichtigen Charakteristika authentischer naturwissenschaftlicher Arbeit[5] bieten. Unzureichende schulische Rahmenbedingungen - Zeit, Geld, Expertise, Räumlichkeiten und Ausrüstung betreffend - sowie ein aufgrund enger Lehrpläne oft auf Inhalte konzentrierter Unterricht lassen eine Vermittlung authentischer Forschung in Schulen häufig nicht zu. Es bleibt damit oft bei Vorführexperimenten oder beim einfachen Experimentieren, das sich bezüglich der ablaufenden kognitiven Prozesse gemäß Chinn und Malhotra (2002b) deutlich von authentischer naturwissenschaftlicher Forschung unterscheidet (s. Abbildung 2.1).

Diese Diskrepanz können außerschulische Lernorte überbrücken, deren Authentizitätsgrad je nach Konzeption der Lernorte zwischen den beiden in der Tabelle dargestellten Extremen anzusiedeln ist. Außerschulische Lernorte stellen durch ihre Verortung an Universitäten, Forschungszentren und Wirtschaftsunternehmen zum einen eine hervorragende Möglichkeit dar, Schülerinnen und Schülern wissenschaftliche Inhalte und aktuelle Forschung nahe zu bringen und dadurch ihr Interesse für die Naturwissenschaften zu wecken. Zum anderen sind sie darüber hinaus aufgrund der zeitlich intensiven Betreuung durch Wissenschaftlerinnen und Wissenschaftler und aufgrund ihrer Ausstattung in der Lage, den Schülerinnen und Schülern auch Merk-

5 Unter authentischer naturwissenschaftlicher Arbeit werden hier alle Aktivitäten verstanden, die Wissenschaftlerinnen und Wissenschaftler innerhalb ihrer Forschung ausführen. Diese Arbeit ist gekennzeichnet durch komplexe Aktivitäten, bei der oft teure Gerätschaften, elaborierte Prozeduren und Theorien, hoch spezialisierte Expertise und fortgeschrittene Techniken der Datenanalyse und Modellierung zum Einsatz kommen (Chinn & Malhotra, 2002b)

Tabelle 2.1: Vergleich der kognitiven Anforderungen authentischer Forschung und einfacher Experimente

Kognitive Prozesse	Authentische Forschung	Einfache Experimente
Aufstellen von Forschungsfragen	Wissenschaftler generieren eigene Forschungsfragen.	Forschungsfragen werden den Schülern vorgegeben.
Studien planen Variablen aussuchen	Wissenschaftler suchen Variablen aus oder erfinden Variablen zur Untersuchung. Es gibt viele mögliche Variablen.	Schüler untersuchen eine oder zwei vorgegebene Variablen.
Variablen kontrollieren	Wissenschaftler wenden unterschiedliche Kontrollen an. Es kann schwierig sein zu entscheiden, welche Kontrollen auf welche Weise greifen sollen.	Es gibt lediglich eine Kontrollgruppe. Schülern wird vorgegeben, welche Variablen zu kontrollieren sind und wie ein kontrolliertes Experiment auszusehen hat.
Beobachtungen anstellen	Wissenschaftler wenden elaborierte Techniken an, um Beobachterfehler zu vermeiden.	Beobachterfehler werden nicht thematisiert, obwohl Messinstrumente, wie z.B. Lineale benutzt werden.
Indirektes Schlussfolgern	Beobachtungen werden durch komplexe Schlussfolgerungen mit der Forschungsfrage verbunden. Beobachtete Variablen und die interessierenden theoretischen Variablen sind nicht identisch.	Beobachtungen werden direkt mit den Forschungsfragen verbunden. Die beobachteten Variablen sind die interessierenden Variablen.

male des Prozesses und den Ablauf wissenschaftlicher Erkenntnisgewinnung zu vermitteln. Dadurch bieten außerschulische Lernorte die Möglichkeit, die naturwissenschaftliche Grundbildung von Schülerinnen und Schülern zu verbessern und positiv Einfluss auf ihre erkenntnistheoretischen Überzeugungen zu nehmen. Die Schülerinnen und Schüler erhalten Einblick in die Entstehung, in die Veränderung und die Bedingungen der Änderung von Wissen aufgrund wissenschaftlicher Erkenntnisgewinnung.

Außerschulische Lernorte stellen informelle Lernumgebungen dar. Gerber (2001) konnte zeigen, dass sinnvoll eingebettete informelle Lernaktivitäten außerhalb des

Klassenraums mit besseren Fähigkeiten, wissenschaftlich zu schlussfolgern, verbunden sind. Gleiches zeigte Gerber bei einem Vergleich zwischen einer experimentellen und einer nicht-experimentellen Klassenumgebung: Schülerinnen und Schüler, die sich experimentell wissenschaftliches Wissen aneignen konnten, besaßen bessere schlussfolgernde Fähigkeiten als die Schülerinnen und Schüler in einer nicht-experimentellen Klassenumgebung. Einen ähnlichen Einfluss sollte eine offene experimentelle Lernumgebung eines außerschulischen Lernortes haben.

Außerschulische Lernorte nehmen dabei nicht per se positiven Einfluss auf die naturwissenschaftliche Grundbildung der Schülerinnen und Schüler. Beispielsweise führt reines Bestreben, ihnen ein positives Bild der Naturwissenschaften zu vermitteln, dazu, dass Personen nach dem Besuch eines außerschulischen Lernortes zwar eine positivere Einstellung gegenüber Naturwissenschaften haben, ihre Vorstellungen aber weniger wissenschaftlich als vor dem Besuch ausfallen (Rennie & Williams, 2002). Um eine Verbesserung der naturwissenschaftlichen Grundbildung zu erreichen, müssen den Schülerinnen und Schülern Forschungsaufgaben geboten werden, die einen möglichst offenen und freien *hands-on*-Umgang[6] mit realen und authentischen Materialien ermöglichen. Schülerinnen und Schüler sollten in die Lage versetzt werden, eine Untersuchung von Beginn an, ausgehend von der Formulierung einer eigenen Forschungsfrage und unter dosierter Hilfestellung anwesender Wissenschaftlerinnen und Wissenschaftler, zu planen, durchzuführen und auszuwerten. Authentizität und Offenheit der Lernumgebung lassen ein aktives Engagement der Schülerinnen und Schüler zu, können auf diese Weise motivierend wirken und bedeutsames Lernen fördern (Roth & Roychoudhury, 1993). Befolgen Schülerinnen und Schüler dagegen bei der Durchführung von Experimenten lediglich rezeptartige Anweisungen, so wird der Prozess wissenschaftlicher Erkenntnisgewinnung reduziert auf ein einfaches Richtig-Oder-Falsch-Spiel. Eine von Kontexten und Zielen losgelöste, einfache praktische Arbeit reicht nicht aus, um eine Verbesserung naturwissenschaftlicher Grundbildung in den Schülerinnen und Schülern zu erreichen (Hart, Mulhall, Berry, Loughran & Gunstone, 2000).

Interviewstudien mit Schülerinnen und Schülern, die in einer offenen Laborumgebung experimentierten, zeigen, dass diese glaubten, durch diese Erfahrungen ein tieferes und konkreteres Verständnis naturwissenschaftlicher Konzepte zu erhalten.

6 Der Begriff *hands-on* wurde in den späten 1960ern geprägt und hatte zunächst die Bedeutung, den Umgang mit Computern durch das *Handanlegen* an die Tastatur, also durch einfaches Benutzen des Computers zu erlernen (Rutherford, 1993). Diese Form des *learning by doing* geht auf Dewey und letztlich auf Pestalozzi und Fröbel zurück (S. Smith, 1979). Heutzutage sind *hands-on*-Aktivitäten Gegenstand von Naturwissenschafts-Curricula und werden für die Vermittlung von Wissenschaft als sehr bedeutsam erachtet (Flick, 1993).

Außerdem gaben sie an, dass es ihr Verständnis naturwissenschaftlicher Prozesse, das Verständnis der Herkunft wissenschaftlichen Wissens und das Verständnis darüber, wie Wissenschaftlerinnen und Wissenschaftler arbeiten, erleichtert habe (Tsai, 1999). Offene, authentische Erfahrungen mit dem Planen, Durchführen und Auswerten eigener Experimente können durchaus zu einer Verbesserung und einem tieferen Verständnis der Prozesse naturwissenschaftlicher Erkenntnisgewinnung führen. Roth und Roychoudhury (1993) ließen in ihrer Studie Schülerinnen und Schüler über einen Zeitraum von sieben Wochen bzw. zwei Monaten (je nach Klassenstufe) in authentischer Umgebung für sie bedeutsame Fragen wissenschaftlich untersuchen. Es zeigte sich, dass sich beispielsweise die Fähigkeit, zu prüfende Variablen zu identifizieren, über die Zeit der Untersuchung deutlich verbesserte. Die Forschungsfragen der Schülerinnen und Schüler wandelten sich von qualitativen Fragen zu spezifischen und wissenschaftlich untersuchbaren Fragen. Dies ging einher mit ihrer Einarbeitung in ein bestimmtes Themengebiet. Es wird deutlich, dass die Verbesserung prozessbezogener naturwissenschaftlicher Fertigkeiten in Abhängigkeit von der Vertrautheit der Schülerinnen und Schüler mit dem jeweiligen Kontext erfolgt. Weiterhin zeigt sich anhand qualitativer Studien, dass das kooperative Arbeiten der Schülerinnen und Schüler sowie der Austausch mit Wissenschaftlerinnen und Wissenschaftlern die genaue Definition von Konzepten und eine genaue Planung von Experimenten fördert sowie in spezieller Weise motivierend wirkt (Roth & Roychoudhury, 1993; Hart et al., 2000). Von ähnlichen Effekten berichten auch Hofstein und Lunetta (1982, 2004): Eine angemessen gestaltete Laborumgebung könne zur Verbesserung von Forschungs- und Problemlösefertigkeiten führen. Dabei wirke diese Umgebung insbesondere unterstützend auf die Planung von Untersuchungen, auf Beobachtungsfertigkeiten und das Verständnis wissenschaftlicher Konzepte. Außerdem würden Fertigkeiten der Kooperation und Kommunikation gefördert.

Wichtig für die Sicherstellung eines Lerneffektes, sei es im Hinblick auf den Prozess naturwissenschaftlicher Erkenntnisgewinnung oder aber im Hinblick auf die Natur der Naturwissenschaften (Nature of Science - NOS), ist ein explizites Ansprechen der Prozesse und Merkmale naturwissenschaftlicher Erkenntnisgewinnung. Die Schülerinnen und Schüler bedürfen einer gewissen Anleitung, um über Sinn und Zweck ihrer wissenschaftlichen Handlungen nachzudenken (Schwartz, Lederman & Crawford, 2004). Ein einfaches *hands-on* ohne ein gleichzeitiges Reflektieren und ein Bewusstsein der Vorgänge reicht nicht aus, um die naturwissenschaftliche Grundbildung der Schülerinnen und Schüler zu verbessern.

Zusammenfassend seien zum Abschluss noch einmal die Voraussetzungen genannt,

die gemäß der genannten Ergebnisse aus Studien erfüllt sein müssen, damit außerschulische Lernorte die naturwissenschaftliche Grundbildung positiv beeinflussen können (vgl.Engeln & Rost, 2006).

- Schülerinnen und Schüler müssen Sinn und Zweck der Aufgaben verstehen, die im Rahmen dieser Lernumgebung auf sie zukommen.
- Die experimentelle Arbeit sollte die Schülerinnen und Schüler herausfordern, aber nicht überfordern.
- Die experimentelle Arbeit sollte nicht auf das Befolgen rezeptartiger Anweisungen reduziert sein.
- Die Arbeit sollte theoretisch fundiert in praktischer Weise, also *hands-on*, erfolgen.
- Der inhaltliche Kontext sollte bekannt sein oder die Schülerinnen und Schüler sollten sich in einen bestimmten Kontext einarbeiten können.
- Die zu bearbeitenden Themen sollten für die Schülerinnen und Schüler von persönlicher Relevanz sein.
- Die Lernumgebung sollte möglichst offen und durch die Schülerinnen und Schüler gestaltbar sein.
- Die experimentelle Arbeit sollte nicht einmalig, sondern über einen längeren Zeitraum stattfinden.
- Die Arbeit sollte kooperativ in Teams erfolgen.
- Die Lernumgebung sollte als authentisch wahrgenommen werden.

2.2 PROZESSBEZOGENE NATURWISSENSCHAFTLICHE GRUNDBILDUNG

Nachdem nun anhand des ersten Abschnitt dieses Kapitels eine allgemeine Betrachtung naturwissenschaftlicher Grundbildung erfolgt ist, werden im Folgenden die Inhalte und das Konzept der speziellen *prozessbezogenen naturwissenschaftlichen Grundbildung* dargestellt. Dabei werden die Grundlagen zur Definition dieses Bereiches auf verschiedenen Ebenen gelegt. Ausgehend von dem Kompetenzbegriff, welcher dem

Konstrukt zu Grunde liegt, wird zunächst darauf eingegangen, welche Prozesskomponenten von besonderer Bedeutung sind, um den Bereich experimenteller Erkenntnisgewinnung zu repräsentieren. Um die Auswahl derjenigen Fertigkeitsbereiche zu begründen, die anhand des in dieser Arbeit entwickelten Tests erfasst werden sollen, werden nationale (KMK, 2005a, 2005c, 2005b) wie auch internationale Bildungsstandards (NRC, 1996), das SDDS-Modell (Klahr, 2000; Klahr & Dunbar, 1988), das PISA-Rahmenkonzept 2006 der naturwissenschaftlichen Grundbildung (OECD, 2006) sowie weitere aktuelle Scientific-Literacy-Konzepte (American Association of the Advancement of Science, 2001; University of York Science Education Group, 2006) herangezogen. Den Abschluss dieses Abschnitts bildet eine erste schematische Darstellung der *prozessbezogenen naturwissenschaftlichen Grundbildung*.

2.2.1 ZU GRUNDE LIEGENDER KOMPETENZBEGRIFF

Der Begriff der *Kompetenz* ist ein in Wissenschaft und Alltagssprache viel gebrauchter, wobei definitorische Grenzen meist verschwimmen oder von Beginn an gar nicht erst klar gezogen werden. Der häufige Gebrauch lässt diesen Begriff schon fast abgenutzt erscheinen und dennoch wird er noch immer gern als Modebegriff bezeichnet. Doch weder die Metapher der Abnutzung noch die Bezeichnung als Modebegriff werden der Bedeutung des Kompetenzbegriffs im Rahmen der Erfassung menschlichen Denkens und Handelns gerecht, sondern werten sie sogar ab. Wenn dieses Denken und Handeln theoretisch begründet und in empirischer Weise gemessen werden soll, so ist es unmöglich, den Begriff der *Kompetenz* unbeachtet zu lassen.

An dieser Stelle soll auf einen geschichtlichen Rückblick auf den Begriff der Kompetenz verzichtet werden. Ziel dieses Abschnitts ist es vielmehr, den Kompetenzbegriff, der dieser Testentwicklung zu Grunde liegt, zu definieren und von anderen Kompetenzdefinitionen abzugrenzen. Im Folgenden werden auch die Begriffe *Fähigkeit* und *Fertigkeit* definiert, da sie für die Erstellung des Rahmenkonzepts zur Testentwicklung und für die Erstellung des späteren Messmodells von Bedeutung sein werden. Eine präzise Modellierung der Kompetenz wird die später folgende Operationalisierung und die Entwicklung valider Testaufgaben ermöglichen.

Der Kompetenzbegriff, der die Testentwicklung im Rahmen dieser Arbeit begründet, lehnt sich an ein Konzept von Weinert (1999) an, der in seinem OECD-Bericht *Concepts of Competence* bestehende Kompetenzdefinitionen aufgriff und versuchte, essentielle theoretische Bausteine zu finden, die eine anwendbare Definition bilden können. In Anlehnung an die unterschiedlichen Konzeptualisierungen, die Weinert unterschied, werden Kompetenzen betrachtet als:

> *„funktional bestimmte, auf bestimmte Klassen von Situationen und Anforderungen bezogene kognitive Leistungsdispositionen, die sich psychologisch als Kenntnisse, Fertigkeiten, Strategien, Routinen oder auch bereichsspezifische Fähigkeiten beschreiben lassen."* (Klieme et al., 2001, S.182).

Aus dieser Konzeptualisierung erarbeiteten Klieme et al. (2001) eine Arbeitsdefinition von Kompetenz, die in Richtung einer Anwendung von Kenntnissen, Fähigkeiten und Fertigkeiten zur Bewältigung bestimmter Anforderungen und Aufgaben zielt:

> *„Kompetenzen sind Systeme aus spezifischen, prinzipiell erlernbaren Fertigkeiten, Kenntnissen und metakognitivem Wissen, die es erlauben, eine Klasse von Anforderungen in bestimmten Alltags-, Schul- oder Arbeitsumgebungen zu bewältigen."* (Klieme et al., 2001, S.182).

Verbindet man die Konzeptionen von Weinert (1999) und Klieme (2001), so machen folgende Merkmale den Kompetenzbegriff aus, welcher der Testentwicklung im Rahmen dieser Arbeit zu Grunde liegt:

- Der Begriff der Kompetenz wird für Kenntnisse, kognitive Fähigkeiten, Fertigkeiten sowie für Strategien und Routinen gebraucht. Motivationale Orientierungen werden getrennt davon erfasst (Weinert, 1999; Klieme et al., 2001).
- Kompetenzen sind prinzipiell bereichsspezifisch, d.h. auf einen begrenzten Sektor von Kontexten und Situationen und auf bestimmte Klassen von Situationen und Anforderungen bezogen (Weinert, 1999; Klieme et al., 2001).
- Kompetenzen sind Dispositionen (also Leistungsmöglichkeiten) und bilden das Leistungspotential einer Person. Die Zuschreibung einer Kompetenz geht über die Feststellung einzelner konkreter Leistungen (*Performanz*) hinaus (Weinert, 1999; Klieme et al., 2001).
- Kompetenzen sind funktional definiert, d.h. Indikator einer Kompetenz ist die Bewältigung bestimmter Anforderungen (Weinert, 1999; Klieme et al., 2001).
- Kompetenzen umfassen das bereichsspezifische Leistungspotential einer Person und dessen Anwendung in bestimmten Anforderungssituationen (Weinert, 1999).
- Kompetenzen sind erlernbar: Sie können durch Lernen erworben, sowie durch Interventionen beeinflusst und durch langjährige Praxis in einem bestimmten Feld ausgebaut werden (Klieme & Hartig, 2008).

Die genannten Punkte heben insbesondere hervor, dass Personen bestimmte Kompetenzen nicht unabhängig von Kontexten besitzen und dass Kompetenz damit keine allein an die Person gekoppelte Eigenschaft ist. Diese Bereichsspezifität und die Eigenschaft der Erlernbarkeit grenzen diesen Kompetenzbegriff vom Begriff der *Allgemeinen Intelligenz* ab.

Was die Abgrenzung des Kompetenzbegriffs von motivationalen Einflüssen angeht, so ist es zunächst wichtig festzuhalten, dass sie oft als Teilaspekte von Kompetenz und damit als untrennbar mit ihr verbunden betrachtet werden. Dennoch lässt sich Motivation als Konstrukt von der Kompetenz in Form von messbarer Leistung trennen. Metaanalysen zeigen, dass ein durchschnittlicher Zusammenhang zwischen allgemeiner Motivation und Leistung von $r = 0,12$[7] (Fraser, Walberg, Welch & Hattie, 1987) besteht. Dieser Wert ist vergleichsweise gering. Wird speziell der Zusammenhang zwischen intrinsischer[8] Lernmotivation und Schul- und Studienleistung betrachtet steigt die Korrelation auf $r = 0,23$ (Schiefele & Schreyer, 1994). In der Organisationspsychologie wird Motivation trotz der geringen Korrelationen als eine der Determinanten für die Entwicklung und vor allem für die Ausschöpfung des Kompetenzpotentials in Form von Leistung betrachtet (Rosenstiel, 2003). Die Leistung (L) wird dabei z.B. als Funktion von Fähigkeit bzw. Fertigkeit (F) und Motivation (M) interpretiert, wobei F und M multiplikativ verknüpft sind (Vroom, 1964).

Es ist wichtig zu beachten, dass Fähigkeit bzw. Fertigkeit und Motivation in diesen Untersuchungen als unterschiedliche Konstrukte betrachtet werden. Die gleiche Trennung soll auch für das Kompetenzkonzept dieser Arbeit gelten. Soll der Einfluss von Motivation messbar sein, so sollte sie nicht als Teil des Kompetenzkonzepts definiert werden. Im Rahmen dieser Testentwicklung wird von einer motivationszentrierten Definition von Kompetenz Abstand genommen.

Nachdem nun der Kompetenzbegriff definiert und von anderen Konzepten abgegrenzt wurde, ist es nötig, die Begriffe *Fähigkeit* und *Fertigkeit* für den Kontext dieser Testentwicklung zu definieren. Ohne bereits detailliert und inhaltlich auf die Dinge einzugehen, die der Test letztendlich messen soll, werden die Begriffe *Fähigkeiten* und *Fertigkeiten* zunächst allgemein definiert.

Die meisten Autoren sehen Kompetenzen als Bündel von Fähigkeiten und Fertig-

7 Der Zusammenhang zwischen Variablen wird in Form von Korrelationen dargestellt, die mit r abgekürzt werden. Variablen können in positivem oder negativem Zusammenhang stehen und das Maß kann Werte zwischen -1 und +1 annehmen. Der Zusammenhang ist um so größer, je mehr sich die Korrelation dem Wert 1 nähert.

8 Von intrinsischer Motivation wird gesprochen, wenn eine Handlung aus eigenem Antrieb, Interesse, Bedürfnis oder eigener Freude erfolgt und nicht hauptsächlich durch äußere Anreize bestimmt wird. (E. R. Smith & Mackie, 2007)

keiten (Weinert, 1999; Klieme et al., 2001; Frey, 2006). Diese Definition siedelt den Kompetenzbegriff hierarchisch über den beiden anderen Begriffen an. Kompetenz und Fähigkeit werden als unterschiedliche theoretische Konstrukte gesehen, die nur über Fertigkeiten strukturiert und diagnostiziert werden können (Frey, 2006). Fertigkeiten werden als Indikatoren für die zu Grunde liegende Kompetenz gesehen. Sie sind erlernbar und stellen die messbare Anwendung verfügbarer Fähigkeiten zur Erfüllung konkreter Aufgaben dar (Weinert, 2001). Dabei werden die Handlungen, das konkrete und inhaltlich bestimmbare Können zur Bewältigung der Anforderungen, mit Präzision und gewisser Leichtigkeit aufgrund von Übung ausgeführt (Weinert, 1999; Frey, 2006). Durch messbare Fertigkeiten werden die theoretisch begründeten Fähigkeiten praktisch definiert. Die Fähigkeiten selbst werden als theoretische Konzepte, als spezialisierte Leistungspotentiale gesehen, die in ihrer Gesamtheit eine Kompetenz ausmachen. Sie bündeln alle psychischen und physischen Fertigkeiten (Frey, 2006).

Abbildung 2.1 gibt in Anlehnung an Frey (2006) zum Abschluss dieses Abschnitts die strukturelle Beziehung von Kompetenz, Fähigkeit und Fertigkeit wieder, die dieser Arbeit zu Grunde liegt. Dabei ist zu beachten, dass es sich lediglich um eine schematische Darstellung der Beziehung zwischen Kompetenz, Fähigkeit und Fertigkeit handelt.

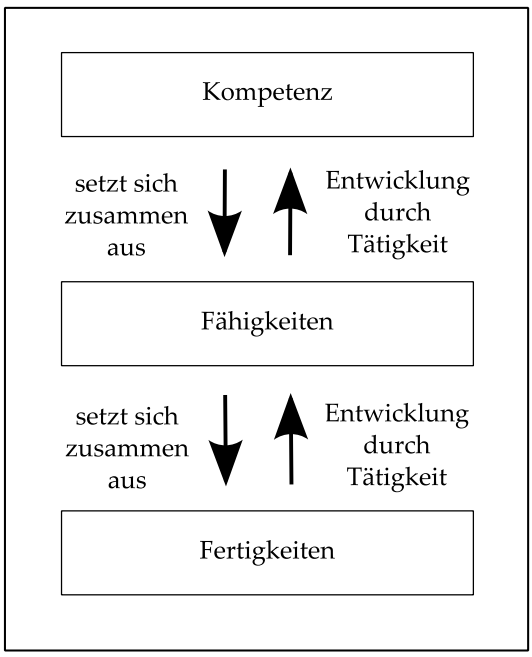


Abbildung 2.1: Verhältnis von Kompetenz, Fähigkeit und Fertigkeit

DER BEGRIFF DER PERFORMANZ

Wenn man sich mit dem Begriff und der Messung von Kompetenz beschäftigt, ist eine Auseinandersetzung mit dem Begriff der *Performanz* als Form der praktischen Umsetzung zu Grunde liegender Kompetenz unumgänglich. Wie im vergangenen Abschnitt bereits erwähnt, lässt sich Kompetenz als theoretisches Leistungspotential erst anhand bestimmter Handlungen und anhand der Bewältigung bestimmter Aufgaben erschließen. Aus diesem Grund ist es für das Verständnis des Performanzbegriffs wichtig, ihn für den weiteren Gebrauch im Rahmen dieser Arbeit zu definieren.

Der Linguist Noam Chomsky (1965) stellte die beiden Begriffe der Kompetenz und Performanz in seinem Werk *Aspects of the Theory of Syntax* als erster gegenüber. Aus seiner Sicht unterscheidet sich die Erforschung der Sprachverwendung in keiner Weise von der Erforschung anderer komplexer Phänomene. Aus diesem Grund kann die Unterscheidung von Performanz und Kompetenz im Rahmen von Sprache und Sprachverwendung hier auch beispielhaft für andere Kompetenzen und Kontexte stehen. Chomsky unterscheidet in seiner Untersuchung von Sprache und Grammatik die Begriffe *Sprachkompetenz* und *Sprachverwendung (Performanz)* folgendermaßen: *Kompetenz* beschreibt er als Kenntnis der eigenen Sprache, als unbewusstes Wissen darüber, wie man Sprache produziert. *Performanz* dagegen definiert er als aktuellen Gebrauch der Sprache in konkreten Situationen, als wirklichen sprachlichen Output. Die Performanz könne die Sprachkompetenz exakt reflektieren, jedoch sei dieser Fall relativ unwahrscheinlich. Begrenztes Gedächtnis, Zerstreutheit, Verschiebung in der Aufmerksamkeit und (zufällige oder typische) Fehler würden eine vollständige Widerspiegelung der Kompetenz in der Performanz weitgehend unmöglich werden lassen. Chomsky zufolge besteht ein so direktes Verhältnis zwischen Kompetenz und Performanz nicht. Dennoch geht er davon aus, dass es eine nicht zugängliche mentale Realität gibt, die einem aktuellen Verhalten zu Grunde liegt. Beobachtungen des Sprachgebrauchs könnten Evidenzen für die Beschaffenheit der mentalen Realität und für grundlegende Strukturen liefern, seien jedoch nicht mit ihnen gleichzusetzen. Die Sprachkompetenz sei keiner direkten Beobachtung zugänglich und könne nicht durch induktive Prozeduren aus gesammelten Daten extrahiert werden. Um signifikante Informationen über die Sprachstruktur überhaupt einholen zu können, bedürfe es zuverlässiger Techniken und Verfahren.

Es zeigt sich, dass Chomskys Kompetenzbegriff insofern mit der Definition dieser Arbeit übereinstimmt, als er Kompetenz als ein Konstrukt sieht, das Verhalten zu Grunde liegt und keiner direkten Messung zugänglich ist. Performanz, in seinem Fall der sprachliche Output, also das gezeigte Verhalten, ermöglicht es, auf die zu Grunde

liegende Kompetenz zu schließen. Allerdings zweifelte er an einer direkten Verbindung zwischen Kompetenz und Performanz und an den Verfahren, die gezeigtes Verhalten erfassen. Beide Punkte müssen jedoch erfüllt sein, wenn Kompetenzdiagnostik funktionieren soll. Eine definitorische Trennung von Kompetenz und Performanz macht Sinn, auch wenn sie willkürlich erscheint (Weinert, 1999). Jedoch müssen die beiden korrelativ verbunden sein, um von der Performanz als gezeigtem und damit messbarem Verhalten auf die Kompetenz schließen zu können. Diese Annahmen beschreiben das Verhältnis von Kompetenz und Performanz als eine der Grundlagen dieser Arbeit.

Die Betrachtungen zur Beziehung zwischen Kompetenz und Performanz zielen auf einen weiteren bedeutsamen Aspekt dieser Arbeit ab, nämlich auf die Validität. Ein Verfahren zur Erfassung eines bestimmten Verhaltens muss genau das messen, was es zu messen beansprucht. Instrumente zur Erfassung von Verhalten müssen valide sein, da ansonsten Überlegungen zur Verbindung von Performanz und Kompetenz überflüssig werden. Das Entwicklungsverfahren im Rahmen dieser Arbeit soll ein solches Testinstrument hervorbringen. Die Performanz besteht im Testverhalten, das anhand des *Tests zur Erfassung prozessbezogener naturwissenschaftlicher Grundbildung* erfasst wird. Von den erreichten Testwerten wird anschließend auf die Ausprägung des zu Grunde liegenden Leistungspotentials, also auf die Kompetenz, in diesem Falle auf die prozessbezogene naturwissenschaftliche Grundbildung, geschlossen.

LATENTE UND MANIFESTE VARIABLEN

In den vergangenen Abschnitten wurde immer wieder von Kompetenz und messbarer Leistung oder auch Performanz gesprochen. Da angenommene theoretische Konstrukte, wie z.B. Kompetenzen, keiner direkten Beobachtung oder Messung zugänglich sind, müssen sie operationalisiert, also messbar gemacht werden. Es muss also beobachtbares Verhalten gefunden werden, das mit dem definierten Konstrukt in Verbindung steht. Dazu muss ein Messmodell generiert werden, welches die Annahmen über die Verbindung von beobachtbarem Verhalten und zu Grunde liegendem Konstrukt darstellt.

An dieser Stelle ist es notwendig, in testtheoretische Begrifflichkeiten einzuführen und die Begriffe *latent* und *manifest* zu definieren. Der Abschnitt dient nur einer kurzen Einführung in diese Begrifflichkeiten, die im Rahmen der Strukturierung der prozessbezogenen naturwissenschaftlichen Grundbildung in diesem Kapitel eine Rolle spielen. Tiefergehende testtheoretische Betrachtungen folgen in Abschnitt 3.1.

Der Begriff *latent* beschreibt Konstrukte, die theoretisch angenommen werden und

keiner direkten Messung zugänglich sind. Als *manifest* werden Variablen bezeichnet, die als messbare Indikatoren für latente Konstrukte fungieren. So können in Anlehnung an Abbildung 2.1 die Kompetenz und die Fähigkeiten, durch die sie repräsentiert wird, als latent bezeichnet werden, während die Fertigkeiten als Operationalisierung der Fähigkeiten als manifest bezeichnet werden können. Auf diesen Annahmen fußt die Item-Response-Theorie (IRT): Aus einem Testverhalten, der Beantwortung bestimmter Testitems, wird auf Merkmalsausprägungen in nicht beobachtbaren dahinterliegenden Fähigkeiten und Kompetenzen geschlossen, von welchen das manifeste Verhalten als abhängig gesehen wird (Rost, 2004a; Moosbrugger, 2007).

2.2.2 DER PROZESS EXPERIMENTELLER ERKENNTNISGEWINNUNG

Neben der grundsätzlichen Definition des Kompetenzbegriffs für diese Testentwicklung besteht einer der ersten Schritte zunächst darin festzustellen, auf welche Komponenten wissenschaftlicher bzw. genauer experimenteller Erkenntnisgewinnung sich die Testentwicklung stützen wird und welche Fertigkeiten auf Seiten der Testpersonen damit erfasst werden. Eine Einschränkung des zu testenden Bereichs ist hier unbedingt notwendig, da die Testentwicklung im Rahmen dieser Arbeit nicht den gesamten Bereich experimenteller Erkenntnisgewinnung abbilden kann. Um eine Einschränkung vornehmen zu können, gilt es in diesem Abschnitt herauszustellen, welche Elemente für den Zweck der Abbildung des Prozesses experimenteller Erkenntnisgewinnung sinnvoll sind. Weiterhin wird ausgeführt, welche Bedeutung das Wissen über Merkmale und Abläufe wissenschaftlicher Erkenntnisgewinnung im Rahmen naturwissenschaftlicher Grundbildung besitzt.

KOMPONENTEN

Je nach Schwerpunktlegung gibt es viele Möglichkeiten, den Prozess experimenteller Erkenntnisgewinnung zu umschreiben. Daher sind die im Folgenden genannten Punkte nicht als Norm zu verstehen. Diese Art der Erkenntnisgewinnung wird in dieser Arbeit als eine dem Problemlösen ähnliche Aktivität verstanden. Sie wird als ein Findeverfahren für einen Übergang von einem Ausgangspunkt zu einem Zielpunkt gesehen (Dörner, 2006), wobei es in diesem Fall darum geht, vom Ausgangspunkt eines Problems, eines Phänomens oder einer wissenschaftlichen Frage zur Erklärung desselben bzw. derselben als Zielpunkt zu gelangen. Der Weg dorthin wird im Fall experimenteller Erkenntnisgewinnung gestaltet durch das Aufstellen von Hypothesen, das Planen und Durchführen von Experimenten, die Interpretation von Testergebnis-

sen und die Revision von Hypothesen (Klahr & Dunbar, 1988). Schon Gagné verstand wissenschaftliche Erkenntnisgewinnung als:

> *„... set of activities characterized by a problem-solving approach in which each newly encountered phenomena becomes a challenge for thinking. Such thinking begins with a careful set of systematic observations, proceeds to design the measurements required, clearly distinguishes between what is observed and what is inferred, invents interpretations which are under ideal circumstances brilliant leaps, but always testable, and draws reasonable conclusions."* (Gagné, 1963, S. 145)

Hierbei ist zu beachten, dass experimentelle Erkenntnisgewinnung nicht mit einfachem Problemlösen gleichgesetzt werden soll. Es geht lediglich darum, einen möglichen Weg zu beschreiben, der von einem definierten Ausgangspunkt in der experimentellen Erkenntnisgewinnung zu einem definierten Zielpunkt führt. Dieser Zielpunkt stellt nicht das Ende des Forschungsprozesses dar. Das *Problem* ist nicht gelöst, sondern es gibt vorläufige Erkenntnisse, die Auswirkungen auf zukünftige Forschung haben und in zukünftiger Forschung auf dem Prüfstand stehen.

Für Gagné (1963) bestand der Prozess aus folgenden Bereichen: Beobachten, Klassifizieren, Beschreiben, Kommunizieren, Messen, Erkennen, Nutzen von Relationen, Schlussfolgern, Erstellen operationaler Definitionen, Formulieren von Hypothesen, Kontrollieren von Variablen, Interpretieren von Daten und Experimentieren. Als kurzer Abriss kann der Prozess wissenschaftlicher Erkenntnisgewinnung folgendermaßen aussehen (die hervorgehobenen Elemente sind Teil des *Tests zur Erfassung prozessbezogener naturwissenschaftlicher Grundbildung*, wobei die beiden zuletzt genannten Punkte dort zusammengefasst werden):

- **Identifikation wissenschaftlicher Fragen/ Fragestellungen**
- Formulierung wissenschaftlicher Hypothesen
- **Planung eines Experiments zur Prüfung der Hypothesen**
- Durchführung des Experiments
- Aufzeichnung der Daten
- Interpretation der Daten
- **Schlussfolgerungen aus Daten ziehen**

- **Beibehalten oder Verwerfen der Hypothesen**

Da es sich bei der Liste nicht um eine vorgegebene Norm des Prozesses wissenschaftlicher Erkenntnisgewinnung handelt, kann und will sie keinen Anspruch auf Vollständigkeit der Tätigkeiten erheben. Sie soll lediglich einen Überblick über mögliche Elemente geben, auf die sich die Testentwicklung beziehen konnte. Weiterhin ist zu beachten, dass die Bezeichnungen *Ausgangs-* und *Zielpunkt* für die Beschreibung dieses Prozesses nur bedingt geeignet sind, da der hier definierte Zielpunkt wiederum der Ausgangspunkt für das neuerliche Aufstellen und Prüfen von Hypothesen sein kann. Es handelt sich um einen fortlaufenden zirkulären Prozess und somit ist die Festlegung eines Anfangs- und Endpunktes zwar logisch, aber auch willkürlich.

Die vorgenommene Auswahl der Elemente ist dadurch begründet, dass sie den Prozess experimenteller Erkenntnisgewinnung gut abbilden. Es handelt sich um Elemente vom Beginn, der Mitte und dem Ende des Prozesses. Außerdem stehen die Elemente in einem starken logischen Zusammenhang. Ohne ein Identifizieren wissenschaftlich untersuchbarer Fragestellungen ist keine Planung wissenschaftlicher Untersuchungen, inklusive der Festlegung zu prüfender Variablen, möglich. Außerdem wird es ohne eine auf die Fragestellung abgestimmte Planung der wissenschaftlichen Untersuchung nicht möglich sein, die Ergebnisse richtig zu interpretieren und adäquate Schlussfolgerungen zu ziehen. Zeigen Testpersonen gute Fähigkeiten in den ausgewählten Bereichen, so sollte man von einem guten allgemeinen Verständnis der Merkmale und Abläufe experimenteller Erkenntnisgewinnung sprechen können.

Die Begründung der Auswahl dieser Elemente als Ankerpunkte für die Testentwicklung folgt in Abschnitt 2.2.3. An dieser Stelle sei nur kurz erwähnt, dass die ausgewählten Komponenten in nationalen wie auch internationalen Bildungsstandards eine Rolle spielen und sich fächerübergreifend als bedeutsam erwiesen haben.

SPEZIELLE BEDEUTUNG DES PROZESSES EXPERIMENTELLER ERKENNTNISGEWINNUNG

Das Wissen um Merkmale und Prozesse experimenteller Erkenntnisgewinnung und die Fertigkeiten, die für diese Art wissenschaftlicher Arbeit benötigt werden, stellen wichtige Bereiche naturwissenschaftlicher Grundbildung dar. Dabei ist die Bedeutsamkeit dieses Bereiches keine neue Entdeckung. Schon Dewey fokussierte in seinem Hauptwerk *Democracy and Education* (1916/1993) besonders auf wissenschaftliche Methoden und den Weg der Erkenntnisgewinnung. Er vertrat die Auffassung, dass Schüler die wissenschaftliche Herangehensweise an Problemstellungen erfahren sollten, anstatt aus Schulbüchern wissenschaftliche Fakten, also die Ergebnisse

zu lernen, *„zu denen die Männer der Wissenschaft gelangt sind“*. Da die meisten Schüler nie wissenschaftliche Experten würden, sei es wichtiger, dass sie einen gewissen Einblick in Wesen und Bedeutung der wissenschaftlichen Methoden gewinnen. Auf diese Weise würden sie zwar weniger Stoff lernen, aber ein grundlegendes Verständnis der Wissenschaft erlangen. 1932 definierte die *National Society for the Study of Education* in ihrem Jahrbuch *A Program for Teaching Science* naturwissenschaftliche Bildung als eine allgemeine Bildung mit der Absicht, dass entsprechend gebildete Menschen sich in der modernen Welt zurechtfinden können. Dies bedeutete insbesondere, dass ein Verständnis der Methoden für eine allgemeine naturwissenschaftliche Bildung von essentieller Bedeutung ist:

> *„The method of science emphasizes careful and accurate observation of controlled and uncontrolled phenomena as a means for determining truth, and it also requires the formulation of hypotheses to explain the relationship of observed phenomena.“* (Education, 1932, S. 38).

Auch für Gagné (1963) bestand das letztendliche Ziel wissenschaftlicher Bildung in der Vermittlung des Prozesses wissenschaftlicher Erkenntnisgewinnung. Dieser sei erlernbar und von domänenübergreifender Bedeutung. Shamos (2002) zufolge ist es gerade der Prozess, *„der die Kraft des logischen Denkens in den Naturwissenschaften zum Ausdruck bringt“*. Er fordert, dass im naturwissenschaftlichen Unterricht der Prozess der Naturwissenschaft stärker hervorgehoben werden sollte, dass Schülerinnen und Schüler lernen sollten, wie man zu dem Wissen gelangt ist, das man heute über die Natur besitzt.

Studien zeigen, dass insbesondere das Wissen um den Ablauf und die Merkmale wissenschaftlicher Erkenntnisgewinnung als Teil naturwissenschaftlicher Grundbildung zu einem verbesserten Lernen und einem umfassenderen Verständnis wissenschaftlicher Sachverhalte führt (Songer & Linn, 1991). Schülerinnen und Schüler, die Erkenntnisse der Wissenschaft nicht als Fakten, sondern als Ergebnisse von Prozessen sehen, sind in der Lage, die Änderung oder Erweiterung aktueller wissenschaftlicher Erkenntnisse sowie die Ursache wissenschaftlicher Entwicklungen einschätzen zu können. Sie sind darauf eingestellt, dass Wissen veränderbar ist, dass aktuelle wissenschaftliche Erkenntnisse den theoretischen Hintergrund für neue Hypothesen und ihre Prüfung anhand von Experimenten bilden und dass sie somit ständig auf dem Prüfstand stehen. Dieses Verständnis darüber, wie wissenschaftliches Wissen sich über die Zeit verändert, kann Schülerinnen und Schüler in die Lage versetzen, ihre eigenen wissenschaftlichen Ideen zu entwickeln (Solomon, 1991). Weitere Studien zeigen Zusammenhänge zwischen erkenntnistheoretischer Überzeugung und

der Lernstrategie von Schülerinnen und Schülern (Edmondson & Novak, 1993; Tsai, 1998). Konstruktivistische Sichtweisen der Personen in Bezug auf den Forschungsprozess waren mit eher aktiven Lernstrategien verbunden und auf die Erlangung eines tieferen Verständnisses von Sachverhalten ausgerichtet, während Personen mit vorwiegend statischer Sichtweise von Forschung die Strategie bevorzugten, Inhalte auswendig zu lernen. Über diesen gefundenen Zusammenhang äußert sich auch Waterman (1982):

> *„If, for example, science is thought of as a body of proven facts, a person might study and memorize facts, and might think that he or she can absolutely prove things in science. If a person thinks of science as an ongoing process of concept development, he or she might learn concepts and their variants."* (Waterman, 1982, S. 5).

Durch Vermittlung einer aktiven Rolle bei der Bearbeitung erkenntnistheoretischer Fragen besteht die Möglichkeit, Schülerinnen und Schüler wegzuführen von einer einfachen passiven Rezeption von Fakten hin zu einer aktiven Konstruktion und Reflexion von Wissen. Der bewusste Umgang mit der Situation wissenschaftlicher Erkenntnisgewinnung kann Schülerinnen und Schüler von der Strategie des Auswendiglernens zu einer Strategie bedeutsamen Lernens führen (Edmondson & Novak, 1993; Anderson, 2007). Erkennen sie den Prozess wissenschaftlicher Erkenntnisgewinnung und insbesondere die Entwicklung und Bewertung wissenschaftlicher Theorien als rationalen und nicht als zufälligen Prozess, so kann dies die Entwicklung eigener Theorien positiv beeinflussen (Duschl, Hamilton & Grandy, 1992). Explizites Wissen darüber, wie Theorie und Evidenz zusammenhängen, kann Schülerinnen und Schüler befähigen, wissenschaftlich zu argumentieren (Carey & Smith, 1993; Kuhn et al., 1988).

Auch wenn es laut Lederman (2007) noch an systematischen Befunden hinsichtlich der Wirkungen des Wissens um Ablauf und Merkmale wissenschaftlicher Erkenntnisgewinnung fehlt und es ebenso sein kann, dass dieses Wissen *nur* zu einem besseren Verständnis der Naturwissenschaften als Disziplin führt, so zeigen diese Darstellungen doch, dass erkenntnistheoretisches Wissen als ein wichtiges Element naturwissenschaftlicher Grundbildung gesehen wird und schon sehr früh gesehen wurde. Die Aktualität und Bedeutsamkeit dieses Bereiches wird dadurch verdeutlicht, dass sie in Form der *Scientific Enquiry* Eingang in das *Scientific Literacy Framework* gefunden hat, das dem naturwissenschaftlichen Schwerpunkt der PISA-Untersuchung (Prenzel, Artelt et al., 2007a) zu Grunde liegt. Die Tatsache, dass dieser Bereich, der Kompetenzen wie das Identifizieren wissenschaftlicher Fragen, das wissenschaftliche

Erklären von Phänomenen und das Nutzen wissenschaftlicher Evidenzen beinhaltet, einen Testbereich dieser bedeutsamen internationalen Studie darstellt, verdeutlicht, welchen Stellenwert er national wie auch international einnimmt.

Zusammenfassend zeigt sich, dass der Prozess wissenschaftlicher Erkenntnisgewinnung als Bereich naturwissenschaftlicher Grundbildung sowohl im lebensweltlichen Alltag, wenn es beispielsweise um die Einordnung und Bewertung wissenschaftlicher Erkenntnisse geht, als auch als Grundlage für das Erlernen weiterer wissenschaftlicher Inhalte von großer Bedeutung ist. Allerdings gibt es bisher, abgesehen von den vereinzelt frei verfügbaren PISA-Items, noch keinen validen, aussagekräftigen und ökonomischen Test zur Erfassung des speziellen Bereiches, der im Rahmen dieser Arbeit als *prozessbezogene naturwissenschaftliche Grundbildung* bezeichnet wird. Ein Test ist jedoch notwendig, um feststellen zu können, in welchem Maße diese Grundbildung in Personen ausgeprägt ist und um die Wirkung eventueller Interventionen auf das Ausmaß dieser Grundbildung messen zu können.

2.2.3 GRUNDLAGEN FÜR BESTIMMUNG UND DEFINITION DER ZU MESSENDEN FERTIGKEITEN

Wie bereits in Abschnitt 2.2.2 angedeutet, bestand eine der Grundlagen für die Testentwicklung zunächst darin festzustellen, welche Elemente experimenteller Erkenntnisgewinnung von besonderer Bedeutung sein könnten. Dabei ging es darum, eine Auswahl zu treffen, die den Prozess der Erkenntnisgewinnung gut abzubilden vermag. Ein wichtiger Punkt, der diese Auswahl unter anderem leitete, war die übergreifende Bedeutsamkeit der Elemente für die drei Naturwissenschaften Physik, Chemie und Biologie.

Letztendlich wurden drei Elemente ausgewählt: das *Identifizieren wissenschaftlicher Hypothesen*, das *Planen einer wissenschaftlichen Untersuchung* und das *Nutzen wissenschaftlicher Ergebnisse*. Eine Beschränkung des zu testenden Bereichs war unbedingt notwendig, da die Testentwicklung im Rahmen dieser Arbeit nicht den gesamten Bereich experimenteller Erkenntnisgewinnung abbilden konnte.

Zur Erläuterung der Auswahl dieser Elemente werden nun die nationalen Bildungsstandards der Fächer Physik, Chemie und Biologie für den mittleren Bildungsabschluss (KMK, 2005), die *National Science Education Standards* des amerikanischen *National Research Council* (NRC, 1996), das so genannte *SDDS-Modell* (Klahr & Dunbar, 1988) und aktuelle *Scientific-Literacy*-Konzeptionen wie das der *PISA-Studie* 2006, das *Project 2061* (American Association of the Advancement of Science, 2001) und *GCSE Additional Applied Science - Scientific Detection* (University of York Science Educati-

on Group, 2006) betrachtet. Diese Quellen werden verdeutlichen, welche Bestandteile experimenteller Erkenntnisprozesse nationale wie auch internationale Expertengruppen für bedeutsam halten. Der *Test zur Erfassung naturwissenschaftlicher Grundbildung* orientierte sich an ihnen.

Am Ende dieses Abschnitts findet sich eine erste schematische Darstellung der *prozessbezogenen naturwissenschaftlichen Grundbildung*, welche eine erste Grundlage der Entwicklung von Testaufgaben bildete. Diese Aufgabenentwicklung stellt den zentralen und folgenreichsten Punkt dieser Arbeit dar. Nur aufgrund genauer Definitionen der ausgewählten Elemente und der dementsprechend zu messenden Fähigkeiten kann es gelingen, die Aufgaben so zu entwickeln, dass sie valide das erfassen, was sie erfassen sollen.

EXPERIMENTELLE ERKENNTNISGEWINNUNG IN NATIONALEN UND INTERNATIONALEN BILDUNGSSTANDARDS

Die Kultusministerkonferenz der Länder (KMK) beschloss im Dezember 2004 die Bildungsstandards der Fächer Biologie, Physik und Chemie für den mittleren Schulabschluss (KMK, 2005a, 2005c, 2005b). Hier wurde festgelegt, welchen Bildungsstand Schülerinnen und Schüler am Ende der neunten Klasse in den entsprechenden Fächern erreicht haben sollten. Der Bereich wissenschaftlicher Erkenntnisgewinnung spielt in diesen Standards neben den Bereichen Fachwissen, Kommunikation und Bewertung eine wichtige Rolle.

Der folgende Abschnitt wird die Bedeutsamkeit der ausgewählten Elemente experimenteller Erkenntnisgewinnung unterstreichen und erläutern, in welcher Form sie in die Bildungsstandards eingegangen sind. Die Inhalte wurden als Quellen für eine genaue Beschreibung der Fertigkeiten herangezogen, die die Grundlage der Aufgabenentwicklung bildeten.

Tabelle 2.2 zeigt, dass die aus dem Prozess experimenteller Erkenntnisgewinnung ausgewählten Elemente in sehr ähnlicher Form übergreifend in allen drei Fachbereichen vorkommen. Es fällt auf, dass das *Identifizieren wissenschaftlicher Fragestellungen* in der Liste der Biologie-Standards fehlt. In den ausformulierten Standards wird jedoch auf diesen Punkt eingegangen. Betrachtet man internationale Standards für naturwissenschaftliche Bildung, so wird deutlich, dass es nicht nur national fächerübergreifende Ähnlichkeiten in der Formulierung der Standards gibt, sondern auch länderübergreifende. Als Beispiel werden in Tabelle 2.2 stellvertretend die *Science Education Standards* (9.-12. Klasse) des Amerikanischen *National Research Council* genannt (NRC, 1996). Die deutschen Standards weisen deutliche Parallelen zum Bereich

Tabelle 2.2: Der Bereich *Erkenntnisgewinnung* der deutschen Bildungsstandards (Physik, Biologie, Chemie) und der Bereich *Science as Inquiry* der amerikanischen Bildungsstandards

Elemente / Quelle	**Identifizieren wissenschaftlicher Hypothesen**	**Planen einer wissenschaftlichen Untersuchung**	**Nutzen wissenschaftlicher Ergebnisse**
Bildungsstandards Physik (für den mittleren Bildungsabschluss)	Die Schüler stellen an einfachen Beispielen Hypothesen auf.	Die Schüler planen einfache Experimente, [...].	Die Schüler werten gewonnene Daten aus, [...].
Bildungsstandards Biologie (für den mittleren Bildungsabschluss)		Die Schüler planen einfache Experimente, [...].	Die Schüler erörtern Tragweite und Grenzen von Untersuchungsanlage, -schritten und -ergebnissen.
Bildungsstandards Chemie (für den mittleren Bildungsabschluss)	Die Schüler erkennen und entwickeln Fragestellungen, [...].	Die Schüler planen geeignete Untersuchungen zur Überprüfung von Vermutungen und Hypothesen.	Die Schüler finden in erhobenen oder recherchierten Daten Trends, Strukturen und Beziehungen, erklären diese und ziehen geeignete Schlussfolgerungen.
National Research Council (NRC) Science as Inquiry	Identify questions and concepts that guide scientific investigations: Students should formulate a testable hypothesis and demonstrate the logical connections between the scientific concepts guiding a hypothesis and the design of an experiment. [...].	Design and conduct scientific investigations: [...]. The investigation may also require student clarification of the question, method, controls and variables; student organization and display of data; student revision of methods and explanations; and a public presentation of the results with a critical response.	Formulate and revise scientific explanations and models using logic and evidence: Student inquiries should culminate in formulating an explanation or model. [...]. These discussions should be based on scientific knowledge, the use of logic, and evidence from their investigation [...].

Science as Inquiry der amerikanischen Standards auf. Es zeigt sich, dass die amerikanischen Standards in den drei ausgewählten Bereichen sehr genaue Angaben darüber machen, welche Anforderungen an Schülerinnen und Schüler gestellt werden. Diese Angaben wurden für die genaue Definition der Fertigkeiten, die anhand des *Tests zur*

Erfassung prozessbezogener naturwissenschaftlicher Grundbildung erfasst werden sollen, ebenso herangezogen, wie die Informationen aus dem *SDDS-Modell* und aktuellen *Scientific-Literacy*-Konzeptionen, die in den folgenden beiden Abschnitten dargestellt werden.

DAS SDDS-MODELL

Das *Scientific-Discovery-As-Dual-Search (SDDS)-Modell* wurde als weitere Quelle für eine Definition der zu testenden Fertigkeiten herangezogen. Es stellt den Prozess wissenschaftlicher Erkenntnisgewinnung als eine Form des Problemlösens dar (vgl. Abschnitt 2.2.2). Dieser Vorgang besteht darin, einen Ausgangszustand durch einen Problemlöseprozess in einen gewünschten Zielzustand zu überführen. Da es sich beim SDDS-Modell um ein domänenübergreifendes und nicht auf spezielle Inhalte festgelegtes Modell handelt, passt es gut in die bisherigen Überlegungen. Die dem Modell zu Grunde liegende Annahme besteht darin, dass wissenschaftliche Erkenntnisgewinnung aus der Suche in zwei Problemräumen, dem Hypothesen- und dem Experimentraum, besteht. Die Wissensbasis über eine bestimmte Domäne stellt den Ausgangszustand dar. Der Zielzustand wird durch die Hypothese repräsentiert, die dieses Wissen möglichst prägnant und universell erklärt. Der Weg dorthin besteht im Problemlösen. Die Suche innerhalb und zwischen den Problemräumen wird durch die Hauptkomponenten geleitet, die in Abbildung 2.2 dargestellt sind (Klahr, 2000, S. 31). Die *Suche im Hypothesenraum* hat das Ziel, eine wissenschaftlich untersuchbare

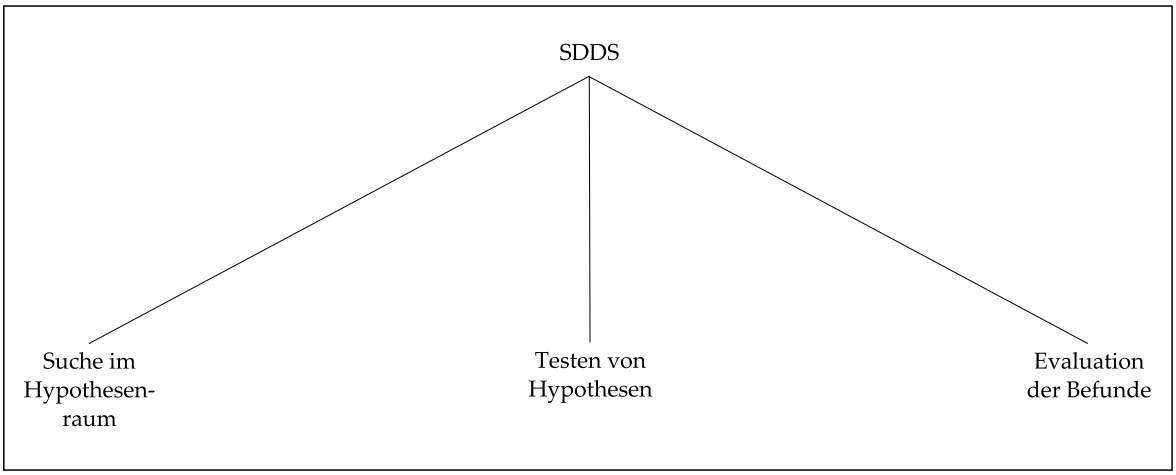


Abbildung 2.2: Die drei Hauptkomponenten des SDDS-Modells

Hypothese zu finden. Dabei gibt es zwei Quellen für das Suchen nach neuen Hypo-

thesen. Zum einen können sie aus gespeichertem Wissen resultieren, zum anderen können sie experimentellen Befunden oder Beobachtungen entstammen.

Das *Testen von Hypothesen* beinhaltet das Generieren eines zur Hypothese passenden Experiments, hier findet also ein Abgleich zwischen der Hypothese und der Suche im Experimentraum statt. Weitere Komponenten sind das Aufstellen einer Vorhersage, das Durchführen des Experiments und die Feststellung der möglichen Diskrepanz zwischen Vorhersage und Ergebnis des Experiments.

Die *Evaluation der Befunde* besteht in einer Entscheidung, ob sie ausreichen, um die Hypothese zu verwerfen oder beizubehalten. Ist dies nicht der Fall, müssen weitere Experimente durchgeführt werden bis eine ausreichende Entscheidungsgrundlage vorhanden ist.

Auch hier wird deutlich, dass die gleichen Elemente im Vordergrund des Modells stehen, die auch der Testentwicklung zu Grunde liegen. Die Definitionen der einzelnen Bereiche sind ebenso wie die Bildungsstandards richtungsweisend für die Definition der im Rahmen des Tests zu messenden Fertigkeiten. Sie werden in der Erstellung des Strukturmodells der *prozessbezogenen naturwissenschaftlichen Grundbildung* am Ende dieses Abschnitts noch einmal aufgegriffen.

AKTUELLE SCIENTIFIC-LITERACY-KONZEPTIONEN

Als dritte Quelle für die Definition von Inhalt und Struktur der *prozessbezogenen naturwissenschaftlichen Grundbildung* dienten aktuelle *Scientific-Literacy*-Konzeptionen, die im Folgenden dargestellt werden.

Den Anfang macht das in Abbildung 2.3 dargestellte Rahmenkonzept der PISA-Untersuchung 2006 (OECD, 2006), dessen Schwerpunkt auf dem Bereich der Naturwissenschaften lag. Die Abbildung macht deutlich, dass zwei der drei genannten Kompetenzen den Elementen gleichen, die der Entwicklung des *Tests zur Erfassung prozessbezogener naturwissenschaftlicher Grundbildung* zu Grunde liegen sollten: das Erkennen naturwissenschaftlicher Fragestellungen und das Nutzen wissenschaftlicher Evidenz. Ein weiterer, für die vorliegende Testentwicklung bedeutsamer, Aspekt des Rahmenkonzepts bestand in einer der Variablen, die mit den genannten naturwissenschaftlichen Kompetenzen zusammenhängen. Gemeint ist damit das Interesse an den Naturwissenschaften. Dieser Bereich konnten der Arbeit als Quellen für die Suche nach Validierungsvariablen dienen und werden im Rahmen dieses Kapitels in Abschnitt 2.6 noch einmal aufgenommen und diskutiert.

Eine amerikanische Scientific-Literacy-Konzeption besteht im Rahmen des *Project 2061* (American Association of the Advancement of Science, 2001) und beruht auf

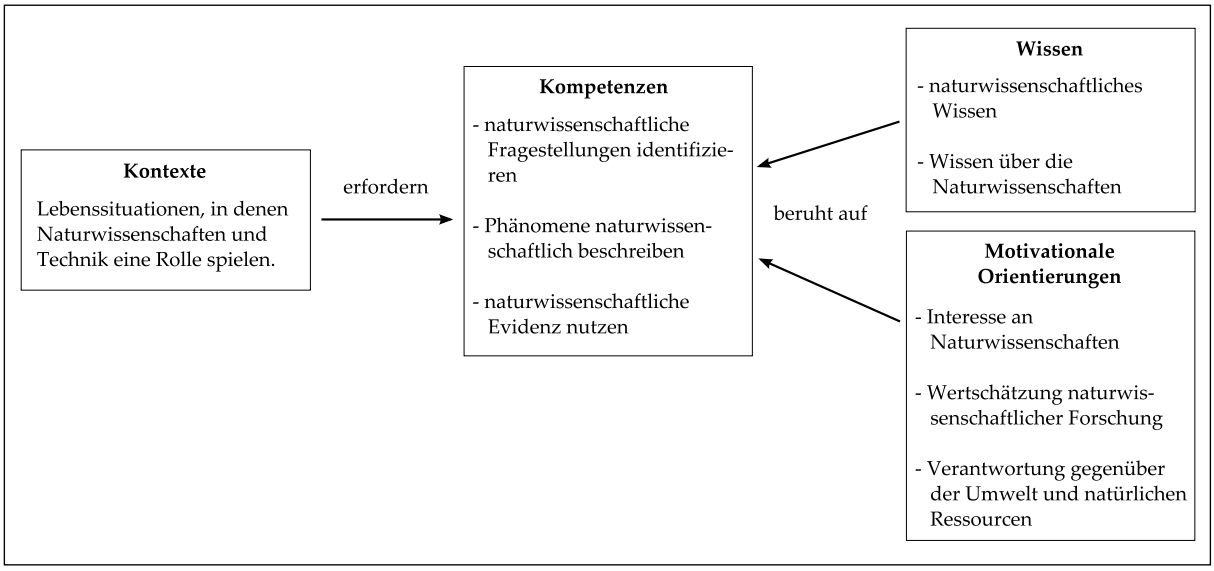


Abbildung 2.3: Das PISA-Rahmenkonzept für den Bereich Naturwissenschaften

dem Report *Science for all Americans* (American Association of the Advancement of Science, 1989). Sie enthält in ihrer Definition der Scientific Literacy das Verständnis:

- der Natur der Naturwissenschaften, der Mathematik und der Technologie,
- der Welt, wie sie aktuell durch Naturwissenschaft und Mathematik abgebildet und durch Technologie beeinflusst wird,
- zentraler Begebenheiten in der Geschichte wissenschaftlichen Strebens,
- der Themen, die die Naturwissenschaften, Mathematik und Technologie durchziehen und dadurch beleuchten, wie die Welt funktioniert,
- der Denkprozesse, die für eine naturwissenschaftliche Grundbildung essentiell sind.

Für den Rahmen der aktuellen Betrachtungen ist hier vor allem das Verständnis der *Natur der Naturwissenschaften* von Bedeutung. Abbildung 2.4 zeigt einen Ausschnitt der die Struktur verdeutlicht. Die Darstellung konzentriert sich auf die Komponenten, die Gemeinsamkeiten mit bereits vorgestellten Konzeptionen aufweisen. Es ist zu erkennen, dass zur *Scientific-Literacy*-Erhebung des *Project 2061* das Prüfen von Hypothesen und das Ziehen von Schlussfolgerungen aus Daten eine Rolle spielen. Die Kontrolle von Variablen findet sich in den bisherigen Ausführungen auch unter dem Planen einer wissenschaftlichen Untersuchung (s. *Science as Inquiry* in Abbildung 2.2 auf S. 32).

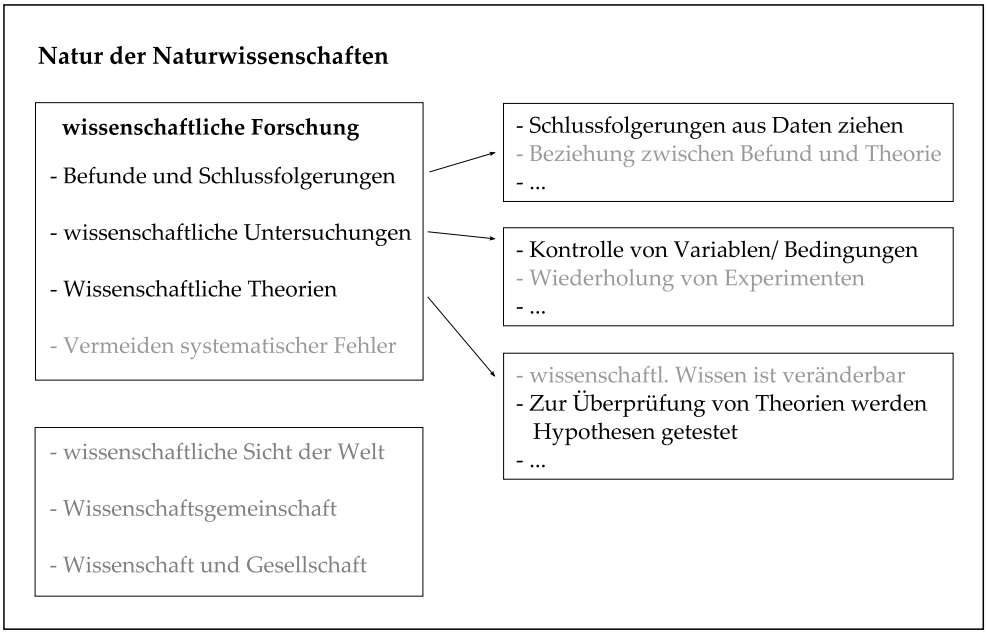


Abbildung 2.4: Ausschnitt der *Scientific-Literacy*-Konzeption der *American Association for the Advancement of Science*

Das letzte Beispiel einer aktuellen Konzeption zur Messung naturwissenschaftlicher Grundbildung ist das der *21st Century Science* (University of York Science Education Group, 2006). Hier gliedert sich der Begriff *Scientific Literacy* in die Bereiche *Vorstellungen von Naturwissenschaften (Reflexion über naturwissenschaftliches Wissen)* und *Erklärungen in den Naturwissenschaften (naturwissenschaftliche Konzepte)*, eine Unterteilung, welche an die PISA-Rahmenkonzeption erinnert, die *Wissen über die Naturwissenschaften* und *naturwissenschaftliches Wissen* unterscheidet.

Betrachtet wird an dieser Stelle der Punkt *Vorstellungen von den Naturwissenschaften*. Er ist durch die Fähigkeit definiert, über naturwissenschaftliches Wissen nachdenken zu können und umfasst die Methoden, die das Wissen hervorgebracht haben, schlussfolgerndes Denken, um wissenschaftliche Argumente zu entwickeln und alle Sachverhalte, die eine Rolle spielen, wenn das Wissen zur praktischen Anwendung kommt. Das Konzept umfasst konkret folgende Punkte:

- Daten und ihre Grenzen (Interpretieren von Daten, Bedeutung von Daten für Theorien),
- Korrelationen und Einflussgrößen,
- Theorien (Verständnis davon, wie wissenschaftliche Evidenzen genutzt werden, um Hypothesen aufzustellen und zu testen),

- die Wissenschaftsgemeinschaft,
- Risiko,
- Entscheidungen hinsichtlich Naturwissenschaft und Technologie treffen.

Im Falle diese Konzepts zeigen sich Gemeinsamkeiten mit bereits genannten Konzepten in den Punkten *Daten und ihre Grenzen* sowie *Theorien*.

Zusammenfassend ist festzuhalten, dass sich die genannten Quellen zwar oft hinsichtlich ihrer Gesamtkonzeption unterscheiden, sich aber bezüglich ihrer Konzeption der prozessbezogenen Bereiche naturwissenschaftlicher Grundbildung ähneln. Es ist zwar davon auszugehen, dass sich die Konzeptionen gegenseitig beeinflussen bzw. beeinflusst haben. Dennoch können ihre Ähnlichkeiten ebenso als Bestätigung des aktuellen Bildes vom prozessbezogenen Bereich naturwissenschaftlicher Grundbildung gesehen werden.

2.2.4 STRUKTUR DER «PROZESSBEZOGENEN NATURWISSENSCHAFTLICHEN GRUNDBILDUNG»

Abschließen soll diesen Abschnitt die Definition der *prozessbezogenen naturwissenschaftlichen Grundbildung*. Um dies zu ermöglichen, hat sich der vergangene Abschnitt ausgehend von der grundlegenden Feststellung des zu Grunde liegenden Kompetenzbegriffs mit dem Prozess experimenteller Erkenntnisgewinnung beschäftigt, auf den sich die Testentwicklung bezieht. Die Entscheidung darüber, welche Elemente dieses Prozesses repräsentativ erfasst werden sollten, wurde unterstützt durch aktuelle Beispiele dessen, was nationale und internationale Bildungsstandards sowie aktuelle Konzeptionen zur Erfassung naturwissenschaftlicher Grundbildung fordern. Obwohl der Wortlaut der einzelnen Quellen natürlich nicht immer exakt gleich ausfiel, konnten doch inhaltlich das *Identifizieren wissenschaftlicher Hypothesen*, das *Planen einer wissenschaftlichen Untersuchung* und das *Nutzen wissenschaftlicher Ergebnisse* übergreifend als bedeutsame Elemente identifiziert werden. Weiterhin konnten anhand der Quellen wichtige Hinweise für die Definition der konkret zu messenden Fertigkeiten gewonnen werden.

Die zu messende Kompetenz der *prozessbezogenen naturwissenschaftlichen Grundbildung* stellt ein latentes Konstrukt dar, das auf theoretischer Ebene aus einzelnen latenten Fähigkeiten und auf praktischer Ebene aus manifesten und damit messbaren Fertigkeiten besteht. Die in diesem Abschnitt dargestellte Struktur umfasst die drei repräsentativ ausgewählten Fähigkeiten und definiert die Fertigkeiten, die der Aufgabenentwicklung als Basis dienen. Die Kompetenz ist insofern bereichsspezifisch, als

sie sich auf den Prozess experimenteller Erkenntnisgewinnung bezieht und sie ist domänenübergreifend, weil sie sich allgemein über die Naturwissenschaften erstreckt. Desweiteren ist sie erlernbar, also durch Interventionen beeinflussbar. Die Kompetenz stellt lediglich ein Leistungspotential dar. Die messbare, manifeste Leistung ist ihr nicht gleichzusetzen. Aufgrund des angenommenen positiven Zusammenhangs zwischen latenter Kompetenz und manifester Leistung kann jedoch auf die Ausprägung der Kompetenz geschlossen werden. Die manifeste Leistung wird hier als Performanz in Form der Anwendung eines Leistungspotentials auf eine bestimmte Anforderungssituation, in diesem Fall in Gestalt experimenteller Erkenntnisgewinnung, gesehen.

Abbildung 2.5 gibt Auskunft über Inhalt und Struktur der *prozessbezogenen naturwissenschaftlichen Grundbildung*. Sie gibt in zusammenfassender Form die Überlegungen der vergangenen Abschnitte wieder. Die Struktur ist bestimmt durch eine Einteilung in latente und manifeste Bestandteile und eine Unterteilung in Kompetenz, Fähigkeit und Fertigkeit. Die Tätigkeiten, die hier auf Ebene der Fertigkeiten aufgeführt wurden, sind so konkret formuliert, dass die Konstruktion von Aufgaben zu ihrer Messung ermöglicht wird.

2.3 KOGNITIVE GRUNDLAGEN DES ERWERBS DER ZU MESSENDEN FERTIGKEITEN

Nachdem in den vorherigen Abschnitten die inhaltlichen Überlegungen und Begründungen der Ziele der Testentwicklung im Vordergrund standen, beschäftigt sich der folgende Abschnitt mit kognitiven Grundlagen, die auf Seiten der Testpersonen vorhanden sein müssen, um die beschriebenen Fertigkeiten messen zu können. Die Fertigkeiten werden in Lernprozessen erworben (vgl. Abschnitt 2.2.1), die erst durch bestimmte kognitive Grundlagen ermöglicht werden. Im Rahmen der Beschreibung dieser Grundlagen wird darauf eingegangen, ab welchem Alter Testpersonen in etwa über sie verfügen.

Zur Integration der Ausführungen dieses Abschnitts in den Zusammenhang der Testentwicklung werden in den darauf folgenden drei Unterabschnitten die kognitiven Grundlagen mit den drei durch den Test zu messenden Fertigkeiten H, P und N verknüpft. Vor dem Hintergrund der kognitiven Entwicklung werden Befunde dargestellt, die Hinweise darauf geben, ab welcher Altersstufe die in Abbildung 2.5 dargestellten Fertigkeiten theoretisch messbar sein sollten.

Allgemein ist festzuhalten, dass die für den Prozess naturwissenschaftlicher Er-

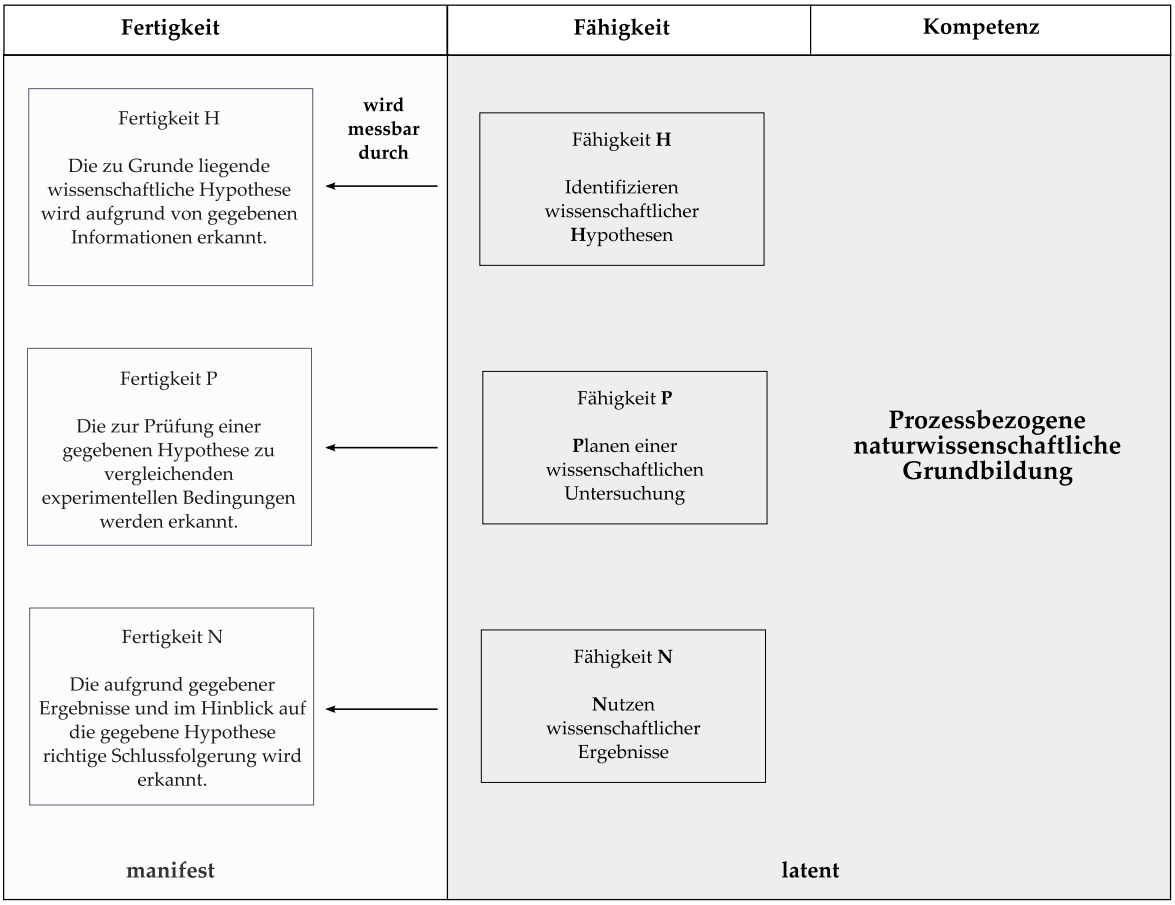


Abbildung 2.5: Struktur der *prozessbezogenen naturwissenschaftlichen Grundbildung*

kenntnisgewinnung notwendigen kognitiven Grundlagen einer Entwicklung über die Kindheit bis hin ins Erwachsenenalter unterliegen, die durch Reifungs- und Lernprozesse bestimmt ist. Damit sind unter anderem auch die mit steigendem Alter zunehmenden Lerngelegenheiten und die damit erweiterten Möglichkeiten der Enkodierung von Informationen angesprochen (Siegler & Alibali, 2005). Weiterhin betrifft diese Entwicklung die zunehmenden Kapazitäten des Arbeitsgedächtnisses, Informationen gleichzeitig bzw. parallel zu verarbeiten, also miteinander in Beziehung zu setzen, zu vergleichen und metakognitiv zu überwachen (Kail & Bisanz, 1992). Die Informationsverarbeitung wird in der Entwicklung vom Kind zum Erwachsenen gründlicher und schneller und die Kontrolle über eigenes Denken und Lernen nimmt zu, wobei erste Anzeichen einer Planung und Überwachung von Handlungen in komplexeren Problemsituationen bereits ab einem Alter von fünf Jahren erkennbar sind (Sternberg, 2003; Deloache, Miller & Pierroutsakos, 1998). Übertragen auf die im Test zu messenden Fertigkeiten sind diese kognitiven Grundlagen z.B. für das Erler-

nen des Abgleichs experimenteller Ergebnisse mit zuvor aufgestellten Hypothesen von Bedeutung.

Weiterhin stellt die Fähigkeit zur mentalen Repräsentation von Sachverhalten im Rahmen der allgemeinen kognitiven Grundlagen eine wichtige Grundlage dar. Die Zielgruppe des Tests muss in der Lage sein, mental mit abstrakten Konzepten zu arbeiten, sie mental zu manipulieren und auf diese Weise Lösungen zu finden. Gemäß Piaget (1974) sind Kinder dazu ab der Stufe des formalen Denkens, spätestens ab einem Alter von ungefähr zwölf Jahren, in der Lage. Studien zeigen, dass Kinder diese Fähigkeiten sogar schon vor der von Piaget angenommenen Altersstufe besitzen (Deloache et al., 1998). Denkoperationen können ab diesem Alter mit abstrakten, nicht mehr konkret vorstellbaren Inhalten durchgeführt werden. Diese Denkoperationen sind die Voraussetzung für eine weitere wichtige grundlegende Fähigkeit: das Problemlösen. Das *Problemlösen* ist deshalb von Bedeutung, da es dem Prozess der wissenschaftlichen Erkenntnisgewinnung ähnelt (Klahr, 2000). Definiert man ein Problem als bestehend aus einem Anfangszustand, einem Zielzustand und einer bestimmten Zahl zulässiger Operatoren, die den Anfangszustand zum Zielzustand bringen und die bestimmten Beschränkungen unterliegen (Newell & Simon, 1972), so können einzelne Bereiche wissenschaftlicher Erkenntnisgewinnung durchaus als eine Art des Problemlösens betrachtet werden. Das *Scientific-Discovery-as-Dual-Search (SDDS)-Modell* (Klahr & Dunbar, 1988; Klahr, 2000) bezieht das klassische Problemlösen auf wissenschaftliche Erkenntnisgewinnung, indem es als Suche in zwei Problemräumen, dem Hypothesenraum und dem Experimentraum, beschrieben wird (vgl. Abschnitt 2.2). In diesem Sinne stellt die Problemlösefähigkeit eine wichtige allgemeine Grundlage für das Erlernen experimenteller Fertigkeiten dar.

Weitere wichtige kognitive Grundlagen für den Prozess der wissenschaftlichen Erkenntnisgewinnung sind die Fähigkeiten, Schlussfolgerungen zu ziehen und kausal zu argumentieren, die insbesondere bei der Entscheidung über das Verwerfen oder Beibehalten einer Hypothese eine Rolle spielen. Dabei geht es unter anderem darum, aus einer vorliegenden Datenbasis Schlüsse ziehen zu können. Hier sind insbesondere deduktives Denken und epistemologisches Verständnis von Bedeutung. *Deduktives Denken* wird ebenso bei einem wissenschaftlichen Vorgehen benötigt, bei dem Hypothesen aus einer vorhandenen Wissensbasis abgeleitet werden, um sie dann in speziell darauf abgestimmten Experimenten anhand erhobener Daten zu prüfen. Es wird von den allgemein vorhandenen Informationen eines Sachgebiets auf eine spezielle Hypothese geschlossen, die bisherige wissenschaftliche Erkenntnisse widerspiegelt. Bei dem *epistemologischen Verständnis* geht es um das Wissen, wie wissenschaftliche

Erkenntnisse erworben wurden und wann sie sich ändern (Zimmermann, 2007). Im Hinblick auf den Prozess wissenschaftlicher Erkenntnisgewinnung bezieht es sich unter anderem auf das Verständnis, dass die Ergebnisse eines Experiments das in der Wissenschaftsgemeinschaft vorhandene Wissen ändern können. Auch das *induktive Denken* spielt in diesem Zusammenhang eine wesentliche Rolle. Hier geht es darum, vom Besonderen auf das Allgemeine zu schließen (Sternberg, 2003). Es wirkt sich insofern auf den Bereich wissenschaftlicher Erkenntnisgewinnung aus, als eine erfolgreiche wissenschaftliche Strategie, z.B. die zu untersuchende Variable zu variieren und die übrigen Variablen bzw. Bedingungen konstant zu halten oder zu kontrollieren, auch in anderen wissenschaftlichen Untersuchungen anzuwenden ist. Weiterhin schließt man induktiv, z.B. aus dem anhand eines Experiments bestätigten Zusammenhang zweier Variablen, dass diese auch in Situationen außerhalb des Experiments zusammenhängen, dass also eine gewisse Gesetzmäßigkeit besteht.

Neben den genannten Grundlagen ist es für alle drei Fertigkeitsbereiche des Tests von Bedeutung, bestimmte *Heuristiken* zur Verfügung zu haben. Dies sind Entscheidungsstrategien, Denkmodelle, Analogien oder einfache Regeln, unter deren Zuhilfenahme komplexe Probleme, die sich nicht vollständig lösen lassen, vereinfacht werden können. Heuristiken reduzieren die Komplexität eines Problems, schaffen damit eine Art kognitive Entlastung (Sternberg, 2003) und tragen dadurch zu der im vorletzten Abschnitt angesprochenen Beschleunigung der Informationsverarbeitung bei (Kail & Bisanz, 1992). Erwachsenen gelingt der Einsatz solcher Heuristiken in Forschungszusammenhängen aufgrund ihrer Erfahrungen besser als Kindern und Jugendlichen. Eine dieser Heuristiken besteht bezüglich wissenschaftlicher Erkenntnisgewinnung darin, sowohl bei der Suche nach Hypothesen als auch bei der Einordnung von Ergebnissen, die Plausibilität zu beurteilen und auf dieser Grundlage Entscheidungen zu treffen. Diese Heuristik kann als domänenübergreifend bezeichnet werden (Klahr, Fay & Dunbar, 1993). In die gleiche domänenübergreifende Kategorie fällt die Heuristik zur Kontrolle von Variablen, um ein unkonfundiertes Experiment durchzuführen.

Sicherlich stellen die hier genannten kognitiven Grundlagen keine erschöpfende Liste dar. Es bleibt allgemein zusammenzufassen, dass die Wahrscheinlichkeit der Messbarkeit prozessbezogener Fertigkeiten aufgrund kognitiver Entwicklungen und zunehmender Lerngelegenheiten mit wachsendem Alter der Testpersonen steigt. Die folgenden Unterabschnitte setzen die genannten Befunde in Beziehung zu den anhand des Tests zu messenden Fertigkeiten.

2.3.1 Identifizieren wissenschaftlicher Hypothesen (Fertigkeitsbereich H)

Durch die Beantwortung von Testaufgaben des Fertigkeitsbereiches *H* sollen die Testpersonen zeigen, inwiefern sie verstehen, dass Wissenschaftlerinnen und Wissenschaftler ihre Hypothesen aufbauend auf das in der Wissenschaftsgemeinschaft vorhandene Wissen formulieren. Dies bedeutet, dass Testpersonen in der Lage sein müssen, Informationen über einen bestimmten wissenschaftlichen Bereich aufzunehmen und mental zu repräsentieren. Sie müssen diese Informationen mit einem präsentierten Versuch in Beziehung setzen und entscheiden können, welcher Vermutung die Wissenschaftlerinnen und Wissenschaftler in einem Aufgabenbeispiel nachgehen. Dies erfordert allgemein eine gewisse Kapazität des Arbeitsgedächtnisses zur parallelen Verarbeitung von Informationen und von Überlegungen zur Lösung. Weiterhin ist deduktives Denken notwendig, um eine wissenschaftliche Hypothese aus der vorhandenen Wissensbasis abzuleiten.

Studien zeigen, dass es für junge Testpersonen nicht einfach ist, sich auf die wesentlichen Sachverhalte zu konzentrieren, um die passende Hypothese zu finden. Dies betrifft insbesondere Schülerinnen und Schüler bis zur dritten Klasse (Klahr et al., 1993). Ältere Schülerinnen und Schüler und Erwachsene haben damit weniger Probleme. Weitere Schwierigkeiten bestehen für Schülerinnen und Schüler bis einschließlich der fünften Klasse darin, eine Hypothese über eine kausale Ursache-Wirkungs-Beziehung aufzustellen. Wenn der Kontext allerdings begrenzt ist und eine möglichst perfekte Kovariation bzw. einen kausalen Zusammenhang zwischen einer potentiellen Ursache und einem Effekt vorgegeben ist, können Sechstklässlerinnen und Sechstklässler eine Hypothese darüber aufstellen, dass ein bestimmter Faktor kausal verantwortlich ist (Zimmermann, 2007). Es zeigt sich, dass diese jüngeren Schülerinnen und Schüler immer dann Schwierigkeiten haben, wenn die zu verarbeitenden Informationen zu komplex sind. Diesen Sachverhalt stützt der Befund, dass Erwachsene besser als Schülerinnen und Schüler der dritten und der sechsten Klasse in der Lage sind, im Rahmen eines experimentellen Settings mehr als nur eine Hypothese zu berücksichtigen (Klahr et al., 1993). Darüber hinaus gelingt es Schülerinnen und Schülern um so eher, passende Hypothesen zu finden, je mehr Vorwissen sie über einen bestimmten Themenbereich besitzen oder im Rahmen einer Aufgabe vermittelt bekommen (Zimmermann, 2007).

Insgesamt zeigt sich, dass Schülerinnen und Schüler bis zur sechsten Klasse, also bis etwa zum Alter von zwölf Jahren Schwierigkeiten mit dem Aufstellen von Hypothesen haben. Ab dieser Altersstufe steigt das Vermögen zur Verarbeitung aller

notwendigen Informationen und führt dazu, dass es ihnen immer leichter fällt, Hypothesen zu identifizieren. Vor dieser Altersstufe sollte der *Test zur Erfassung prozessbezogener naturwissenschaftlicher Grundbildung* also nicht eingesetzt werden.

2.3.2 PLANEN EINER WISSENSCHAFTLICHEN UNTERSUCHUNG (FERTIGKEITSBEREICH P)

Durch die Beantwortung von Testaufgaben des Fertigkeitsbereichs *P* sollen die Testpersonen zeigen, inwiefern sie in der Lage sind, experimentelle Bedingungen so zu planen, dass die unkonfundierte Prüfung einer vorgegebenen Hypothese ermöglicht wird. Dazu müssen die Testpersonen zum einen wissen, dass experimentelle Bedingungen aus der Hypothese folgen, die geprüft werden soll. Zum anderen müssen sie wissen, dass es in der Gestaltung experimenteller Bedingungen wichtig ist, die zu untersuchende Variable zu variieren und die übrigen Bedingungen bzw. Variablen konstant zu halten oder zu kontrollieren.

Beim Planen von Experimenten ist es notwendig, sich auf die wichtigsten Sachverhalte und auf die zu untersuchenden Variablen zu konzentrieren und systematisch vorzugehen. Dies gelingt Erwachsenen deutlich besser als Kindern, die noch sehr unsystematisch vorgehen und dadurch meist konfundierte Experimente planen und durchführen. Sie stellen kaum Bemühungen an, verschiedene Variablen systematisch zu vergleichen, wenn sie die Gründe für ein bestimmtes Ergebnis erforschen (Ruffman, Perner, Olson & Doherty, 1993; Klahr et al., 1993; Deloache et al., 1998). Je älter die Schülerinnen und Schüler sind, desto besser schneiden sie in Performanztests ab, in denen sie Experimente entwickeln sollen (Chen & Klahr, 1999). Es geht dabei insbesondere um die Isolation, Kontrolle und systematische Kombination von Variablen. Die Kontrolle von Variablen wird als domänenübergreifende und strategische Fertigkeit betrachtet, da sie die Suche nach möglichen Experimenten begrenzt (Zimmermann, 2007). Chen (1999) bezeichnete diese Strategie als *Control of Variables Strategy (CVS)*. Studien zeigen, dass der Einsatz dieser Strategie um so besser funktioniert, je älter die Schülerinnen und Schüler sind. Grundschülerinnen und Grundschüler haben nur ein fragiles Verständnis der Konzepte und Fertigkeiten der *CVS* (Chen & Klahr, 1999; Klahr et al., 1993). Ab der fünften und sechsten Klasse beginnen sie, ein rudimentäres Verständnis von *CVS* zu entwickeln bzw. können diese Strategie erlernen und einen Transfer auf andere Situationen und Probleme herstellen. Auch Zimmermann (2007) zeigt in ihren Studien, dass Schülerinnen und Schüler etwa ab der fünften Klasse in der Lage sind, angemessene kontrollierte Tests zu produzieren. Ab der sechsten Klasse ist ihre Leistung mit einer Gruppe von Erwachsenen zu

vergleichen. Die Fertigkeit, kontrollierte Tests auszuwählen oder wiederzuerkennen zeigt einen linearen Entwicklungstrend. Selbst wenn die Testpersonen nicht spontan in der Lage waren, einen kontrollierten Test zu planen, so erkannten sie ihn wenigstens, wenn sie wussten, welches Ergebnis produziert werden sollte. Je begrenzter die eigenen Freiheitsgrade in der Planung von Experimenten waren, desto besser schnitten Kinder in Testsituationen ab (Deloache et al., 1998).

Es zeigt sich also auch im Fertigkeitsbereich *P*, dass Schülerinnen und Schüler auf Grundlage der Forschungsergebnisse ab der sechsten Klasse, also ab einem Alter von zwölf Jahren, in der Lage sein sollten, Aufgaben zu lösen, welche die Fertigkeit zur Planung unkonfundierter Experimente messen.

2.3.3 NUTZEN WISSENSCHAFTLICHER ERGEBNISSE (FERTIGKEITSBEREICH N)

Die Beantwortung der Testaufgaben des Fertigkeitsbereichs *N* soll Auskunft darüber geben, inwiefern die Testpersonen in der Lage sind, das Ergebnis eines vorgegebenen Experiments richtig zu interpretieren und darüber zu entscheiden, ob die ebenfalls vorgegebene Hypothese zu verwerfen oder beizubehalten ist. Die Testperson muss also wissen, dass das experimentelle Ergebnis auf die zuvor aufgestellte Hypothese bezogen werden muss, um eine Entscheidung über das Beibehalten oder Verwerfen der Hypothese treffen zu können.

Auch in diesem Bereich zeigen die Forschungsergebnisse, dass sich die Fertigkeiten der Kinder und Jugendlichen mit zunehmendem Alter deutlich verbessern. So zeigen Studien mit Kindergartenkindern, dass jüngere Testpersonen dazu neigen, Ergebnisse zu vernachlässigen, die eine Hypothese nicht stützen. Erwachsene sind dagegen in der Lage, auf die wichtigsten Sachverhalte zu fokussieren (Klahr et al., 1993). Frühe Unternehmungen, das Verständnis von Kindern bezüglich der Beziehung von Hypothesen und Evidenzen zu untersuchen, zeigen, dass Kinder vor der Altersstufe von elf bis zwölf Jahren wenig Einsicht haben, wie Hypothesen durch Evidenzen gestützt oder widerlegt werden. Diese Einsicht bleibt oft sogar bis ins Erwachsenenalter vage (Kuhn et al., 1988). Schülerinnen und Schüler neigen bis zur sechsten Klasse dazu, Evidenzen zu ignorieren und zu behaupten, sie seien konsistent mit ihrer Theorie. Oder sie benutzen die Evidenz, um eine neue Theorie zu konstruieren, ohne zu bemerken, dass diese neue Theorie im Widerspruch zu ihrer alten Theorie steht. Die Beurteilung von Evidenzen hängt vom Vorwissen der Kinder ab und davon, für wie plausibel sie bestimmte Ergebnisse halten. Anders als Erwachsene oder auch College-Studenten haben sie Probleme, Evidenzen zu evaluieren, die im Widerspruch zu ihrem Vorwissen stehen (Zimmermann, 2007; Amsel & Brock, 1996). Allerdings sind

Kinder mit ausreichend Übung in der Lage, Evidenzen unabhängig von ihrem Vorwissen zu beurteilen (Schauble, 1990), was dafür spricht, dass diese Fertigkeit grundsätzlich erlernbar ist.

Abschließend ist festzuhalten, dass das Verständnis des Zusammenhangs zwischen Theorie und Evidenz auf epistemologische Überzeugungen zurückgeht, also darauf, welche Vorstellungen darüber bestehen, was Wissen ist und wie es sich ändert (Chinn & Malhotra, 2002b). Die Koordination von Theorie und Evidenz beinhaltet im Kern, dass das eigene Vorwissen möglichen Änderungen unterliegt. So besteht eine Schlüsselfertigkeit, die über den Fortgeschrittenenlevel in diesem Bereich entscheidet, darin, eine gewisse metakognitive Kontrolle über diesen Änderungsprozess zu besitzen. Es muss ein Bewusstsein darüber geben, dass Theorie und Evidenz einen jeweils eigenen epistemologischen Status besitzen und in einer besonderen Beziehung zueinander stehen (Zimmermann, 2007). Forschungsergebnisse deuten darauf hin, dass Menschen eine grundsätzliche Tendenz aufweisen, aufgestellte Hypothesen eher zu verifizieren als zu falsifizieren. Dies zeigt sich insbesondere bei Schülerinnen und Schülern bis einschließlich der sechsten Klasse. Welche Strategie vornehmlich gewählt wird - eher das Verifizieren oder eher das Falsifizieren - ist zusätzlich abhängig davon, für wie plausibel die Hypothese gehalten wird. Wird sie für plausibel gehalten, ist die Tendenz stärker, sie zu verifizieren, als für den Fall, dass die Hypothese als wenig plausibel betrachtet wird. In dem Fall werden durchaus auch nicht bestätigende Evidenzen in Betracht gezogen (Klahr et al., 1993).

Es wird angenommen, dass Schülerinnen und Schüler ab einem Alter von zwölf Jahren beginnen, Experimente zu nutzen, um eine Hypothese zu untersuchen. Vor dieser Altersstufe geht es ihnen eher darum, mit einem Experiment eine bestimmte Überzeugung zu zeigen (Penner & Klahr, 1996a). Dies deutet darauf hin, dass die Bearbeitung von Testaufgaben des Bereichs *N* auch erst dann sinnvoll sein wird, wenn die Testpersonen mindestens ein Alter von zwölf Jahren erreicht haben.

Zusammenfassend kann aufgrund der genannten Befunde davon ausgegangen werden, dass Schülerinnen und Schüler in etwa ab einem Alter von zwölf Jahren in gewisser Ausprägung über Teile der Fertigkeiten verfügen können. Allgemein zeigt sich hinsichtlich aller drei Bereiche die Tendenz, dass die Fertigkeiten um so eher entwickelt und damit theoretisch messbar sind, je älter die Testpersonen sind. Diese Feststellung hat Auswirkungen auf die Auswahl der Zielgruppe und damit auch auf die Auswahl der Stichprobe für die Testentwicklung, auf die in Abschnitt 2.7.1 noch genauer eingegangen wird.

2.4 Versuche und Grenzen der Erfassung prozessbezogener naturwissenschaftlicher Grundbildung

Neben inhaltlichen Überlegungen hinsichtlich zentraler Elemente experimenteller Erkenntnisgewinnung und bezüglich einer Definition der Kompetenz prozessbezogener naturwissenschaftlicher Grundbildung sind auch technische Aspekte der Testentwicklung von zentraler Bedeutung. Hier gilt es zu entscheiden, welche Art von Verfahren am besten zur Erfassung dieser speziellen Kompetenz geeignet ist. Eine solche Entscheidung ist zum einen vor dem Hintergrund der Frage zu treffen, auf welche Weise möglichst genaue und valide Informationen zu gewinnen sind. Zum anderen fließen Überlegungen zur Testökonomie und zum speziellen Anwendungskontext des Testverfahrens in den Entscheidungsprozess ein.

Die Bemühungen in den Bereichen *Public Understanding of Science (PUS)*, *Public Understanding of Research (PUR)*, *Nature of Science (NOS)* haben eine große Anzahl von offenen und geschlossenen Fragebögen, Interviews, Multiple-Choice-Tests und anderen Verfahren hervorgebracht. Bei diesen Verfahren ging es vorrangig darum, das Wissen über die *NOS* oder über spezielle naturwissenschaftliche Inhalte zu testen oder die Einstellungen gegenüber oder das Interesse an Naturwissenschaften zu erfragen. Die Verfahren, die speziell Wissen oder Fertigkeiten in Bezug auf den Prozess naturwissenschaftlicher Erkenntnisgewinnung abbilden, sind im Vergleich zu diesen Verfahren in der Minderzahl. Die folgenden Abschnitte nehmen eine Rückschau bisheriger Versuche und Grenzen der Erfassung prozessbezogener Bereiche naturwissenschaftlicher Grundbildung und des Einsatzes unterschiedlicher Methoden zu ihrer Erfassung vor. Die Darstellung kann sicherlich nicht als vollständig bezeichnet werden, vielmehr ist sie als eine Auflistung der derzeit zugänglichen und für diesen Bereich typischen Erhebungsverfahren zu verstehen.

Bei den dargestellten Verfahren handelt es sich um Leistungstests. Diese Klassifizierung erscheint in Bezug auf Multiple-Choice-Tests augenscheinlicher als mit Blick auf Interviews. Dennoch wird in allen Fällen die Kompetenz in Form von Leistung oder Wissen erfasst. Leistungstests zeichnen sich gemäß Rost (2004a, S.43) dadurch aus, dass *„von den Personen die Lösung von Aufgaben oder Problemen verlangt wird, die Reproduktion von Wissen, das Unterbeweisstellen von Können [...]“*. Anhand welches Verfahrens dies geschieht, ist gemäß dieser Definition zunächst irrelevant. Aus diesem Grund beschäftigen sich die kommenden Abschnitte mit Leistungsabfragen anhand von Fragebögen, Interviews, Performanztests, Multiple-Choice-Tests sowie anhand

alternativer Verfahren wie Portfolio- und auch Critical-Incident-Techniken.

2.4.1 FRAGEBÖGEN

Wie bereits in der Einleitung dieses Abschnitts erwähnt, erfasst der größte Teil von Fragebögen im Bereich naturwissenschaftlicher Grundbildung die *Natur der Naturwissenschaften* und die Einstellungen gegenüber Naturwissenschaften. Da sich die Testentwicklung im Rahmen dieser Arbeit auf den prozessbezogenen Teil naturwissenschaftlicher Grundbildung bezieht, wird sich dieser Abschnitt auf die Maße konzentrieren, die sich inhaltlich mit diesem Bereich befassen. Ein Überblick über Verfahren zur Messung des Verständnisses der *Natur der Naturwissenschaften* findet sich beispielsweise bei Lederman et al. (1998).

Bevor auf konkrete Fragebogen-Verfahren eingegangen wird, werden an dieser Stelle einige allgemeine Merkmale von Fragebögen betrachtet. Fragebögen können zunächst aufgrund der verbalen Reize differenziert werden, auf welche die Testpersonen reagieren müssen. Sie können in einfachen Fragen, in Statements oder einzigen Substantiven oder Adjektiven bestehen.

Desweiteren besteht die Möglichkeit, Fragebögen anhand ihres Antwortformats zu differenzieren. Hier sind *offene* und *geschlossene Formate* zu unterscheiden. *Geschlossene Antwortformate* können die Form einfacher Ja-Nein-Antworten annehmen oder in Form von Aussagen bestehen, bei denen sich die Testperson für eine Antwort entscheiden muss. Beide Formate können als *Forced Choice* bezeichnet werden, eine Spezialform des weiter unten beschriebenen Multiple-Choice-Formats. Geschlossene Antwortformate können in Form von *Ratingskalen* vorliegen. In dem Fall werden Statements anhand einer Abstufung eingeschätzt, die beispielsweise in der Bandbreite *stimmt*, *stimmt eher*, *stimmt eher nicht*, *stimmt nicht* vorliegen kann. Es hat sich gezeigt, dass Fragebögen im Rating-Format zwar als ökonomisch und leicht auswertbar gelten, sich aber oft auch als wenig aussagekräftig erweisen können. Die mit dieser Art von Fragebögen erhobenen Daten im Bereich naturwissenschaftlicher Grundbildung lassen nicht immer ein klares Urteil darüber zu, ob die Testpersonen tatsächlich ein Verständnis von Wissenschaft besitzen oder ob ihre Antworten auf auswendig gelernten Inhalten basieren (Carey, Evans, Honda, Jay & Unger, 1989). Ein weiteres Problem ist häufig die Validität dieser Fragebögen. Ein Grund dafür liegt in der mangelnden Qualität der Entwicklung, ein anderer in der Interpretation der Testfragen. Oft stimmen die Interpretationen der Testpersonen nicht mit dem überein, was die Testkonstrukteure ursprünglich beabsichtigt haben. Diese Diskrepanz führt zu einer geringen Validität: Der Fragebogen misst nicht das, was die Entwickler eigentlich zu

messen beanspruchten (Lederman et al., 1998).

Beim Einsatz eines *offenen Antwortformates* dagegen wird der Nachteil möglicher Missverständnisse in Ratingformaten dadurch vermieden, dass die Personen auf eine Frage frei und ohne Einschränkung auf eine Frage antworten können. Dieses Antwortformat ist insofern vorteilhaft als Schülerinnen und Schüler durch Klassenarbeiten daran gewöhnt sind. Dabei kann die Antwort aus einem Wort, einem Symbol oder aber einer längeren Gedankenführung und Begründung des Lösungsweges bestehen. Sogar Zeichnungen, Diagramme und Skizzen sind hier als Antworten denkbar (Duit et al., 2001). In jedem Falle werden der Testperson alle Möglichkeiten gegeben, die maximale Leistung zu zeigen, also alles vorhandene Wissen wiederzugeben. Problematisch bei dieser Art der Datenerhebung sind die sehr aufwändige Durchführung, Auswertung und Interpretation. Hier besteht die Gefahr, dass Aussagen falsch ausgewertet und interpretiert werden, dass also Auswertungs- und Interpretationsobjektivität eingeschränkt sind. Um akzeptable Werte zu gewährleisten, muss zur Auswertung und Interpretation ein gutes Kategorien- oder Kodierungssystem zur Verfügung stehen, welches im Vorfeld der Auswertung zunächst aufwändig zu entwickeln ist. Erst wenn unterschiedliche Personen (*Rater*) ihre Bewertungen anhand des Kategoriensystems abgegeben haben und die Beurteilerübereinstimmung ausreichend hoch ist, kann mit der Interpretation solcher Daten begonnen werden.

Die Geschichte der Erhebung naturwissenschaftlicher Grundbildung oder von Teilen naturwissenschaftlicher Grundbildung anhand von Fragebögen hat in den 60er Jahren begonnen. Anhand von Tabelle 2.3 wird auf die Fragebögen zurückgeblickt, die sich (zum Teil auch nur ansatzweise) mit dem Prozess naturwissenschaftlicher Erkenntnisgewinnung bzw. mit prozessbezogenen Fertigkeiten beschäftigen oder beschäftigt haben. Die dargestellten Fragebögen zeigen die bereits beschriebene Bandbreite von offenen bis geschlossenen Antwortformaten. Das *Wisconsin Inventory of Science Processes (WISP)* des *Scientific Literacy Research Center* (1967) stellt ein Beispiel für ein Ratingverfahren dar, bei dem sich die Testpersonen zwischen den Stufen *zutreffend*, *nicht zutreffend* oder *nicht verstanden* entscheiden müssen. Als Beispiel für ein offenes Format, das zur Auswertung ein Kategoriensystem erfordert, ist hier das *AASPS* genannt (Germann, Aram & Burke, 1998). Diese beiden Verfahren bilden bezüglich der Offenheit des Antwortformats in etwa die beiden Extreme, die Testverfahren in diesem Bereich annehmen können.

Forschungsergebnisse weisen auf die Nachteile der Fragebogenmethode im Bereich der Erfassung *prozessbezogener naturwissenschaftlicher Grundbildung* hin. Fragebögen mit geschlossenem Antwortformat sind aufgrund von Validitätsproblemen als

Tabelle 2.3: Fragebögen, die den Prozess naturwissenschaftlicher Erkenntnisgewinnung erfassen

Jahr	Instrument	Autor(en)	Ziel der Messung	Zielgruppe
1962	Processes of Science Test (POST)	Biological Sciences Curriculum Study (BSCS)	Verständnis von Prozessen in den Naturwissenschaften	
1966	Science Process Inventory (SPI) Anmerkung: ähnlicher Inhalt wie WISP und TOUS-Subskala III 135 Items im Forced-Choice-Format (stimme zu/ stimme nicht zu); keine Skalen, nur ein Gesamtscore Reliabiltät: .86	Welch, W.	Verständnis der Methoden und Prozesse, durch die naturwissenschaftliches Wissen entsteht	Schüler/innen
1967	Wisconsin Inventory of Science Processes (WISP) 93 Statements werden bewertet als *zutreffend*, *nicht zutreffend* oder *nicht verstanden*. Reliabiltät (K.-R.): .76	Scientific Literacy Research Center	Verständnis von Annahmen, Aktivitäten, Zielen und Produkten der Naturwissenschaften, bezogen auf naturwissenschaftliche Prozesse	Schüler/innen
1974	Science Inventory	Hungerford, H. & Walding, H.		
1991	Alternative Assessment of Science Process Skills (AASPS) basiert auf dem „Directed Inquiry Approach to Learning Science Process Skills and Scientific Problem Solving“ (Germann, 1987; 1989); Aufgaben *„In hot water“* Auswertung der Schülerantworten anhand des eigens entwickelten „Science Process Skills Inventory“ (SPSI) 1 Kontext: *„DJ and BJ were getting ready to wash the car“* 10 offene Fragen	Missouri Department of Elementary and Secondary Education	Verständnis des Prozesses naturwissenschaftlicher Erkenntnisgewinnung; prozessbezogene Fertigkeiten	Schüler/innen der 7. Klasse

problematisch einzuschätzen (Lederman et al., 1998). Fragebögen mit offenem Format (wie das hier aufgeführte *AASPS*) sind hinsichtlich ihrer Validität als günstiger einzustufen, erweisen sich aber in Anwendung und Auswertung als sehr unökonomisch. So haften der Fragebogenmethode Nachteile an, die sich nicht ohne Inkaufnahme eines anderen Nachteils überwinden lassen. Dies macht deutlich, dass auf der Suche nach einem geeigneten *Verfahren zur Erfassung prozessbezogener naturwis-*

senschaftlicher Grundbildung ein Kompromiss gefunden werden musste, der sowohl Validitätsansprüche als auch Ökonomieansprüche erfüllt.

2.4.2 INTERVIEWS

Eine andere Möglichkeit der Datenerhebung besteht in der Durchführung von Interviews. Im Gegensatz zu Fragebögen sind sie aussagekräftiger und haben den Vorteil, dass die interviewende Person bei unklaren Antworten der Testpersonen die Möglichkeit hat, vertiefend nachzufragen. Obwohl dies im Hinblick auf die Validität ein Vorteil ist, folgt aus dieser Möglichkeit der Nachteil, dass bei nicht vollkommener Standardisierung und Strukturierung des Interviews die Vergleichbarkeit und somit die Aussagekraft der Daten vermindert wird. Daher ist es bei der Konstruktion eines Interviewleitfadens außerordentlich wichtig, mögliche Antworten der Testpersonen zu antizipieren und schon im Vorfeld entsprechende standardisierte Nachfragen zu formulieren. Sowohl die Konstruktion als auch die Auswertung von Interviews erfordern einen erheblichen Aufwand. Zum einen müssen sie zur weiteren Verwendung transkribiert werden und zum anderen müssen im Vorwege - ebenso wie im bereits beschriebenen Fall von Fragebögen mit offenem Antwortformat - genaue Kodierungen und Auswertungskategorien entwickelt werden, um eine reliable Messung zu ermöglichen.

Um Interviews und eventuelle Anschlussfragen gezielt entwickeln zu können, besteht die Möglichkeit, sie mit einer Fragebogenerhebung zu kombinieren (Aikenhead, 1988; Lederman et al., 1998; Lederman, 2002). Die Antworten des Fragebogens werden zur Konstruktion des Interviews genutzt, um mögliche vertiefende Fragen zu identifizieren, die allen Testpersonen gleichermaßen gestellt werden können. Die Kombination der beiden Messmethoden führt zwar zur Kompensation einiger ihrer Schwächen, sie stellt jedoch eine wenig ökonomische und sehr zeitaufwändige Variante der Datenerhebung dar.

Ein viel zitiertes Beispiel für ein Interview, das der Erfassung eines Teilbereiches naturwissenschaftlicher Grundbildung dient, stellt das klinische Interview zur Erfassung des Verständnisses von der Natur wissenschaftlichen Wissens und wissenschaftlicher Forschung dar, das Susan Carey (1989) mit Zwölfjährigen durchgeführt hat und das Sodian, Thoermer und Kircher (2002) später ins Deutsche übersetzten und adaptierten. Zur Auswertung benutzte Carey ein Kodierungsschema, das die Aussagen der Schülerinnen und Schüler in drei Verständnislevel einteilt. Das Interview umfasst Fragen dazu, wie Wissenschaftlerinnen und Wissenschaftler mit ihren Ideen (*„Is there a relationship between a scientist´s ideas and the rest of the work a scientist*

does?"), Hypothesen (*„Where does a scientist get a hypothesis?"*), Experimenten (*„How does a scientist decide what experiment to do?"*) und Ergebnissen (*„What happens when a scientist is testing his/her ideas, and gets a different result from the one he/she expected?"*) umgehen.

Es ist zu erkennen, dass das Interview den Prozess wissenschaftlicher Erkenntnisgewinnung, wie er bereits in Abschnitt 2.2.2 beschrieben wurde, erfasst. Neben den standardisierten Fragen gibt Carey zusätzlich Wörter vor, denen mit der Frage „Was meinst du mit..." begegnet werden soll, wenn die interviewte Person sie benutzt. Beispiele für solche Wörter sind *Theorie („theory")*, *Erklärung („explanation")* und *Beweis („proof")*. Auf diese Weise wird dem Interview zusätzliche Struktur gegeben und gewährleistet, dass Nachfragen immer an gleicher Stelle und in gleicher Weise erfolgen.

Dieses Interview stellt ein gutes Beispiel für die Erhebung eines Teilbereiches naturwissenschaftlicher Grundbildung dar. Dennoch bleiben die typischen Nachteile, die mit der Durchführung von Interviews verbunden sind, bestehen. Trotz aller Standardisierung und Strukturierung kann nicht vollständig sichergestellt werden, dass Interviews in absolut gleicher Weise erfolgen. Diese Gefahr besteht umso mehr, wenn sie nicht von der gleichen interviewenden Person durchgeführt werden. Dies würde eine Einschränkung der Durchführungsobjektivität bedeuten. Doch der größte Nachteil dieses Instruments liegt im benötigten Zeitaufwand. In einem bestimmten zur Verfügung stehenden Zeitraum wird eine deutlich geringere Anzahl an Testpersonen befragt als es beispielsweise mit Multiple-Choice-Tests möglich ist. Hier stehen sich also Validität und Ökonomie gegenüber. Gelingt es jedoch, einen validen Multiple-Choice-Test zu entwickeln, so kann dieser dem Interview durchaus vorgezogen oder durch Interviews mit einem Teil der Stichprobe ergänzt werden.

2.4.3 PERFORMANZTESTS

Da es im Rahmen der Messung naturwissenschaftlicher Grundbildung, wie bereits beschrieben, nicht allein darum geht, was eine Person weiß, sondern auch darum, ob und inwieweit sie ihr Wissen anwenden kann, sind in diesem Bereich Testverfahren gefordert, die diese speziellen Fertigkeiten situationsspezifisch erfassen können. Diese Anforderungen können durch *Performanztests* erfüllt werden, die Shavelson und Ruiz-Primo (1999) auch als *alternative Tests* (im Gegensatz zu konventionellen Tests) bezeichnen. Sie bestehen aus:

- einer Problemstellung, deren Lösung den Umgang mit konkreten (ersatzweise computersimulierten) Materialien erfordert,

- einem Antwortformat, in dem bestimmte Anforderungen (z.B. den Lösungsplan darzulegen, die Untersuchung zu protokollieren, die Ergebnisse in einem Graphen darzustellen) festgelegt sind und aus

- einem Auswertungsschlüssel, mit dem z.B. die Plausibilität und wissenschaftliche Haltbarkeit des Lösungsweges oder die Genauigkeit der erzielten Ergebnisse bewertet werden.

Es fällt auf, dass es hier nicht darum geht, die richtige Antwort zu finden, sondern vielmehr darum, eine vernünftige und logische Handlung durchzuführen, die ausgehend von einer Problemstellung und einer im Vorfeld zu formulierenden Planung z.B. auch die Kontrolle von Variablen enthält. Baxter et al. (1992) beschreiben drei Kräfte, welche die Suche nach alternativen Erhebungsverfahren angetrieben haben: die Unzufriedenheit mit bestehenden Multiple-Choice-Verfahren, die Fortschritte in der Forschung über Kognition und Instruktion sowie die Reform der Naturwissenschaftscurricula. Um valide und für die Bildungsforschung aussagekräftige Leistungsmaße zur Verfügung stellen zu können, müssen Tests über korrekte Antworten hinausgehen und auf Wissen über naturwissenschaftliche Konzepte und deren Anwendung fokussieren. Beispiele für Performanztests finden sich in der Arbeit von Ruiz-Primo et al. (1996; 2001), Shavelson (1999), Baxter (1992) und Tamir (1982).

Shavelson und Ruiz-Primo (1999) beschreiben die Schwierigkeit und den notwendigen Aufwand, um in Performanztests eine ausreichende Objektivität und Reliabilität zu erreichen. Dieses Ziel kann erreicht werden, indem die Beobachterinnen und Beobachter gut geschult werden und der Aufgabenpool ausreichend groß ist. Weiterhin muss ein Messverfahren vorhanden sein, welches das Verhalten der Testpersonen möglichst gut abbildet und das sich durch hohe Beobachterübereinstimmung (Interrater-Reliabilität) auszeichnet. Als Beispiel für die Entwicklung eines solchen Verfahrens wird an dieser Stelle die Arbeit von Baxter et al. (1992) präsentiert. Sie benutzten für ihre Entwicklung eines Auswertungssystems die bereits für valide befundene Aufgabe *„Paper Towels“* (Department of Education and Science, 1984). Diese Aufgabe ist so aufgebaut, dass den Schülerinnen und Schülern zunächst Equipment und Untersuchungsgegenstände vorgelegt werden. Anhand dieser Gegenstände sollen sie der Frage nachgehen, welches von drei Papiertüchern das meiste Wasser aufsaugt. Um diese Frage zu untersuchen, sollen die Schülerinnen und Schüler ihre eigene Methode entwickeln und das Equipment nutzen, das sie gemäß ihrer Methode benötigen. Eine wichtige Aufgabe der Schülerinnen und Schüler besteht weiter darin, genaue Notizen über ihr Vorgehen zu machen, die Ergebnisse ihrer Untersuchung

anzugeben und zu begründen, wie sie zu diesem Ergebnis gekommen sind. Ihre Ausführungen sollen so detailliert erfolgen, dass eine andere Person ihren Versuch exakt auf die gleiche Weise durchführen kann.

Aufgrund der in Feldtests durch die Testpersonen beschrittenen Wege zur Lösung der Aufgabe entwickelten Baxter et al. (1992) ein Auswertungssystem, das aus den drei Bereichen *Method for getting towels wet*, *Control of variables*, *Method for determining result* und einigen allgemeinen Kategorien besteht. In jedem dieser drei in Abbildung 2.6 dargestellten Bereiche treffen die Testpersonen Entscheidungen, die von den Beobachterinnen und Beobachtern genau festgehalten und benotet werden. Dieser Prozess der Entwicklung eines sinnvollen und reliablen Kategoriensystems sowie die Schulung der Beobachterinnen und Beobachter, damit diese mit dem Kategoriensystem sicher umgehen können, machen Performanztests zu einem aufwändigen Verfahren. Darüber hinaus kann die beobachtende Person jeweils nur eine Testperson begutachten und bewerten, was zu einer Einschränkung der Stichprobenanzahl oder einer Verlängerung des Testzeitraums führt.

Um die Nachteile dieses Verfahrens zu überwinden, gibt es unterschiedliche Ansätze. Baxter et al. (1992) wählten den Zugang, die Schülerinnen und Schüler ihr experimentelles Vorgehen in so genannten *Notebooks* auf präzise Weise protokollieren zu lassen, so dass eine andere Person das Experiment wiederholen könnte. Diese Protokolle sind in der Lage, die Performanz der Schülerinnen und Schüler zu reflektieren, erreichen eine gute Reliabilität und bringen darüber hinaus eine Zeit- und damit Kostenersparnis (es werden nur ca. 1,5 Minuten benötigt, um die Notizen auszuwerten).

Performanztests stellen bei überprüfter Beobachterübereinstimmung und einem exakten Auswertungsschlüssel ein sehr gutes Verfahren dar, um *prozessbezogene naturwissenschaftliche Grundbildung* zu messen. Die Testpersonen können ohne die Vorgabe von Antwortkategorien zeigen, ob sie in der Lage sind, einen Prozess naturwissenschaftlicher Erkenntnisgewinnung zu gestalten. Das Verfahren ist sehr valide und verfügt darüber hinaus über eine hohe *Augenscheinvalidität* [9]. Sie kann die Akzeptanz des Verfahrens bei den Testpersonen erleichtern. Trotz der im Beispiel beschriebenen Ökonomisierung anhand von *Notebooks* bleiben Performanztests aufwändig und für den Zweck eines schnellen und ökonomischen Screenings ungeeignet.

9 Die Augenscheinvalidität (engl. *face validity*) basiert auf der Eigenschaft eines psychologisch-diagnostischen Verfahrens, bei den Testpersonen eine subjektive Annahme über die Messintention desselben hervorzurufen, unabhängig von dessen tatsächlicher Validität. Die Augenscheinvalidität eines Verfahrens wird begünstigt, wenn es Ähnlichkeiten mit Real-Life-Situationen aufweist, die in Zusammenhang mit der diagnostischen Fragestellung stehen. Eine hohe Augenscheinvalidität wirkt sich positiv auf die Akzeptanz des Verfahrens aus (Kersting, 2003)

Student _____ Observer _____

Score _____ Skript _____

1. Method A. Container: pour water in/ put towel in; put towel in/ pour water in; 1 pitcher or 3 breakers/ — B. Drops — C. Tray (surface): towel in tray/ pour water on; pour water on tray/ wipe

2. Saturation A. Yes B. No C. Controlled

3. Determin Result
A. Weigh Towel
B. Squeeze towel/ measure water (weight or volume)
C. Measure water in/ out
D. Time to soak up water
E. No measurement
F. Count # drops until saturated
G. See how far drops spreadout
H. Other _____

4. Care in measuring A. Yes B. No

5. Correct result A. Yes B. No

Grade	Method	Saturate	Determine Result	Care in Measuring	Correct Answer
A	Yes	Yes	Yes	Yes	Yes
B	Yes	Yes	Yes	No	Yes/No
C	Yes	Yes/ Controlled	Error		Yes/No
D	Yes	No	Missing		Yes/No
F	---------- No Attempt----------				

Abbildung 2.6: Auswertungssystem für die Performanzaufgabe *Paper Towels*

2.4.4 WEITERE MÖGLICHE VERFAHREN

Neben den bereits lang etablierten Verfahren gibt es immer mehr alternative Zugänge, die sich der Erfassung naturwissenschaftlicher Grundbildung auf kreative Weise nähern können. Genannt seien an dieser Stelle die *Critical-Incidents (CI)-Technik* und *Portfolios*.

Die *Critical-Incident-Technik* [10] (Flanagan, 1954) stellt eine Methode dar, die nicht nur im Rahmen erziehungswissenschaftlicher Messungen eine Rolle spielt, sondern auch in der Arbeitspsychologie zur Eignungsdiagnostik eingesetzt wird. Die CI-Methode definiert im Rahmen der betrieblichen Eignungsdiagnostik kritische Arbeits- und Entscheidungssituationen, in denen das jeweilige Verhalten aussagekräftig für die Bewertung guter und schlechter Mitarbeiter ist. Um die Entscheidungen der Testpersonen bewerten zu können, müssen im Vorfeld Beispiele für erfolgreiche und für weniger erfolgreiche Verhaltensweisen festgelegt werden.

Zur speziellen Erfassung naturwissenschaftlicher Kompetenzen von Lehrerinnen und Lehrern haben Nott und Wellington (1998) ein CI-Verfahren entwickelt. Sie definierten einen CI dabei als eine Situation, in der eine Lehrkraft über eine bestimmte Vorgehensweise oder bestimmte Formulierungen entscheiden muss, die den Prozess wissenschaftlicher Bemühungen betrifft. Tabelle 2.4 zeigt ein Beispiel einer solchen CI-Aufgabe. Nott und Wellington (1998) konnten zeigen, dass die CI-Methode durchaus geeignet war, die naturwissenschaftlichen Kompetenzen der Lehrkräfte zu erfassen. Besser war sie jedoch geeignet, diese Kompetenzen zu schulen. Durch die CI-Methode wurden den Lehrkräften Probleme in ihren wissenschaftlichen Argumentationen bewusst und sie verbesserten ihre Kompetenzen, indem sie sich über ihre jeweiligen Entscheidungen mit anderen Lehrkräften austauschten.

Ein weiteres Verfahren, das auch hilfreich sein kann, um prozessbezogene naturwissenschaftliche Grundbildung zu erfassen, stellt die *Portfolio-Methode* dar. Der Begriff Portfolio stammt aus dem Lateinischen (lat. portare= *tragen* und folium= *Blatt*) und bezeichnet eine Sammlung von Objekten eines bestimmten Typs. Es kann im übertragenen Sinn auch eine Sammlung von hilfreichen Methoden, Verfahren oder Handlungsoptionen bedeuten. Der Begriff, der ursprünglich aus der Kunst stammt, wird heute in ganz unterschiedlichen Bedeutungsbereichen verwendet. Dazu gehören unter anderem die Bereiche Design, Wirtschaft und Bildung. Im Bildungsbereich, für den der Begriff in Analogie zu den genannten Einsatzgebieten übernommen wur-

10 Bei der von Flanagan (1954) entwickelten Technik handelt es sich um ein halbstandardisiertes Verfahren, das sich nicht - wie sonst üblich - mit typischen Ausschnitten einer Tätigkeit befasst, sondern ganz bewusst im positiven wie negativen Sinne untypische, kritische Ereignisse oder Verhaltensweisen aufgreift, die besonders zum Erfolg oder Misserfolg beitragen (Rosenstiel, 2003).

Tabelle 2.4: Beispiel einer *Critical-Incident*-Technik nach Nott und Wellington (1998)

Aufgabe: "The Magnesium Incident" A class of Y9´s are heating magnesium ribbon in a crucible with a lid. The purpose of the lesson is to test a consequence of oxygen theory that materials gain in weight when burnt. During the summary at the end of the lesson, four groups report a loss in weight, two groups report no difference and two groups report a gain in weight. List the kind of things you could say and do at this point.
Erfolgreiches Verhalten: The rationale for this CI is that teachers may talk about: - the need for everybody to make sure they have followed the same procedure - that experiments don´t always "work" (i.e. fit accepted theory) - that if experiments don´t work then they need to be critically evaluated - that results can be averaged and/ or discounted
Domäne: Domain: Scientists use procedures to check experimental results: experiments are as much the result of the experimenters´ skills as they are a mirror of nature

de, steht *Portfolio* für eine Mappe, in der alle Arten von Materialien (z.B. Texte, Zeichnungen, Fotos, Filme etc.) zusammengetragen und aufbewahrt werden. Portfolios können Ergebnisse von Lernprozessen ebenso gut abbilden wie die Prozesse selbst (Häcker, 2006). Im deutschsprachigen Raum spielen Portfolios seit Beginn der 1990er Jahre eine Rolle.

An dieser Stelle werden Portfolios als Mittel zur Darstellung von Prozessen aufgefasst. Nicht das Produkt steht hier im Vordergrund, sondern der Weg dorthin. Der gesamte Verlauf eines Prozesses wird sichtbar gemacht und dabei wird der Blick auf das gelenkt, was Personen können. Das Portfolio ist damit kompetenz- und nicht defizitorientiert. Paulson und Paulson (1991) definieren den Portfoliobegriff folgendermaßen:

> *„A portfolio is a purposeful collection of student work that exhibits the student´s efforts, progress, and achievement in one or more areas. The collection must include student participation in selecting contents, the criteria for selection, the criteria for judging merit, and evidence of student self-reflection.“* (Paulson & Paulson, 1991, S.60).

Die Bandbreite der Einsatzmöglichkeiten von Portfolios ist groß. Im Rahmen von wissenschaftlichen Erkenntnisprozessen kann das Portfolio als Projektportfolio eingesetzt werden, das aus der Bearbeitung eines komplexen Problems hervorgeht und den Lösungsprozess und die Ergebnisse der Arbeit dokumentiert (Häcker, 2006). Es kann demnach in einer Weise eingesetzt werden, die bereits im Bereich der Performanztests erwähnt wurde. In Anlehnung an die dort erwähnten *Notebooks* (Baxter et al., 1992) können alle Schritte eines Prozesses wissenschaftlicher Erkenntnisgewinnung in einer Form aufgezeichnet werden, die es anderen Personen ermöglicht, diesen Prozess nachzuvollziehen.

Die Bewertung dieser Methode fällt ähnlich aus wie die der Performanztests. Die inhaltliche Validität ist relativ hoch, da seine Inhalte unmittelbar dem Lern- und Arbeitsprozess entstammen, den es zu überprüfen gilt. Problematisch ist allerdings die Reliabilität dieses Verfahrens, die als moderat bis niedrig beschrieben wird. Hier ist die Beobachterübereinstimmung angesprochen, die oft nicht gegeben ist, weil eine vorgefasste Meinung über Schülerinnen und Schüler mit in die Bewertung des Portfolios einfließt (Duit et al., 2001). Weiterhin problematisch bei der Zielsetzung einer ökonomischen Testung ist die Tatsache, dass Portfolios ebenso wie Performanztests und die CI-Methode in der Auswertung sehr aufwändig und damit als unökonomisch zu bewerten sind.

2.4.5 MULTIPLE-CHOICE-TESTS

Einen weiteren Weg, sich der Erfassung prozessbezogener naturwissenschaftlicher Grundbildung zu nähern, stellen *Multiple-Choice-Tests* dar. Diese Tests können nicht nur ökonomisch durchgeführt werden, sondern bieten zusätzlich die Möglichkeit, die erhobenen Daten ökonomisch auszuwerten und zu interpretieren. Testpersonen können sich zwar in gewisser Weise in ihren Antwort- und Ausdrucksmöglichkeiten beschränkt fühlen (Sadler, 2000), jedoch stellen diese Tests im Falle eines – wenn ein detailliertes Manual vorliegt – eine Möglichkeit dar, Daten auf standardisierte und objektive Weise zu erfassen und zu interpretieren. Leider werden Multiple-Choice-Aufgaben oft in ihrem Potenzial, kognitive Leistungen zu erfassen, unterschätzt (Duit et al., 2001).

Das Prinzip von Multiple-Choice (MC)-Aufgaben besteht darin, dass ausgehend von einem Aufgabenstamm, in dem alle notwendigen Informationen und die Aufgabenstellung vermittelt werden, aus einer bestimmten Anzahl vorgegebener Möglichkeiten die richtige Antwort oder die richtige Kombination aus Antworten auszuwählen ist. In der Regel ist nur eine der Antwortmöglichkeiten richtig und alle übrigen

falsch.

Als ein Nachteil von MC-Tests wird oft die im Vergleich zu offenen Formaten hohe Ratewahrscheinlichkeit betrachtet. Bei genauerer Betrachtung liegt diese bei vier Antwortmöglichkeiten im einfachen MC-Format jedoch lediglich bei 25%. In diesem Fall sollte die Ratewahrscheinlichkeit bei entsprechender Berücksichtigung in der Interpretation der Daten ein geringes Problem darstellen. Die Reliabilität von MC-Verfahren liegt über der von offenen Formaten, was durch das objektive Bewertungssystem dieser Verfahren begründet ist und besonders dann ins Gewicht fällt, wenn die Testzeit eine Stunde übersteigt (Haladyna, 1994).

Um zu zeigen, wie vielfältig MC-Tests sein können, werden im Folgenden beispielhaft einige typische MC-Formate vorgestellt, anhand derer die Möglichkeiten und Grenzen dieses Verfahrens verdeutlicht werden.

MC-Tests können Haladyna zufolge (1994) unter anderem folgende Formate annehmen :

- *Konventionelles Format (vgl. Abbildung 2.7)*:
  - Dieses Format stellt das am häufigsten verwendete dar.
  - Es besteht aus drei Teilen: 1) Itemstamm, 2) richtige Antwort, 3) mehrere falsche Antworten, auch Distraktoren genannt.
  - Der Stamm stellt den Stimulus für die Antwort dar und sollte alle Informationen zu dem zu lösenden Problem enthalten.
  - Der Stamm kann die Form einer Frage, eines zu ergänzenden Satzes oder eines ganzen Szenarios inklusive Bildmaterial annehmen.

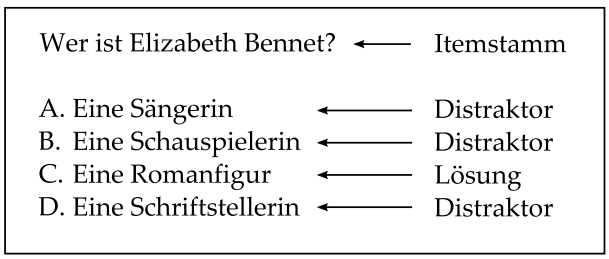


Abbildung 2.7: Beispiel für ein konventionelles MC-Format

- *Zuordnungs-Format (vgl. Abbildung 2.8)*:

- Dieses Format stellt eine beliebte Variation des konventionellen Formats dar.
- Es besitzt zwei oder mehr Optionen (im Beispiel A-E), die von den Itemstämmen (1-6) gefolgt werden.
- Es ist gut geeignet, um Assoziationen, Definitionen oder Charakteristika von Konzepten zu testen.
- Wichtig zu beachten ist, dass es mehr Itemstämme als Optionen geben sollte, um ein zu leichtes Ausschlussverfahren zu verhindern.

Ordne bitte folgende Bundesländer den unten stehenden Beschreibungen zu und trage deine Antworten in das Antwortblatt ein.

A. Sachsen
B. Bayern
C. Berlin
D. Schleswig-Holstein
E. Niedersachsen

1. Wird durch einen Kanal geteilt, der zu den bedeutendsten Wasserstraßen der Welt gehört.
2. In der Landeshauptstadt dieses Bundeslandes steht die Semperoper.
3. Dieses Bundesland heißt wie die Hauptstadt der Bundesrepublik Deutschland.
4. In diesem Bundesland befindet sich der höchste Berg Deutschlands.
5. Dies ist das zweitgrößte Bundesland Deutschlands.
6. Dies ist das bevölkerungsreichste Bundesland Deutschlands.

Abbildung 2.8: Beispiel für ein MC-Format mit Zuordnung

Weitere Formen des MC-Formats sind *komplexe Multiple-Choice-Aufgaben* und *Richtig-Falsch-Formate*. Komplexe MC-Formate stellen insofern eine Erweiterung des einfachen Formates dar, als dass mehr als eine Antwort richtig ist. Untersuchungen haben ergeben, dass Items dieses Formats im Vergleich zu einfachen MC-Aufgaben schwieriger zu entwickeln sind, mehr Zeit zum Lesen erfordern und schlechter diskriminieren (Haladyna, 1994).

Im Richtig-Falsch-Format muss jede einzelne Aussage mit *richtig* oder *falsch* bewertet werden. Anders als komplexe MC-Aufgaben sind diese Aufgaben relativ leicht herzustellen, leicht auszuwerten und erfordern wenig Lesezeit. Nachteilig ist, dass sie oft trivial erscheinen und die Ratewahrscheinlichkeit dadurch erhöht wird.

Um die Nachteile beider Testformen zu überwinden, wurden beide kombiniert und zu einem multiplen Richtig-Falsch-Format zusammengefügt.

Bewerte die Antworten auf die folgende Frage mit **A** für „richtig“ und **B** für „falsch“ und trage deine Bewertungen auf dem Antwortblatt ein.

Was beeinflusst die Reliabilität von Teswerten?

1. Testlänge
2. Varianzhomogenität
3. Itemlänge
4. Variablität der Testwerte
5. Dimensionalität

Abbildung 2.9: Beispiel eines aus komplexem MC und Richtig-Falsch-MC gemischten Formats

- *Multiples Richtig-Falsch-Format (vgl. Abbildung 2.9)*:
  - Das Format hat günstige Eigenschaften in Bezug auf Reliabilität und Validität.
  - Testpersonen akzeptieren dieses Format eher als das einfache MC-Format.
  - Es handelt sich um ein effizientes Format, was die Aufgabenentwicklung, die Zeit der Bearbeitung und damit die Anzahl an Fragen, die in einem bestimmten Zeitraum gestellt werden können, betrifft.
  - Eine gewisse Ratewahrscheinlichkeit muss bei der Interpretation der Testdaten beachtet werden. Diese Ratewahrscheinlichkeit kann durch einen ausreichend großen Itempool verringert werden.

Multiple-Choice-Tests sind insbesondere im nordamerikanischen Raum sehr beliebt, um Schülerleistungen zu prüfen. Das formale Abtesten des Verständnisses naturwissenschaftlicher Inhalte in Form von Multiple-Choice-Tests geht zurück auf das *Boston Survey* von 1845 (Sadler, 2000). Seitdem wurden viele Tests zur Erfassung naturwissenschaftlicher Inhalte und Konzepte entwickelt. Ein nicht geringer Teil von ihnen wurde in der Absicht entwickelt, das Verständnis des Prozesses wissenschaftlicher Erkenntnisgewinnung zu erfassen. Die Anzahl theoretisch ausreichend begründeter und darüber hinaus validierter Verfahren ist jedoch leider bis heute gering geblieben. Die Tabellen 2.5 und 2.6 geben Auskunft über MC-Testverfahren, die sich mit der Erfassung der prozessbezogenen Komponente naturwissenschaftlicher Grundbildung befassen. Diese Aufstellung erhebt keinen Anspruch auf Vollständigkeit, zeigt jedoch die bekanntesten Verfahren. Bei den dargestellten Trennschärfen und Schwierigkeiten handelt es sich um klassische Itemkennwerte. Zwei aktuelle deutschspra-

Tabelle 2.5: Multiple-Choice-Tests, die sich mit dem Prozess naturwissenschaftlicher Erkenntnisgewinnung beschäftigen

Jahr	Instrument	Autor(en)	Ziel der Messung	Zielgruppe
1961	Test on Understanding Science (TOUS), Form W 60 Items Multiple-Choice-Test, 3 Subskalen) Reliabilität (K.-R.): .76	Cooley, W. & Klopfer, L.	Verständnis von der Naturwissenschaft als Unternehmung, von Wissenschaftlern und von Methoden und Zielen der Wissenschaft	Schüler/innen, Lehrer
1970	Test zur Messung des „understanding of the nature of enquiry“ 6 Subskalen á 5 Items 30 Multiple-Choice-Items (Teil I des Tests) Item-Schwierigkeit: .15 - .87 Reliabilität: .86	Jungwirth, E.	Verständnis von der Natur der Forschung die 6 Subskalen messen: Verständnis wissenschaftl. Literatur; Manipulieren gegebener Daten (z.B. zum Vergleichen); Beurteilen der Angemessenheit und Relevanz experimenteller Prozeduren; Erkennen von Annahmen, die einer Hypothese oder einem Experiment zu Grunde liegen; Erkennen, welche Hypothese ein Experiment testet; angemessene Schlüsse aus Daten ziehen	Schüler/innen der 7.-12. Klasse
1971	Test of Science Processes 96 Multiple-Choice-Items (5 Kategorien) 8 Subskalen (Reliabilität von .30 - .80) Reliabilität (K.-R.): .91	Tannenbaum, R.S.	Verständnis von Wissenschaftsprozessen: die 8 Subskalen messen: Beobachten, Klassifizieren, Quantifizieren, Messen, Experimentieren (Planen, Durchführen und Interpretieren eines Experiments), Schlussfolgern und Vorhersagen	Schüler/innen der 7.-9. Klasse
1980	Test of Integrated Science Process Skills (TIPS) 36 Multiple-Choice-Items mittlere Item-Trennschärfe: .40 mittlere Item-Schwierigkeit: .53 Reliabilität (Cronbach's Alpha): .89	Dillashaw, F.G. & Okey J.R.	Fertigkeiten, die das Planen, Durchführen und die Interpretation von Ergbnissen wissenschaftlicher Untersuchungen betreffen: Formulieren von Hypothesen; Definieren, Kontrollieren und Manipulieren von Variablen; Planen von Untersuchungen; Interpretieren von Daten	Schüler/innen der 9.-10. Klasse
1980	Test of Enquiry Skills (TOES) 87 Multiple-Choice-Items (5 Kategorien) 9 Subskalen (3 Teile á 3 Skalen) Reliabilität (Retest): .65 - .82 (im Mittel .73)	Fraser, B.J.	Fertigkeiten, die das Nutzen von Referenzmaterial, das Interpretieren und Verarbeiten von Informationen und kritisches wissenschaftliches Denken (u.a. Design experimenteller Abläufe und Schlussfolgerungen aus Daten ziehen) betreffen.	Schüler/innen der 7.-10. Klasse

chige Verfahren zur Messung des Verständnisses vom Prozess wissenschaftlicher Erkenntnisgewinnung sind der Kompetenztest, den Phan (2007) beschreibt und der *Naturwissenschaftliche-Arbeitsweisen (NAW)-Test*, der von Kieren (2004) und Walpuski (2006) zunächst für die 7. Klasse entwickelt wurde und den Henke (2006) für die 12.

Tabelle 2.6: Forts. Multiple-Choice-Tests, die sich mit dem Prozess naturwissenschaftlicher Erkenntnisgewinnung beschäftigen

Jahr	Instrument	Autor(en)	Ziel der Messung	Zielgruppe
1982	Test of Integrated Science Processes (TISP) 34 Multiple-Choice-Items mittlere Item-Trennschärfe: .32 mittlere Item-Schwierigkeit: .42 Reliabilität (Cronbach's Alpha): .96	Tobin, K.G. & Capie, W.	Fertigkeiten, die das Planen, Durchführen und die Interpretation von Ergebnissen wissenschaftlicher Untersuchungen betreffen, u.a: Identifizieren abhängiger und unabhängiger Variablen	Schüler/innen der 8. Klasse, Lehrer
1985	Test of Integrated Science Process Skills (TIPS II) 36 Multiple-Choice-Items mittlere Item-Trennschärfe: .35 mittlere Item-Schwierigkeit: .53 Reliabilität (Retest): .86	Burns, J.C., Okey, J.R. & Wise, K.C.	Fertigkeiten, die das Planen, Durchführen und die Interpretation von Ergbnissen wissenschaftlicher Untersuchungen betreffen: Formulieren von Hypothesen; Definieren, Kontrollieren und Manipulieren von Variablen; Planen von Untersuchungen; Interpretieren von Daten	Schüler/innen der 6.-12. Klasse
1989	Processes of Biological Investigations Test (PBIT) 35 Items (einfache Multiple-Choice und Multiple-Choice-Zuordnungsaufgaben) mittlere Item-Trennschärfe: .41 mittlere Item-Schwierigkeit: .55 Reliabilität (Cronach's Alpha): .86 Reliabilität (K.-R.): .84	Germann, P.J.	Fertigkeiten, die Schüler in BSCS Biologie-Kursen benötigen: Beurteilung der Plausibilität von Hypothesen; Beurteilung der Logik von Vorhersagen; notwendige Annahmen erkennen; Hypothesen vor dem Hintergrund gegebener Daten beurteilen; Beurteilung der Verallgemeinerung präsentierter Daten; Entscheiden, welche Hypothese durch gegebene Daten gestützt wird; Wahrscheinlichkeit und Plausibilität von Gründen und Effektivität von Handlungen beurteilen	Schüler/innen
2007	Kompetenztest Multiple-Choice-Test mittlere Item-Trennschärfe: .29 - .59 mittlere Item-Schwierigkeit: .52 - .72 Reliabilität (Cronach's Alpha): .63 - .77	Phan, T.T.H.	Fertigkeiten, die das Planen, Durchführen und die Interpretation von Ergbnissen wissenschaftlicher Untersuchungen betreffen: in Anlehnung an das SDDS-Modell von Klahr & Dunbar (1988) bzw. Klahr (2000) werden die Fertigkeiten zur Suche im Hypothesenraum, zum Testen von Hypothesen und zur Evaluation von Befunde getestet	Schüler/innen der 5. und 6. Klasse
2008	Naturwissenschaftliche-Arbeitsweisen (NAW)-Test Reliabilität (Cronach's Alpha Jg. 7): .73 Reliabilität (Cronach's Alpha Jg. 12): .82	Klos, S., Henke, C., Kieren, C., Walpuski, M. & Sumfleth, E.	Fertigkeiten, die die Ideen-/ Hypothesenbildung, die experimentelle Umsetzung der Hypothesen und das Ziehen von Schlussfolgerungen aus einer Daten- bzw. Befundlage betreffen	Schüler/innen der 7. und 12. Klasse.

Jahrgangsstufe weiterentwickelte. Diese beiden Verfahren sollen an dieser Stelle etwas genauer betrachtet werden.

Phan (2007) entwickelte im Rahmen ihrer Dissertation einen Kompetenz- und einen

Wissenstest für den Bereich der Biologie, um zu untersuchen, ob ein MC-Test in Paper-and-Pencil-Form in der Lage ist, unterschiedliche Kompetenzstufen des Experimentierens zu erfassen, auf welche Weise die in Klahrs SDDS-Modell angenommenen Dimensionen interagieren und auf welche Weise das Vorwissen über bestimmte Bereiche und die Kompetenzen des Experimentierens in den gleichen Bereichen zusammenhängen.

Phans Ergebnisse zeigen, dass es möglich ist, anhand eines Multiple-Choice-Tests verschiedene Kompetenzlevel hinsichtlich der Fertigkeiten im Bereich des Experimentierens festzustellen. Weiterhin ergaben die Untersuchungen, dass die einzelnen Dimensionen des Experimentierens stark korrelativ zusammenhängen, was eventuell durch den starken Einfluss des Vorwissens begründet ist. Darüber hinaus konnten bisherige Forschungsergebnisse bestätigt werden, die zeigen, dass das Vorwissen über die ausgewählten biologischen Inhaltsbereiche und das Abschneiden im Kompetenztest signifikant zusammenhängen.

Die Unterscheidung verschiedener Kompetenzlevel anhand eines MC-Tests stellt ein bedeutsames Ergebnis dar. Grenzen dieser Arbeit bestehen darin, dass Kompetenz- und Wissenstest ohne weitere Prüfung der Validität direkt verwendet wurden, um die genannten Forschungsfragen zu beantworten. Ohne die Überprüfung, ob tatsächlich die gewünschte Kompetenz gemessen wurde, ist es jedoch nicht möglich, Aussagen über den Zusammenhang dieser Kompetenz mit anderen Konstrukten zu treffen. Die Kombination von Testentwicklung und Beantwortung von Forschungsfragen schränkt die Aussagekraft insgesamt ein. Ein weiterer Punkt, der bei Betrachtung der thematisch in Einheiten zusammengefassten Aufgaben deutlich wird, ist, dass einzelne Aufgaben dieser Einheiten nicht unabhängig voneinander sind. Das richtige Beantworten einer Aufgabe hat z.B. zur Konsequenz, dass die darauffolgende Aufgabe mit einer höheren Wahrscheinlichkeit auch gelöst wird.

Ein weiteres Verfahren, das bezüglich der Abhängigkeit der Aufgaben ähnliche Probleme aufweist, ist der *Naturwissenschaftliche-Arbeitsweisen (NAW)-Test* (Klos, Henke, Kieren, Walpuski & Sumfleth, 2008). Dieser Test hat bereits eine längere Entwicklung durchlaufen. Die erste Version der Aufgaben entstand im Rahmen einer Staatsexamensarbeit (Kieren, 2004), welche die Entwicklung eines Testverfahrens zum naturwissenschaftlichen Arbeiten zum Ziel hatte. Seitdem haben sich durch Überarbei-

tungen dieses Verfahrens eine Testversion für die 7. Jahrgangsstufe (Walpuski, 2006) und eine für die 12. Jahrgangsstufe (Henke, 2006) entwickelt. Der NAW-Test wurde anders als der Kompetenztest von Phan (2007) bereits auf unterschiedliche Weisen validiert. Anhand von Korrelationen mit dem Kognitiven-Fähigkeiten-Test (KFT) (Heller & Perlet, 2000) und einem Chemie-Fachwissenstest wurde die diskriminante Validität überprüft. Der korrelative Zusammenhang mit dem Fachwissenstest Chemie fiel nur gering und nicht signifikant und die Korrelation mit dem KFT zwar signifikant, aber gering aus. Die Testautoren schlossen aus diesen Ergebnissen, dass der NAW-Test ein von Fachwissen und kognitiven Fähigkeiten verschiedenes Konstrukt misst.

Zur Prüfung der Konstruktvalidität wurde in Ermangelung eines passenden Paper-and-Pencil-Verfahrens mit einer kleinen Gruppe von Schülerinnen und Schülern ein Performanztest durchgeführt, dessen Ergebnisse mit dem des NAW-Tests korreliert wurden. Die resultierende signifikante Korrelation in mittlerer Höhe wurde als weiteres Indiz für die Validität des Instruments betrachtet.

Faktorenanalytisch konnte in beiden Altersstufen bestätigt werden, dass mit dem NAW-Test nur *ein* Konstrukt erhoben wird. Die Ergebnisse zeigen ein bereits validiertes Verfahren, welches das Verständnis vom Ablauf und vom Wesen naturwissenschaftlicher Erkenntnisgewinnung zu erfassen vermag. Die Grenzen dieses Verfahrens liegen wie auch schon bei dem von Phan (2007) entwickelten Verfahren darin, dass Aufgaben nicht unabhängig sind. Damit hat das richtige oder falsche Beantworten einer Aufgabe Auswirkungen auf die Attraktivität der Antwortalternativen folgender Aufgaben. Weiterhin ist kritisch zu betrachten, dass die Antworten auf die Aufgaben teilweise direkt dem Text zu entnehmen sind, was nahe legt, dass an diesen Stellen vorwiegend Lese- und Textverständnis erfasst wird.

Die dargestellten Versuche und Grenzen der Erfassung prozessbezogener naturwissenschaftlicher Grundbildung werden in Abschnitt 2.7.2 noch einmal aufgegriffen, um aus den Informationen Anforderungen an ein zu konstruierendes Testverfahren und die Aufgabenstellung dieser Arbeit abzuleiten.

2.4.6 SPEED ODER POWER?

Neben den Überlegungen, welches Verfahren grundsätzlich für die Erfassung prozessbezogener naturwissenschaftlicher Grundbildung in Form eines Screenings geeignet ist, muss auch eine Entscheidung darüber getroffen werden, ob das Verfahren als *Speed- oder Powertest* (Geschwindigkeits- oder Niveautest) angelegt werden sollte. Im Folgenden werden beide Testformen dargestellt, um beide Varianten gegeneinan-

der abzuwägen.

> *„Geschwindigkeitstests sind dadurch definiert, dass bei unbegrenzter Zeitvorgabe alle Items von allen Probanden gelöst werden, d.h. ihr Schwierigkeitsgrad konvergiert dann gegen Null. Die Differenzierung zwischen den Probanden wird nur durch die Begrenzung der Bearbeitungszeit erreicht."* (Amelang & Zielinski, 2002, S. 118)

Geschwindigkeitstests werden häufig zur Prüfung der Konzentration eingesetzt, erscheinen jedoch zur Messung kognitiver Kompetenzen wenig sinnvoll. Für diesen Zweck müssen Niveautests herangezogen werden.

> *„Niveautests sind dadurch definiert, dass auch bei unbegrenzter Zeitvorgabe von keinem Testteilnehmer alle Aufgaben richtig gelöst werden. Mit derartigen Verfahren wird primär das intellektuelle Niveau oder die „Denkkraft" (Power) ermittelt."* (Amelang & Zielinski, 2002, S. 119)

Reine Powertests sind aus technischen Gründen meist nicht möglich, da jede Testvorgabe eine zeitliche Begrenzung haben muss. Daher sind im Rahmen der Erfassung kognitiver Kompetenzen Mischformen üblich. Hier ist zu beachten, dass mit zunehmender Speedkomponente die Werte für Schwierigkeit, Trennschärfe und Homogenität an Aussagekraft verlieren. Im Fall einer starken Speedkomponente tritt die Erfassung der Verarbeitungszeit gegenüber der Messung der inhaltlichen Kompetenz in den Vordergrund und die Fehlerquote nimmt zu (Schweizer, 2006). Wenn es dagegen darum geht, das Kompetenzpotential einer Person zu erfassen, sollte die Powerkomponente überwiegen. Es sollte so viel Zeit bemessen werden, dass in der Regel alle Testpersonen bis zur letzten Aufgabe vordringen. Auf diese Weise versucht man zu umgehen, dass eine unterschiedliche Anzahl nicht bearbeiteter Aufgaben zu rechnerischen Problemen und zu Interpretationsschwierigkeiten führt (Rost, 2004a).

2.5 ARTEN UND WIRKUNG DER INFORMATIONSDARSTELLUNG IN TESTAUFGABEN

Nachdem sich die vergangenen Abschnitte mit inhaltlichen Definitionen von Kompetenzen und Fertigkeiten, der Definition der prozessbezogenen naturwissenschaftlichen Grundbildung und mit möglichen Messverfahren beschäftigt haben, stehen nachfolgend Überlegungen zur Art der Informationsdarstellung in Testaufgaben im Vordergrund. Die Informationen, die die Grundlagen der Testaufgaben bilden, sollten

möglichst einfach, kurz und klar verständlich dargestellt werden, um ihre Aufnahme und Verarbeitung zu erleichtern. Insbesondere in Bezug auf Graphiken gibt es unterschiedliche Arten der Informationsdarstellung. Die gleiche Information wird je nach Darstellung leicht oder weniger leicht verstanden.

In dem folgenden Abschnitt wird die Darstellung von Informationen anhand von Texten, Bildern, Diagrammen und Tabellen in den Fokus genommen. Die theoretischen und empirischen Betrachtungen dienen allein der Entscheidung, auf welche Weise Informationen im Rahmen des neuen Testverfahrens dargestellt werden sollten. Eine umfassende Betrachtung der Wirkung von Informationsdarstellungen ist an dieser Stelle nicht vorgesehen.

2.5.1 TEXT

Bei der Verwendung von Text in Testaufgaben ist zu beachten, dass dieser möglichst kurz, einfach und unmissverständlich sein sollte. Die einzelnen Sätze sollten kurz, der Satzbau einfach, ohne die Verschachtelung von Nebensätzen sowie ohne doppelte Verneinungen gestaltet werden. Darüber hinaus sollten die verwendeten Ausdrücke möglichst bekannt und der Zielgruppe angemessen sein. Lange Textpassagen sind zu vermeiden, weil sie den Einfluss der Lesekompetenz auf die Testleistung unnötig erhöhen. Eingesetzter Text sollte alle notwendigen Informationen vermitteln und auf überflüssige, ausschmückende Informationen verzichten. Alle Testpersonen sollen möglichst unabhängig von ihrer Lesekompetenz in die Lage versetzt werden, Aufgabeninhalte zu erfassen und zu verarbeiten. Ein gewisser Einfluss der Lesekompetenz wird bei der Erfassung von Aufgabeninhalten in Textform immer vorhanden und nicht vermeidbar sein.

Das Lesen von Texten stellt einen komplexen Vorgang dar, der aus mehreren Teilprozessen besteht. Dazu gehören das Erkennen von Buchstaben, Wörtern sowie die Erfassung von Wortbedeutungen, das Herstellen semantischer und syntaktischer Relationen zwischen Sätzen, die satzübergreifende Integration von Sätzen zu Bedeutungseinheiten und der Aufbau einer kohärenten mentalen Repräsentation der Bedeutung eines Textes (Artelt, Stanat, Schneider & Schiefele, 2001).

Lesen geschieht in Form einer aktiven Auseinandersetzung mit Geschriebenem und das Leseverständnis folgt aus genau dieser aktiven Auseinandersetzung. Dies erfordert auf Seiten des Lesers kognitive Grundfähigkeiten, Sprach- , Welt- und inhaltliches Vorwissen, strategische Kompetenz und auch ein gewisses Interesse. Gerade bezüglich der Motivation des Lesers erscheinen Textdarstellungen oft unattraktiver als Bilder, Graphiken oder auch Tabellen, die einen kurzen Überblick über Sachver-

halte geben.

Um den Einfluss der Lesekompetenz gering zu halten, sind folgende Punkte bei der Gestaltung von Texten zu beachten (Weidenmann, 2001):

- Verständlichkeit
  - Mit diesem Bereich hat sich die Leseforschung ausgiebig beschäftigt.
  - Das *Hamburger Verständlichkeitskonzept* (Langer, Thun & Tausch, 2006) hat versucht, aus den Textmerkmalen *Einfachheit* (Wortwahl, Satzbau usw.), *Gliederung* (Überschriften, Abschnitte usw.), *Kürze* (Knappheit) und *Anregung* (direkte Rede, Beispiele, Humor, Spannung) so genannte Lesbarkeitsindizes zu berechnen.
  - Je höher die Ausprägung der genannten Merkmale, desto besser soll die Verständlichkeit des Textes ausfallen.
- Kohärenz
  - Mit der Kohärenz ist der rote Faden des Textes gemeint.
  - Der Text sollte sinngemäß zusammengehörige Teile auch durch entsprechende Formulierungen deutlich hervorheben (z.B. durch *also*, *deshalb*, *weil* oder *auf Punkt XY gehe ich nun detailliert ein*).
- Reihenfolge
  - Die Anordnung von Informationen sollte einer nachvollziehbaren Logik folgen.
  - Hilfen wie Überschriften, Unterstreichungen, Zusammenfassungen usw. sorgen insbesondere bei längeren Texten für ein klares Nachvollziehen (Ballstaedt, 1997).

2.5.2 BILDER

Bilder sollen Aufmerksamkeit wecken und auch leseschwächeren Personen die schnelle Erfassung von Informationen ermöglichen. Dabei stellt sich die Frage, ob es neben den aus der PISA-Untersuchung bekannten *Literacy*-Konzepten (*Reading*, *Mathematical* und *Scientific Literacy*) auch so etwas gibt wie *Visual Literacy*, also die erlernbare Fähigkeit, Bilder zu *lesen* (Weidenmann, 2001). Diese Bezeichnung wird immer wieder kritisiert. Das Erkennen von *realistischen Bildern* (Schnotz, 1994), wie z.B. Fotos, unterscheidet sich in der Eigenschaft der Erlernbarkeit von den anderen *Literacy*-Konzepten, da z.B. das Identifizieren der Darstellungen solcher Bilder nicht erlernt werden muss. Anders verhält es sich mit dem Erkennen von Perspektiven auf Bildern

sowie mit bildlichen Symbolen, die erst erlernt werden müssen. So haben Mackworth und Bruner (1970) nachgewiesen, dass Erwachsene häufiger informationshaltige Teile eines Bildes fixieren als Kinder.

Die Vorliebe von Lernenden für Bilder ist empirisch belegt (Weidenmann, 2001). Dabei erweisen sich farbige Bilder im Vergleich zu schwarz-weißen als interessanter. Oft werden aus motivationalen Gründen daher bunte Illustrationen eingesetzt, die zusätzlich zum Lesen des Textes anregen sollen. Bild und Text können entweder in redundanter oder aber komplementärer Beziehung stehen. Ergänzen sich Bild und Text sinnvoll, so ermöglicht diese Komplementarität eine tiefere Verarbeitung der Informationen. Entweder löst das Bild dabei den Wunsch aus, mehr über das Dargestellte zu lesen oder aber das Gelesene weckt umgekehrt den Wunsch, es in Form eines Bildes veranschaulicht zu betrachten (Weidenmann, 2001; Peeck, 1994). Die Wirkungen von Texten und Bildern sind für die Entwicklung von Testaufgaben wichtig. Denn obwohl Bilder häufig als erstes betrachtet werden, investieren Personen weniger mentale Anstrengung in sie als in einen Text. Text und Bild müssen zusammenpassen und dürfen die Aufmerksamkeit des Betrachters nicht in eine Richtung ablenken. Zur Vermittlung eines bestimmten Sachverhaltes in Testaufgaben sind vor allem darstellende Bilder (einfache Zeichnungen oder Fotografien) geeignet. Diese Art von Bildern kann dazu dienen, Textinformationen zu konkretisieren, Objekte, Ereignisse oder Aktionen zu visualisieren, schwer zu verarbeitende Texte verständlich zu gestalten oder die Struktur und den Zusammenhang von Textinhalten zu verbessern (Peeck, 1994). Damit können sie insbesondere leseschwächere Personen beim Verstehen von Texten unterstützen.

2.5.3 DIAGRAMME

Bei realistischen Bildern besteht eine konkrete Form der strukturellen Übereinstimmung mit dem Gegenstand, wobei sich hier die Darstellung natürlich auf die Möglichkeiten des zweidimensionalen Raumes beschränken muss. Demgegenüber besteht bei Diagrammen, die auch als *logische Bilder* bezeichnet werden (Schnotz, 1994), eine abstrakte Form der Übereinstimmung mit dem dargestellten Gegenstand, da in diesem Fall die Informationen über diesen Gegenstand z.B. in Form von Balken oder Säulen dargestellt sind. Da Diagramme im Vergleich zu realistischen Bildern keine Ähnlichkeit mit dem dargestellten Gegenstand besitzen, sondern ähnlich wie Symbole eine konventionalisierte Struktur besitzen, ist es dem ungeübten Leser nicht möglich, auf alltägliche kognitive Wahrnehmungsschemata zurückzugreifen. Das Verstehen von Diagrammen ist eine Kulturfertigkeit, die erst erlernt werden muss (Schnotz,

1997).

Die Effektivität von Diagrammen hängt von der Interaktion der Darstellung mit dem menschlichen kognitiven System ab. Man unterscheidet hier zwischen *subsemantischen* und *semantischen* Verarbeitungsprozessen (Schnotz, 1997). Die *subsemantische* Verarbeitung besteht im Erkennen, in der Diskrimination und Identifikation sowie in der Gruppierung der graphischen Komponenten eines Diagramms. Diese Prozesse verlaufen weitestgehend automatisch und sind nur in geringem Maße vom Vorwissen abhängig. Bei der Wahrnehmung der Komponenten eines Diagramms spielen vielmehr die sogenannten Gestaltgesetze[11] eine Rolle. So werden beispielsweise diejenigen Komponenten zusammengefasst, die nahe beieinander liegen oder die sich ähnlich sind (z.B. gleiche Farbe, Form oder Textur). Auf diese Weise werden einzelne Komponenten zu größeren Einheiten zusammengefasst. Dieser erste Schritt der Diagrammverarbeitung auf Wahrnehmungsebene mündet dann in die eigentlichen Verstehensprozesse. Um den Testpersonen die erfolgreiche Bewältigung dieses ersten Schrittes zu ermöglichen, ist es notwendig, dass Striche, Farben, Formen und Struktur deutlich und von ausreichender Größe sind.

Nach der subsemantischen folgt die *semantische* Verarbeitung des Diagramms, die in einer konzeptgeleiteten Analyse der wahrgenommenen graphischen Konfiguration besteht und im Verstehen der Darstellung mündet. Diese Verarbeitung ist abhängig vom Vorwissen und durch Instruktionen bzw. Bildunterschriften beeinflussbar. Sie kann in Form der Vorgabe einer bestimmten Verarbeitungsreihenfolge erfolgen. In diesem Zusammenhang spielen auch kulturspezifische Verarbeitungsgewohnheiten, wie z.B. das Lesen von links nach rechts, eine Rolle. Um Diagramme verstehen zu können, müssen Betrachter die Schemata erlernt haben, mit denen Werte und Zusammenhänge zwischen Variablen und Einzelwerten dargestellt werden. Befunde zeigen, dass Lernende mit höherem Vorwissen auf diesem Gebiet eher in der Lage sind, nach übergreifenden visuellen Mustern zu suchen als Lernende mit geringem Vorwissen, die bevorzugt nach lokal begrenzten Einzelinformationen suchen (Schnotz, 1997).

Zur Gestaltung von Diagrammen und Text geben Schnotz (1994, 1997) und Lowrie und Diezmann (2007) folgende Hinweise:

11 Die Gestaltgesetze entstanden im Rahmen der Gestaltpsychologie, die zu Beginn des 20. Jahrhunderts eine neue psychologische Strömung darstellte. Ein Gestaltgesetz bezeichnet die Art des Zusammenschlusses von erlebten Teilen zu einer erlebten Ganzheit. Dieser Zusammenschluss erfolgt so, dass ein möglichst einfaches, einheitliches, geschlossenes und symmetrisches Ganzes entsteht. Einer der Vertreter dieser Strömung, Max Wertheimer formulierte sechs Gestaltgesetze: Gesetz der Nähe, der Ähnlichkeit, der guten Gestalt, der guten Fortsetzung, der Geschlossenheit und des gemeinsamen Schicksals (Wertheimer, 1938)

- Das Merkmal Farbe sollte benutzt werden, um qualitative Unterschiede darzustellen. Es ist ungeeignet für die Darstellung quantitativer Unterschiede.
- Zur Darstellung von Quantitäten sollten möglichst Balken oder Säulen verwendet werden.
- Die Darstellung sollte inhaltlich und formal sparsam erfolgen:
  - Es sollten nur die Variablen dargestellt werden, die den Sachverhalt beschreiben. Auf die Darstellung zusätzlicher Variablen ohne Informationsgehalt sollte verzichtet werden.
  - Es sollte auf visuelle Effekte verzichtet werden, die nicht der Informationsdarstellung dienen.
- Die Wahrnehmung der graphischen Konfiguration sollte möglichst gut mit der zu vermittelnden Struktur des Sachverhaltes übereinstimmen.
- Der Komplexitätsgrad sollte der Zielgruppe angepasst sein.
- Die Einbettung von Diagrammen sollte anhand von klaren Instruktionen und Verarbeitungshinweisen im Text erfolgen, um einer oberflächlichen Verarbeitung durch den Leser entgegenzuwirken.
- Graphische und verbale Informationen sollten in räumlicher Nähe zueinander stehen.
- Die Darstellung von Sachverhalten sollte den üblichen Konventionen folgen:
  - *Kreisdiagramme* sollten vorwiegend zur Darstellung der Zusammensetzung einer Variable genutzt werden.
  - *Säulen-* und *Balkendiagramme* sollten zur Visualisierung von quantitativen Merkmalsausprägungen eingesetzt werden, bei denen sich die Merkmalsträger nur qualitativ voneinander unterscheiden.
  - Die Visualisierung von Entwicklungsverläufen sollte anhand von *Liniendiagrammen* erfolgen. Dies gilt insbesondere, wenn es um die Darstellung mehrerer Entwicklungsverläufe in einem Diagramm geht.

2.5.4 TABELLEN

Das Lesen von Tabellen gehört ebenso wie das Lesen von Graphiken zu Kulturtechniken, die erlernt werden müssen und sich dadurch vom Lesen realistischer Bilder

abgrenzen. Dies ist unter anderem dadurch begründet, dass die Informationsdarstellung in Form von Zahlen oder Texten ohne Lese- oder mathematische Grundbildung nicht lesbar und verständlich sein kann. Tabellen werden nicht genutzt, um Informationen oder Daten in eine andere, bildhafte Sprache zu übersetzen, sondern stellen eine Form der graphischen Strukturierung dar, um vorliegendes Datenmaterial übersichtlich und zusammenfassend darzustellen.

Anders als Diagramme werden Tabellen meist genutzt, wenn es um die Darstellung von Informationen mit geringerem Komplexitätsgrad geht. Für die Darstellung derartiger Informationen haben sie sich als durchaus geeignet erwiesen (Wainer, 1980). Werden die dargestellten Daten zu komplex, so müssen andere Formen der Darstellung gefunden werden.

Für die Erstellung von Testaufgaben ist es wichtig, den Komplexitätsgrad von Tabellen gering zu halten. Die Extraktion der zur Aufgabenbearbeitung nötigen Informationen sollte möglichst einfach sein. Tabellen sollen das Lesen langer Textpassagen ersparen und Informationen übersichtlich präsentieren. Wainer (1992) gibt folgende Hinweise, um die Darstellung und Vermittlung von Informationen anhand von Tabellen möglichst effektiv und übersichtlich zu gestalten:

- Zeilen und Spalten sollten sinnvoll und logisch geordnet sein.
- Dargestellte Informationen sollten nach Größe, Alphabet oder Jahreszahl aufsteigend sortiert sein; sie sollten gemäß derjenigen Information sortiert sein, um die es in der Testaufgabe vorrangig geht.
- Zahlen sollten gerundet dargestellt werden.
- Überflüssige Informationen, die für die Aufgabe nicht benötigt werden, sollten nicht dargestellt werden.
- Wichtige Informationen sollten hervorgehoben werden.

Diese relativ kurz gehaltenen Hinweise zur Darstellung von Informationen in Tabellen sollen an dieser Stelle genügen, da Tabellen in den Aufgaben dieser Testentwicklung nur selten eingesetzt wurden und dadurch einen geringen Stellenwert einnehmen.

2.6 KRITERIEN ZUR TESTVALIDIERUNG

Anhand des folgenden Abschnitts wird das Heranziehen des Leistungskonstrukts *Schulnote* und des Motivationskonstrukts *Interesse* zur Validierung des neu entwi-

ckelten Tests theoretisch begründet. Ohne dabei zu sehr auf die Bedeutung und die unterschiedlichen Ansätze der Testvalidierung einzugehen, (dies erfolgt erst im Rahmen des Methodenteils), sei an dieser Stelle recht allgemein erklärt, dass die Validierung eines Tests dazu dient, festzustellen, ob er tatsächlich misst, was er zu messen beansprucht. Eine Möglichkeit, diesen Sachverhalt zu klären, besteht darin, die erreichten Testwerte in einen korrelativen Zusammenhang mit solchen Variablen zu setzen, welche mit dem Testverhalten in einem theoretisch begründeten und engen Zusammenhang stehen. Üblicherweise werden zur Prüfung der Validität Tests herangezogen, die das Gleiche messen. Da jedoch die Testentwicklung im Rahmen dieser Arbeit gerade in Ermangelung eines solchen Tests durchgeführt wurde, mussten andere Variablen zur Validierung gefunden werden.

Bei der Betrachtung des Motivationskonstrukts *Interesse* als möglicher Zugang zur Validierung geht es einerseits um das allgemeine naturwissenschaftliche Fachinteresse und andererseits um das spezielle Interesse an wissenschaftsbezogenen Aktivitäten und Forschungsprozessen. Im Folgenden wird begründet, welches Interessekonzept der Validierung zu Grunde liegt und welche Indikatoren zur Messung von Interesse herangezogen werden. Grundsätzlich stellt das Interesse an Naturwissenschaften deshalb ein Kriterium zur Validierung dar, da die Forschung einen Zusammenhang zwischen Interesse und Leistung zeigt. Über unterschiedliche Schularten, Jahrgangsstufen und Schulfächer hinweg liegt die Interesse-Leistungs-Korrelation durchschnittlich bei $r = .30$ (Krapp, 1992; Wild, Krapp & Winteler, 1992). Dieser Zusammenhang nimmt noch zu, wenn nicht das allgemeine Interesse an einem Fach, sondern das Interesse für spezielle Teilbereiche eines Faches zur Korrelation herangezogen werden.

2.6.1 DAS LEISTUNGSKRITERIUM „SCHULNOTE“

Als Leistungskriterium zur Testvalidierung wurden die *Schulnoten in naturwissenschaftlichen Fächern* herangezogen. Schulnoten als Leistungskriterium zur Validierung von Tests zu verwenden hat eine lange Tradition, die sich je nach Zielvariable, die vorhergesagt werden soll, als mehr oder weniger sinnvoll erwiesen hat.

Problematisch ist vor allem die Tatsache, dass Schulnoten nicht nur fachliche Kompetenzen erfassen, sondern ein Produkt vieler Einflüsse sind. Beispielsweise spielen auch der Referenzrahmen des Schulniveaus, der Schule, der Klasse, Schülermerkmale und Erwartungen, Voreinstellungen und Diagnosekompetenz des Lehrers eine bedeutsame Rolle (Schrader & Helmke, 2001). Statt objektiver Kriterien (z.B. Lernzielen) spielt das klasseninterne Bezugssystem oft eine so große Rolle, dass vergleich-

bare Leistungen in unterschiedlichen Klassen nicht zur gleichen Beurteilung führen. Somit ist es nicht unproblematisch, die Noten in naturwissenschaftlichen Fächern als Prädiktor für das Abschneiden im Test zur Erfassung prozessbezogener naturwissenschaftlicher Grundbildung heranzuziehen, zumal dieser Bereich der Grundbildung sehr speziell ist und nur einen kleinen Teil der Note ausmachen kann.

Die Schulnoten werden hier also in dem Wissen als externes Validitätskriterium herangezogen, dass sie aufgrund der genannten Umstände von unterschiedlichem prädiktivem Wert sein können. Die Korrelation zwischen Abiturnoten und Studienleistungen liegen metaanalytisch bei $r = 0,46$. Die Korrelationen zwischen Haupt- bzw. Realschulnoten und Ausbildungserfolg erreichen lediglich eine Höhe von $r = 0,26$. Schuler merkt an, dass der Wert von $0,46$ von einzelnen eignungsdiagnostischen Verfahren nur schwerlich erreicht wird (Schuler, 2000). Die Korrelationen einzelner Fachnoten liegen noch unterhalb der genannten Werte. Die Mathematiknote erscheint mit Korrelationen um $0,30$ im Durchschnitt als bester Prädiktor, die Korrelationen mit den Naturwissenschaftsnoten liegen zwischen $0,26$ und $0,31$ (Schuler, 2006).

Es ist zu erkennen, dass Schulnoten ein schwieriges Validitätskriterium darstellen, da sie hinsichtlich ihrer Prognosekraft eine große Spannweite möglicher Korrelationen zeigen. Dennoch werden sie als gute Prädiktoren weiterer Bildung, also als Prädiktoren von Lernleistungen betrachtet. In etwas geringerem, aber nicht zu gering zu schätzendem Maße, scheinen sie für die Vorhersage von Ausbildungs- und Berufserfolg geeignet zu sein (Schuler, 2006).

Da der zeitliche Rahmen der Testung keine Erhebung eines eventuell angemesseneren Leistungsmaßes erlaubten, wurde hier auf das Leistungskriterium Schulnoten zurückgegriffen. Um für die Validierung einen Wert zu bestimmen, den die Korrelationen zwischen Naturwissenschaftsnoten und naturwissenschaftlicher Kompetenz erreichen sollten, wurden neben den genannten Forschungsergebnissen vor allem die PISA-Ergebnisse 2006 als Grundlage herangezogen (Schütte, Frenzel, Asseburg & Pekrun, 2007). Hier zeigten sich Korrelationen zwischen naturwissenschaftlicher Kompetenz und Naturwissenschaftsnoten, die Werte $< 0,4$ erreichten (Biologie: $r = 0,36$; Physik: $r = 0,34$; Chemie: $r = 0,35$). Aus diesem Grund werden innerhalb der externen Validierung des im Rahmen dieser Arbeit entwickelten Tests Korrelationen mit den Naturwissenschaftsnoten erwartet, die Werte $> 0,3$ annehmen sollten.

2.6.2 Zu Grunde liegendes Interessekonzept

Interesse wurde in der Bildungsforschung viel beschrieben und für diverse Studien herangezogen. Seit Herbart zu Beginn des 19. Jahrhunderts eine erste, noch nicht empirisch belegte, Interessentheorie aufstellte (Herbart, 1806/1965), hat sich im Laufe der Zeit eine eigene Interessenforschung entwickelt und etabliert. Diese Entwicklung führte dazu, dass der Begriff des Interesses immer weiter differenziert wurde. Eine grobe Einteilung, die an dieser Stelle als theoretische Ausführung genügen soll, besteht in den beiden Konzepten des *individuellen* und des *situationalen Interesses*. Ersteres wird als persönlichkeitsspezifisches Merkmal des Lerners betrachtet, als stabile Präferenz für bestimmte Lerngegenstände, Wissens- oder Handlungsgebiete. Letzteres stellt einen situationsspezifischen, motivationalen Zustand dar, der aus der Interessantheit einer bestimmten Lernsituation resultiert (Krapp, 2000, 1992). In beiden Fällen richtet sich das Interesse auf bestimmte Objekte, es ist also gegenstandsspezifisch.

Für die Validierung der Testitems dieser Arbeit wird das *individuelle* Interesse an Naturwissenschaften und an naturwissenschaftsbezogenen Tätigkeiten als stabiles Persönlichkeitsmerkmal zur Validierung herangezogen, weil es ebenso zur naturwissenschaftlichen Grundbildung gehört wie das Verständnis von naturwissenschaftlichen Konzepten und Vorgehensweisen (Prenzel, Schütte & Walter, 2007). Die Vorliebe für bestimmte Sachgebiete führt zu einer stärkeren Auseinandersetzung mit bestimmten Inhalten. Die ausdauernde Beschäftigung mit einem Thema wird gefördert und dadurch ein Lernerfolg erreicht, der in einer Art Rückkopplung auf die Interessenentwicklung im Sinne einer Steigerung wirkt (Krapp, 1996). Man spricht in diesem Zusammenhang auch von der epistemischen Orientierung (Prenzel, 1988). Wer sich für eine Sache interessiert, möchte mehr darüber erfahren und sein Wissen in diesem Bereich erweitern. Es werden also zunehmend differenzierte Wissensstrukturen über Lerngegenstände und die mit ihnen realisierbaren Handlungsmöglichkeiten aufgebaut (Krapp, 2000).

Auch wenn der Kausalcharakter des positiven Zusammenhangs zwischen Interesse und Leistung komplex ist, gründet sich die Validierung auf die beschriebenen theoretischen Zusammenhänge. Personen mit einem höher ausgeprägten Interesse an Naturwissenschaften und Forschungsprozessen sollten eine gewisse Kompetenz auf diesem Gebiet erreicht haben und damit eine größere Anzahl an Testitems lösen als Personen mit weniger stark ausgeprägtem Interesse.

Als Indikatoren des Interesses der Schülerinnen und Schüler werden innerhalb der vorliegenden Testentwicklung verschiedene Variablen erhoben. Diese sind das *Schulfachinteresse*, das *Interesse an naturwissenschaftlichen Tätigkeiten* und das *Interesse an naturwissenschaftsbezogenen Aktivitäten*. Sicherlich ist das *Interesse an naturwissenschaftlichen Fächern* nicht mit dem Interesse an den entsprechenden naturwissenschaftlichen Disziplinen gleichzusetzen. Da Schülerinnen und Schüler jedoch dazu neigen, diese beiden Bereiche gleichzusetzen (Prenzel, Schütte & Walter, 2007) und es für sie am leichtesten ist, bezüglich der naturwissenschaftlichen Fächer im Schulkontext zu denken, wurde das Interesse konkret in Bezug auf Schulfächer abgefragt.

Um aussagekräftige Informationen über die Interessenlage der Schülerinnen und Schüler zu erhalten, ist es wichtig, differenzierte und spezialisierte Aspekte des Interesses an Naturwissenschaften und naturwissenschaftlicher Forschung zu erfassen (Prenzel, 1988; Krapp, 2000). Schülerinnen und Schüler können beispielsweise dem Fach Physik negativ gegenüberstehen, während sie jedoch Spaß daran haben, Fragestellungen zu entwickeln oder sich die Umsetzung von Fragestellungen in Versuche zu überlegen. Die Fragen bezüglich bestimmter Aspekte der Naturwissenschaften müssen demnach so formuliert sein, dass die Schülerinnen und Schüler eine klare Vorstellung der Dinge erhalten, die sie beurteilen sollen. Aus diesem Grund erfolgte eine Abfrage des *Interesses an naturwissenschaftlichen Tätigkeiten*. Dieses Vorgehen ermöglicht es, zur Validierung der Testitems eine direkte Verbindung zu dem Inhaltsbereich herzustellen, auf den die Testentwicklung fußt. Haben die Schülerinnen und Schüler großes Interesse an diesen Tätigkeiten, so sollten sie im Sinne der epistemischen Orientierung öfter ausführen oder diesbezüglich eine größere Wissensbasis besitzen als Schülerinnen und Schüler mit geringerem Interesse. Diese Wissensbasis sollte bei den interessierten Schülerinnen und Schülern zu einem höheren Anteil gelöster Testitems führen als bei ihren weniger interessierten Altersgenossen.

Einen weiteren Indikator für naturwissenschaftliches Interesse, der zur Validierung herangezogen werden kann, stellen *naturwissenschaftsbezogene Aktivitäten* dar. Hier wird der Blick darauf gerichtet, wie häufig Schülerinnen und Schüler bestimmte Tätigkeiten ausüben, wie zum Beispiel das Anschauen von Fernsehsendungen über Naturwissenschaften oder aber das Lesen entsprechender Zeitungsartikel. Die Ergebnisse der PISA-Untersuchung 2006 (Prenzel et al., 2008) zeigen einen positiven Zusammenhang zwischen den naturwissenschaftsbezogenen Aktivitäten und den Naturwissenschaftsleistungen. Aus diesem Grund wurde die Variable zur Testvalidierung herangezogen.

Welche Skalen konkret zur Messung der genannten Variablen verwendet wurden,

wird in Kapitel 4 dargestellt.

2.7 DER TEST ZUR ERFASSUNG PROZESSBEZOGENER NATURWISSENSCHAFTLICHER GRUNDBILDUNG

Nachdem alle Theoriebereiche beleuchtet worden sind, welche für die Entwicklung des Tests zur Erfassung prozessbezogener naturwissenschaftlicher Grundbildung von Bedeutung sind, dient dieser letzte Abschnitt einer kurzen Zusammenfassung. Diese erfolgt in Form von Schlussfolgerungen bezüglich der Zielgruppe des Tests und in Gestalt von Anforderungen, die an den Test gestellt werden. Den Abschluss des Theoriekapitels bildet das Rahmenkonzept, das auf einen Blick noch einmal Auskunft über das Konstrukt, die zu messenden Fertigkeiten und die Variablen gibt, die zur Validierung herangezogen werden.

2.7.1 ZIELGRUPPE DES TESTS

Die Zielgruppe und damit auch die Merkmale der Stichprobe, die für die Testentwicklung herangezogen werden, ergeben sich als Resultat der theoretischen Ausführungen.

Die ersten zu beachtenden Merkmale der Zielgruppe ergeben sich daraus, dass der Test schulisch wie auch außerschulisch einsetzbar sein sollte und daher alle Schulformen umfassen sollte. Aus diesem Grund richtet sich der Test zur Erfassung prozessbezogener naturwissenschaftlicher Grundbildung gleichermaßen an Schülerinnen und Schüler der Haupt-, Realschule und des Gymnasiums.

Die Auswahl der Altersgruppe folgt zum Teil aus den Abschnitten über die Entwicklung kognitiver Grundlagen, die zur Ausführung der zu messenden Fertigkeiten vorhanden sein sollten. Die angeführten Befunde zeigen, dass die Schülerinnen und Schüler ein Alter von mindestens zwölf Jahren erreicht haben sollten, um ihre kognitiven Grundlagen für die Planung und Durchführung experimenteller Erkenntnisgewinnung einsetzen zu können. Da es sich hier um ein Mindestalter handelt und es im Hinblick auf Lesefertigkeiten und auf ein gewisses Grundwissen im Bereich der Naturwissenschaften günstiger ist, ältere Schülerinnen und Schüler für die Testentwicklung heranzuziehen, wurden schließlich Schülerinnen und Schüler der neunten Klasse (Altersstufe 15-16 Jahre) ausgewählt. Diese Klassenstufe ist die letzte, die es ermöglicht, Schülerinnen und Schüler aller drei Schulformen zu erfassen.

2.7.2 ANFORDERUNGEN

Kern dieser Arbeit sollte es als Ergebnis der Ausführungen sein, einen ökonomischen Paper-und-Pencil-Test zur Erfassung prozessbezogener naturwissenschaftlicher Grundbildung zu entwickeln und erste Schritte der Validierung zu unternehmen. Der Test sollte in einer Weise konstruiert werden, die sowohl Zustands- als auch Veränderungsmessungen in Bezug auf die genannte Kompetenz ermöglicht. Im Vordergrund steht zunächst ein Einsatz des Instruments als schnelles Screening, also zum Zwecke einer ökonomischen Zustandsmessung von Gruppen. Die Individualdiagnostik soll demnach zunächst im Hintergrund stehen.

Die entwickelten Aufgaben sollten es ermöglichen, die Fertigkeiten der Testpersonen im Hinblick auf das *Identifizieren wissenschaftlicher Hypothesen*, auf das *Planen einer wissenschaftlichen Untersuchung* und auf das *Nutzen wissenschaftlicher Ergebnisse* zu erfassen. Von den manifesten und messbaren Fertigkeiten sollten Rückschlüsse auf zugrunde liegende Fähigkeiten und schließlich auf die zu Grunde liegende Kompetenz möglich sein.

Eine wichtige Anforderung, die an das Testverfahren gestellt wird, besteht darin, dass es in der Durchführung, Auswertung und Interpretation ökonomisch ist und in kurzer Zeit eine valide Aussage ermöglicht. Diese Eigenschaft ist sowohl im schulischen als auch im außerschulischen Einsatz von großer Bedeutung. Die normalen Abläufe sollten hier also möglichst wenig gestört werden.

Die Aufgaben sollten unabhängig voneinander lösbar sein. Auch wenn sie inhaltlich zu einem bestimmten Gebiet der Naturwissenschaften gehören, so darf die Beantwortung einer Aufgabe sich nicht auf die Beantwortung anderer Aufgaben auswirken. Die Lösungswahrscheinlichkeit einer Aufgabe darf nicht durch vorangehende oder folgende Aufgaben beeinflusst werden.

Der *Test zur Erfassung prozessbezogener naturwissenschaftlicher Grundbildung* muss sich der Herausforderung stellen, den Schülerinnen und Schülern alle für die Bearbeitung der Testaufgaben notwendigen Informationen zu vermitteln. Die Beantwortung der Aufgaben soll ein Maß ihrer prozessbezogenen naturwissenschaftlichen Fertigkeiten und nicht ihres Wissens zu einem bestimmten Themengebiet darstellen (Germann, 1989). Dazu gehört auch, dass die in den Aufgaben präsentierten Kontexte möglichst breit gefächert sein und Aspekte aller drei Naturwissenschaften beinhalten sollten, um niemandem Vor- oder Nachteile zu verschaffen, der sich auf einem bestimmten Gebiet besonders gut oder schlecht auskennt.

Die Testpersonen sollten die Aufgaben ohne großen Zeitdruck bearbeiten, um ihre maximale Leistung zeigen zu können. Der Zeitrahmen sollte demnach so gesteckt

sein, dass die Mehrheit der Testpersonen alle Testaufgaben bearbeiten kann. Daraus folgt, dass der Test als Mischung zwischen Speed- und Powertest angelegt sein sollte, in der die Speedkomponente gering gehalten ist. Eine gewisse Zeitbegrenzung ist unvermeidbar, wenn der Einsatz des Instruments ökonomisch bleiben soll.

Die Zielgruppe der Testentwicklung sollten Schülerinnen und Schüler der neunten Klasse darstellen, da in diesem Alter die zur Bearbeitung notwendigen kognitiven Voraussetzungen vorhanden sein sollten. Es werden gleichermaßen Schülerinnen und Schüler der Haupt-, Realschule sowie des Gymnasiums angesprochen. Der neue Test sollte alle Schulniveaus ansprechen, da die durch den Test erfassten Fertigkeiten für Schülerinnen und Schüler aller Schulformen bedeutsam sind.

2.7.3 Abgrenzung des Verfahrens

Um zu verdeutlichen, in welcher Weise sich die vorliegende Testentwicklung von bereits bestehenden kleineren und größeren Verfahren zur Erfassung naturwissenschaftlicher Grundbildung abgrenzen lässt, werden an dieser Stelle die wichtigsten Punkte genannt, die das Besondere dieses Verfahrens ausmachen.

Von den kleineren Verfahren, zu denen der *Test zur Erfassung prozessbezogener naturwissenschaftlicher Grundbildung* gehört, und insbesondere von den aktuellen deutschsprachigen Verfahren (vgl. Tabelle 2.6, S. 62), ist er dadurch abzugrenzen, dass er zum einen zur Testung einer anderen Altersstufe angelegt ist. Zum anderen ist er auf Grundlage der probabilistischen Testtheorie entwickelt, wodurch Aussagen hinsichtlich der internen Validität ermöglicht werden, die so anhand der klassischen Testtheorie nicht möglich sind: Es kann festgestellt werden, inwiefern entwickelte Items das gleiche Konstrukt messen und es gibt eine theoretische Grundlage für den Zusammenhang messbaren Verhaltens und zu Grunde liegender Kompetenz (Hartig, Frey & Jude, 2007). Darüber hinaus unterscheidet sich die vorliegende Testentwicklung dadurch, dass der gesamte Entwicklungsprozess eines Tests, inklusive der Validierung, in umfassender Form abgebildet wird. Das beschriebene Vorgehen bündelt Verfahrensschritte aktueller Testentwicklungsverfahren und kann insofern auch als Grundlage für weitere Testentwicklungen dienen.

Von den größeren Verfahren, wie dem Naturwissenschaftstest der PISA-Untersuchung (OECD, 2007) und anderen internationalen Ansätzen zur Messung naturwissenschaftlicher Grundbildung (American Association of the Advancement of Science, 2001; University of York Science Education Group, 2006), ist die vorliegende Testentwicklung insofern abzugrenzen, als sie sich auf einen speziellen, nämlich den prozessbezogenen Anteil naturwissenschaftlicher Grundbildung, konzentriert. Während

sich die großen Studien einer umfangreichen Erfassung sowohl des konzeptuellen Wissens (naturwissenschaftliches Wissen) als auch des prozessbezogenen Wissens (Wissen über die Naturwissenschaften) widmen, handelt es sich im Falle des vorliegenden Tests um die punktuelle Erfassung eines Bereiches, der sich gemäß der in Abschnitt 2.2.2 dargestellten Befunde als besonders bedeutsam und in Bezug auf weitere Aspekte naturwissenschaftlicher Bildung als einflussreich erwiesen hat. Damit erfordert die *prozessbezogene naturwissenschaftliche Grundbildung* eine spezialisierte Betrachtung, die anhand dieser Arbeit vorgenommen wird. Die Erfassung eines eingegrenzten Bereichs naturwissenschaftlicher Grundbildung ist unter anderem deshalb vorteilhaft, da sie die Prognosekraft des Verfahrens in Bezug auf spezielle Außenkriterien, wie z.B. das experimentelle Arbeiten, erhöht (Rost, 2004a).

2.7.4 RAHMENKONZEPT

Zum Abschluss des Theoriekapitels werden nun noch einmal die wichtigsten Informationen in einem Rahmenkonzept dargestellt. Abbildung 2.10 fasst diese zusammen und schließt das Theoriekapitel ab. Es folgt das Methodenkapitel, das alle statistischen und testtheoretischen Grundlagen für die Operationalisierung, also für die Umsetzung der theoretischen Grundlagen in Testaufgaben, legt.

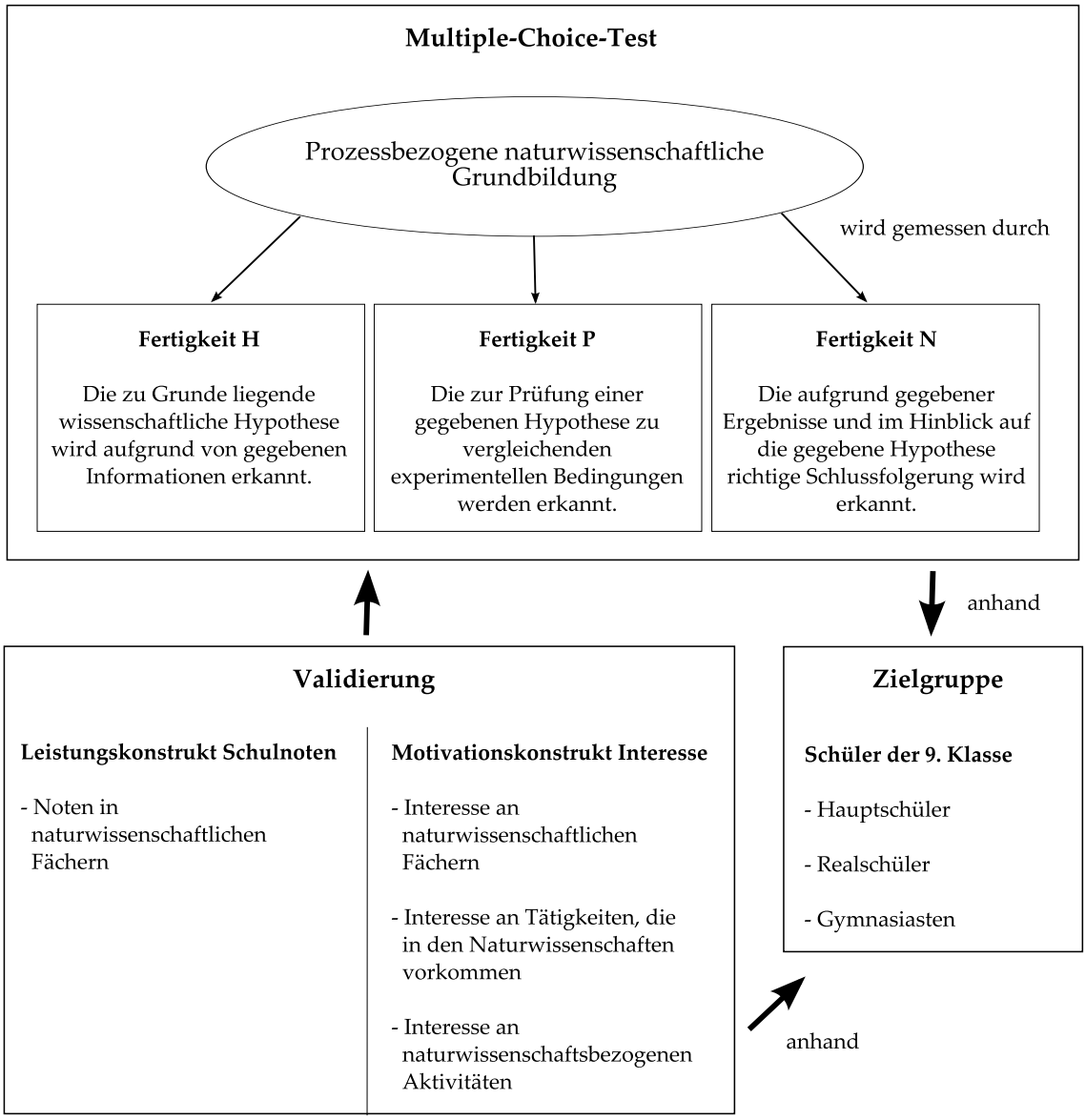


Abbildung 2.10: Rahmenkonzept der Testentwicklung

3 Testtheoretische und verfahrenstheoretische Grundlagen

Dieses Kapitel wird alle test- und verfahrenstheoretischen Informationen liefern, die notwendig sind, um die Methoden der Testentwicklung, der Qualitätsprüfung und der Itemüberarbeitung nachvollziehen und beurteilen zu können.

Ausgehend von den Grundlagen der probabilistischen Testtheorie folgt die Beschreibung der Testmodelle und der Teststatistiken, die der Prüfung der klassischen und probabilistischen Gütekriterien und damit der Entwicklung des Tests dienen. Sie werden am Ende der Testentwicklung darüber Auskunft geben, inwiefern es dem Test gelingt, die prozessbezogene naturwissenschaftliche Grundbildung in objektiver, reliabler und valider Form zu messen.

Es folgen verfahrenstheoretische Ausführungen zum Test- und Antwortformat, zu möglichen logischen Abhängigkeiten sowie Positions- und Reihenfolgeeffekten von Items innerhalb eines Tests und zum Umgang mit fehlenden Werten. Den Abschluss des Kapitels bildet die Beschreibung eines exemplarischer Testentwicklungsansatzes, der den Zweck eines Zwischenfazits verfolgt, indem er auf die theoretischen Grundlagen der Arbeit zurückblickt und auf das Operationalisierungskapitel überleitet.

3.1 Testtheoretische Grundlagen

Lienert (1998, S. 1) definiert einen Test als *„[...] ein wissenschaftliches Routineverfahren zur Untersuchung eines oder mehrerer empirisch abgrenzbarer Persönlichkeitsmerkmale mit dem Ziel einer möglichst quantitativen Aussage über den relativen Grad der individuellen Merkmalsausprägung.“*. Diese Merkmale sind latent und damit keiner direkten Beobachtung zugänglich. Beobachtbar sind lediglich die manifesten Äußerungen der Testpersonen in Gestalt der Itemantworten, die nach der Bearbeitung eines Tests in Form von Datenmustern vorliegen. Die Verbindung zwischen latenten und manifesten Merkmalen wird durch die Testtheorie beschrieben, die Rost (2004a, S. 21) folgendermaßen definiert:

> *„Die Testtheorie beschäftigt sich mit dem Zusammenhang von Testverhalten und dem zu erfassenden psychischen Merkmal.“*.

Um schließlich von Datenmustern auf die Ausprägung der latenten Kompetenz schließen zu können, sind *Testmodelle* nötig. Diejenigen Testmodelle, welche für die vorliegende Testentwicklung von Bedeutung sind, werden in Abschnitt 3.1.2 genauer betrachtet. Im Anschluss daran wird ein Ansatz vorgestellt, der zeigt, auf welche Weise die am Ende festgestellten Kompetenzwerte zu inhaltlich beschreibbaren Kompetenzstufen zusammengefasst werden können.

Zunächst werden jedoch die Besonderheiten der Testtheorie ausgeführt, die dieser Arbeit zu Grunde liegt. Hierbei handelt es sich um die *Item-Response-Theorie*.

3.1.1 ITEM-RESPONSE-THEORIE

Die *probabilistische* oder auch *Item-Response-Theorie (IRT)* wird oft als Alternative zur *klassischen Testtheorie (KTT)* gesehen. Der folgende Abschnitt wird erklären, dass das Verhältnis der Item-Response-Theorie zur klassischen Testtheorie vielmehr als komplementär bezeichnet werden kann (Moosbrugger, 2007).

Die *Klassische Testtheorie* ermöglicht es, die Messgenauigkeit (Reliabilität) zu prüfen und mit Hilfe der Reliabilität und des Messfehlers Konfidenzintervalle für den wahren Wert anzugeben. Auch Fragestellungen der kriterienbezogenen Validität[1] können anhand der KTT beantwortet werden. Wenn es allerdings darum geht, die interne Validität[2] zu prüfen, so liefert die KTT keine geeigneten Antworten.

Als Ergänzung zur Klassischen Testtheorie ist die *Item-Response-Theorie* in der Lage zu klären, welche Rückschlüsse von den Itemantworten der Testpersonen auf das Einstellungs-, Persönlichkeits- oder Fähigkeitsmerkmal gezogen werden können. Um dies zu erreichen, unterscheidet die IRT zwischen zwei Ebenen von Variablen: der latenten und der manifesten Ebene (vgl. Abschnitt 2.2.1). Die Grundannahme der IRT besteht darin, dass die Wahrscheinlichkeit einer bestimmten Antwort einer Testperson auf ein Item im Idealfall als einfache Funktion der für das Konstrukt stehenden latenten Variable (z.B. Ausprägung einer bestimmten Kompetenz) und den Merkmalen des Items (z.B. die Schwierigkeit des Items) gesehen werden kann. Es werden also Annahmen über Zusammenhänge zwischen der individuellen Merkmalsausprägung und der Wahrscheinlichkeit für das Auftreten bestimmter Antworten formuliert. Besteht ein solcher Zusammenhang, so sollten die Antworten auf verschiedene Items zur Messung eines Merkmals hohe Korrelationen aufweisen. Im Falle von Leistungs-

1 Die kriterienbezogene Validität gehört in den Bereich der externen Validierung, in dem es darum geht, vom Testwert erfolgreich auf ein Verhalten außerhalb der Testsituation zu schließen (Beispiel: Schulreifetests).

2 Im Rahmen der internen Validierung wird geprüft, inwiefern vom manifesten Testverhalten auf das latente psychologische Merkmal geschlossen werden kann.

tests sollten Personen mit hoher Ausprägung des latenten Merkmals mehr Items lösen als Personen mit niedriger Ausprägung des Merkmals.

Die starke Korrelation zwischen Itemantworten stellt zwar eine notwendige, jedoch keine hinreichende Bedingung dar, um eine valide Aussage bezüglich des latenten Merkmals treffen zu können. Um als hinreichend gelten zu können, muss zusätzlich die Bedingung der *Itemhomogenität* erfüllt sein. Dies bedeutet, dass das Antwortverhalten der Testpersonen tatsächlich nur von diesem einen latenten Merkmal systematisch beeinflusst wird.

Um wiederum von Itemhomogenität ausgehen zu können, muss die Bedingung der *lokalen stochastischen Unabhängigkeit* erfüllt sein. Diese kann anhand von Korrelationen untersucht werden. Zu diesem Zweck wird das latente Merkmal auf einem bestimmten Wert konstant gehalten und die Korrelationen der Antwortvariablen nur an Personen mit dieser Ausprägung des Merkmals untersucht. Verschwinden die Korrelationen unter dieser Bedingung, so kann von lokaler stochastischer Unabhängigkeit gesprochen werden. Testitems, welche die Bedingung der lokalen stochastischen Unabhängigkeit erfüllen, werden auch als *Indikatoren* der latenten Variablen bezeichnet (Moosbrugger, 2007).

3.1.2 TESTMODELL

Testmodelle dienen der reduzierten Darstellung der Wirklichkeit. Sie sind *„[...]spezielle formale Modelle, die durch die Art der empirischen Daten, auf die sie sich anwenden lassen, definiert sind."* (Rost, 2004a, S.28). Sie lassen sich aus diesem Grund auch nur auf genau diese Daten anwenden.

Bei der Auswahl eines Testmodells aus unterschiedlichen, theoretisch möglichen Modellen ist es wichtig, dasjenige zu wählen, das die Annahmen der inhaltlichen Theorie am besten widerspiegelt. Ein solches Modell muss eine Interpretation der theoretisch angenommenen Beziehung zwischen Testpersonen und Itemantworten und der Abstände unterschiedlicher Itemantworten sowie unterschiedlicher Testpersonen ermöglichen. Um den Abstand zwischen Personen- und Itemparameter beziffern zu können, wird die Wahrscheinlichkeit einer bestimmten Antwort herangezogen, die für unterschiedliche probabilistische Modelle in Form einer Gleichung folgendermaßen ausgedrückt werden kann.

$$\boxed{p(x_{vi}) = f(\theta_v - \sigma_i)}. \tag{3.1}$$

Die Position der Testperson v auf dem latenten Merkmal (Fähigkeit) wird hier durch

θ_v und die Position der Antwort auf ein Item (die Itemschwierigkeit) i durch σ_i repräsentiert. Die Wahrscheinlichkeit einer Antwort $p(x_{vi})$ wird ausgedrückt als Funktion f der Differenz zwischen Fähigkeit der Testperson und Schwierigkeit des Items. Dies kann folgendermaßen interpretiert werden (Wilson, 2003):

- Eine Differenz von 0 bedeutet, dass die Lösung des Items mit einer bestimmten Wahrscheinlichkeit (50%) erfolgt.

- Im Falle einer positiven Differenz erfolgt die Lösung mit größerer Wahrscheinlichkeit.

- Im Falle einer negativen Differenz erfolgt die Lösung mit geringerer Wahrscheinlichkeit.

Ein passendes Testmodell soll durch die Einführung einer latenten Variablen (in diesem Fall der *prozessbezogenen naturwissenschaftlichen Grundbildung*) alle systematischen Zusammenhänge zwischen den Itemantworten der Testpersonen erklären. Das Testmodell, das im Rahmen dieser Testentwicklung Anwendung findet, und dies leisten kann, ist das Rasch-Modell. Es handelt sich um ein Modell mit quantitativer Personenvariable. Jede Testperson wird hinsichtlich ihrer Ausprägung der *prozessbezogenen naturwissenschaftlichen Grundbildung* auf einem Fähigkeitskontinuum verortet. Das Rasch-Modell nimmt genau den in Abbildung 3.1 dargestellten Zusammenhang an, der typisch für Leistungstests ist. Die Abbildung zeigt eine so genannte *Itemcharakteristik (engl.: ICC für Item Characteristic Curve)*. Für jedes Item kann die bereits in Gleichung 3.1 formulierte Wahrscheinlichkeit einer bestimmten Itemantwort in Form einer solchen Kurve dargestellt werden. Hier wird die Lösungswahrscheinlichkeit einer Aufgabe *p(x)* auf der Y-Achse und die Personenfähigkeit, die mit θ bezeichnet wird, auf der X-Achse abgetragen. Mit zunehmender Ausprägung der Personenfähigkeit θ steigt die Wahrscheinlichkeit *p(x)* der Lösung an. Um die Entstehung einer solchen ICC nachzuvollziehen, wird lediglich ein Item unter der künstlichen Annahme betrachtet, dass die wahren Fähigkeitswerte der Testpersonen bekannt sind. Sie verteilen sich über eine Spanne von Fähigkeitsausprägungen auf der Fähigkeitsskala und werden diesbezüglich in unterschiedliche Gruppen eingeteilt. Alle Testpersonen einer Gruppe j besitzen den gleichen Fähigkeitslevel θ_j und es befinden sich m_j Personen in dieser Gruppe. Innerhalb dieser Gruppe beantworten r_j Personen das betrachtete Item richtig. Es folgt, dass auf genanntem Fähigkeitslevel θ_j der Anteil

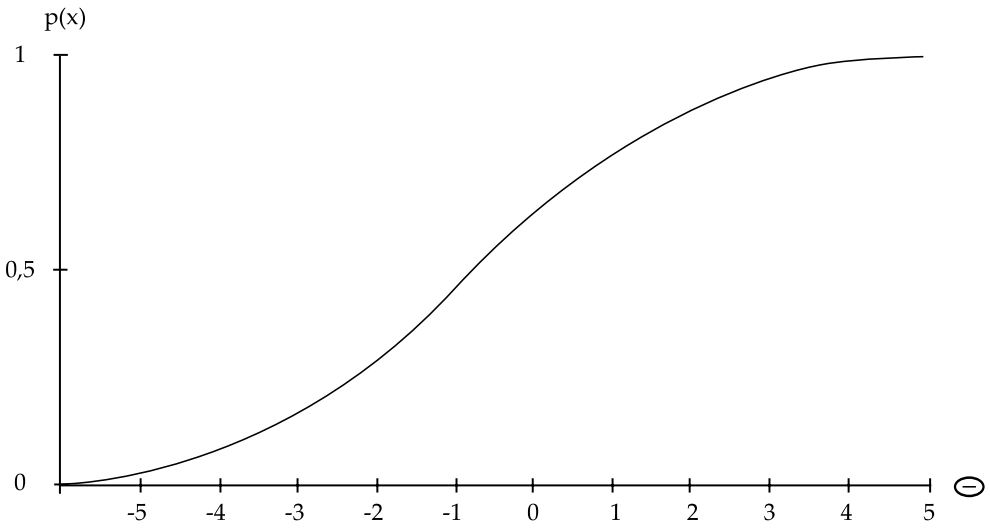


Abbildung 3.1: Itemcharakteristik-Kurve (ICC)

richtiger Antworten folgendermaßen berechnet wird:

$$\boxed{p(\theta_j) = \frac{r_j}{m_j}} \tag{3.2}$$

Die Anzahl von Personen, die das Item richtig beantworten und damit auch die Wahrscheinlichkeit einer richtigen Antwort kann für dieses Item nun für jede Fähigkeitsausprägung berechnet werden. Die beobachteten Anteile richtiger Antworten auf jedem Fähigkeitslevel werden als Funktion der Personenfähigkeit im Koordinatensystem abgetragen.

Aufbauend auf das Konzept der Itemcharakteristik kann im folgenden Abschnitt auf die zentralen Parameter der probabilistischen Testtheorie eingegangen werden, auf die *Itemschwierigkeit* und *-trennschärfe* sowie auf die *Personenfähigkeit*. Im Anschluss an diese allgemeinen Definitionen der Parameter wird vertiefend auf das Rasch-Modell und seine Besonderheiten eingegangen.

ITEM- UND PERSONENPARAMETER

Weder die Item- noch die Personenparameter eines Testmodells sind bekannt. Dies ist dadurch begründet, dass sie sich auf die Populationsebene beziehen, vorliegende Testdaten jedoch einer Stichprobe, also nur einem Teil der Population entstammen. Die Parameter können nicht berechnet werden, sondern müssen aus den in Form einer Datenmatrix zusammengefassten Reaktionen der Probanden auf die Testaufga-

ben geschätzt werden. Die Parameterschätzungen liefern Informationen zu den technischen Eigenschaften der Testitems und zu den Personenmesswerten.

Die Itemparameter, die im Rahmen der Item-Response-Theorie geschätzt werden, sind die Itemschwierigkeit und die Itemtrennschärfe. Die *Schwierigkeit* von Items ist durch die Lage der Itemcharakteristik (ICC) relativ zur X-Achse definiert. Das Item, dessen ICC am weitesten links liegt ist das leichteste und das Item, dessen ICC am weitesten rechts liegt, ist das schwierigste. Um ein solches Item lösen zu können, benötigen die Testpersonen eine höhere Fähigkeit θ als zur Lösung eines weiter links liegenden Items. Per Konvention ist festgelegt, dass der Abszissenwert der 50% Wahrscheinlichkeit die Schwierigkeit des Items definiert. Itemschwierigkeiten um $0,5$ können bei dichotomen Items im Hinblick auf die Maximierung der Varianz und der Trennschärfe als optimal gelten (Kelava & Moosbrugger, 2007). Bei Betrachtung eines Itempools ist es wichtig, dass die Itemschwierigkeiten sich insgesamt so verteilen, dass sie die Fähigkeiten der Testpersonen abbilden können: Es sollte also möglichst zu jeder Fähigkeitsausprägung passende Items geben.

Die *Trennschärfe* drückt die Aussagekraft des Items hinsichtlich unterschiedlicher Eigenschaftsausprägungen der Testpersonen aus. Besitzt ein Item eine hohe Trennschärfe, so vermag es sehr gut zwischen Personen mit unterschiedlichen Eigenschaftsausprägungen zu unterscheiden. Items mit geringer Trennschärfe vermögen dies nur schwer (Rost, 2004a). Im Rahmen der Item-Response-Theorie ist die Itemtrennschärfe in der Itemcharakteristik als Steigung der Kurve in ihrem steilsten Punkt abzulesen. Je steiler der Kurvenanstieg, desto höher die Trennschärfe. Item 2 in Abbildung 3.2 ist demnach trennschärfer als Item 1. Haben Items, wie die hier dargestellten, unterschiedliche Trennschärfen, so überschneiden sich ihre Itemfunktionen. Daraus folgt, dass zwei Items für verschiedene Personen x und y eine unterschiedliche Reihenfolge ihrer Lösungswahrscheinlichkeiten aufweisen können. Abbildung 3.2 zeigt, dass Person x eine höhere Lösungswahrscheinlichkeit für Item 1 als für Item 2 besitzt, wohingegen es sich bei Person y umgekehrt verhält. Im Rahmen des einparametrischen Rasch-Modells, das der vorliegenden Testentwicklung zu Grunde liegt und das im nachfolgenden Abschnitt näher beschrieben wird, spielt die Trennschärfe als zu schätzender Modellparameter keine Rolle, da in diesem Fall von gleichen Trennschärfen der Items ausgegangen wird. Die Homogenität der Trennschärfen stellt allerdings eine im Rahmen der Testentwicklung zu prüfende Voraussetzung dieses speziellen Modells dar.

Der auf Personenseite zu schätzenden Parameter ist die *Personenfähigkeit*. Ziel der Durchführung eines Tests ist es, Personen auf einer Fähigkeitsskala zu verorten. Wird

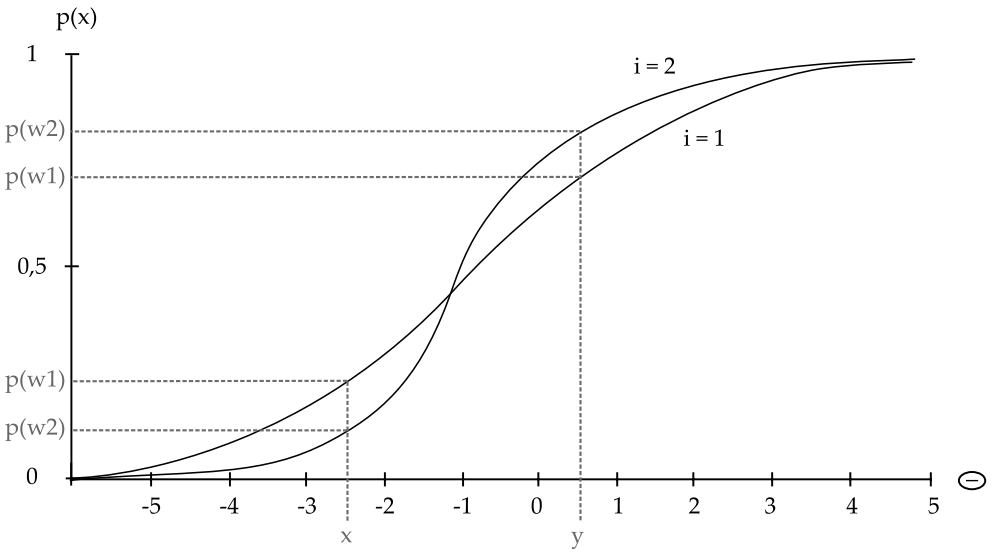


Abbildung 3.2: Itemcharakteristiken von Items unterschiedlicher Trennschärfe

dies für jede Testperson durchgeführt, so können zwei Ziele damit verfolgt werden. Zum einen erhält jede einzelne Person eine Einschätzung ihrer Fähigkeit. Zum anderen können Vergleiche zwischen Testpersonen angestellt werden.

Die nachfolgenden Abschnitte werden zum einen das einfache Rasch-Modell beschreiben und Auskunft darüber geben, auf welche Weise die Parameterschätzungen in diesem Modell erfolgen. Zum anderen werden das mehrdimensionale Raschmodell und das Partial-Credit-Modell beschrieben, da sie im Rahmen der Testvalidierung bzw. zur Schätzung der Parameter mehrstufiger Validierungsvariablen benötigt werden.

DAS RASCH-MODELL

Das dichotome[3] Rasch-Modell, das im Rahmen dieser Testentwicklung zur Anwendung kommt, stellt das einfachste und am weitesten verbreitete Modell dar. Es nimmt für alle Items die gleiche logistische Funktion an und es erklärt das Auftreten einer Datenmatrix, die dichotom bewertete Antworten einer Stichprobe von n Personen auf eine fixe Anzahl von Items enthält, die alle dasselbe latente Merkmal messen. Das Modell enthält folgende Parameter:

- Jede Person besitzt einen Personenparameter θ_v, der die Ausprägung der Person bezüglich des latenten Merkmals markiert.

3 Die Testitems dieses Modells können die Werte 0 für *nicht gelöst* und 1 für *gelöst* annehmen.

- Jedes Item besitzt *einen* Parameter σ_i, der die Schwierigkeit des Items ausdrückt (Molenaar, 1995).

Personen- und Itemparameter werden auf derselben eindimensionalen Skala dargestellt. *Logits* stellen die Maßeinheit zur Bezifferung der Parameter dar.

Die logistische Itemcharakteristik-Funktion zwischen der individuellen Merkmalsausprägung θ_v und der Lösungswahrscheinlichkeit eines Items i in Abhängigkeit von dessen Schwierigkeit σ_i sieht folgendermaßen aus:

$$\boxed{p(x_{vi} = 1) = \frac{exp(\theta_v - \sigma_i)}{1 + exp(\theta_v - \sigma_i)}} \tag{3.3}$$

Die Gleichung stellt die Lösungswahrscheinlichkeit eines Items i durch eine Person v dar. Entscheidend für diese Lösungswahrscheinlichkeit ist die Differenz zwischen der individuellen Personenfähigkeit θ_v und der Schwierigkeit σ des jeweiligen Items i. Sind Personenfähigkeit und Itemschwierigkeit gleich groß, so liegt die Lösungswahrscheinlichkeit bei 50%.

Das Modell enthält nur einen Itemparameter, den Schwierigkeitsparameter, und wird deshalb auch als Einparameter-Logistisches (1PL)-Modell bezeichnet. Bezüglich des nicht berücksichtigten Trennschärfeparameters macht das Modell die Annahme, dass er für alle Items gleich ist (*Itemhomogenität*). Erweisen sich Items hinsichtlich ihrer Trennschärfe als nicht homogen, so wird dies so interpretiert, dass diese Items ein anderes Merkmal messen als eigentlich intendiert und somit aus dem Itempool entfernt oder überarbeitet werden sollten (Rost, 2004a).

Die Feststellung abweichender Trennschärfen wird getroffen, indem die einzelnen Itemcharakteristik-Kurven (ICCs) betrachtet werden. Da sich die Trennschärfen als Steigung der Funktionskurve in ihrem steilsten Punkt zeigen und diese laut genannter Voraussetzung gleich sein sollten, müssten die Kurven der Items parallel verlaufen, wenn diese das gleiche Merkmal messen. Somit sollten die Kurven lediglich parallel bezüglich der X-Achse verschoben sein (s. Abbildung 3.3). Daraus folgt, dass die Items für jeden Fähigkeitslevel der Testpersonen die gleiche relative Reihenfolge besitzen. Die Parallelität der Itemfunktionen stellt ein bedeutsames Merkmal des Rasch-Modells dar. Wenn das Modell gilt, so besitzen alle Items dieselbe Trennschärfe (Rost, 2004a). Weichen beobachtete ICCs von den anhand des Modells zu erwartenden ICCs ab und überschneiden sich Itemfunktionen, so besitzen die betrachteten Items also nicht die gleiche Trennschärfe.

Einen mit der Überprüfung dieser Modellvoraussetzungen nicht zu verwechselnden Punkt stellt die Betrachtung der klassischen Itemtrennschärfen im Laufe der Test-

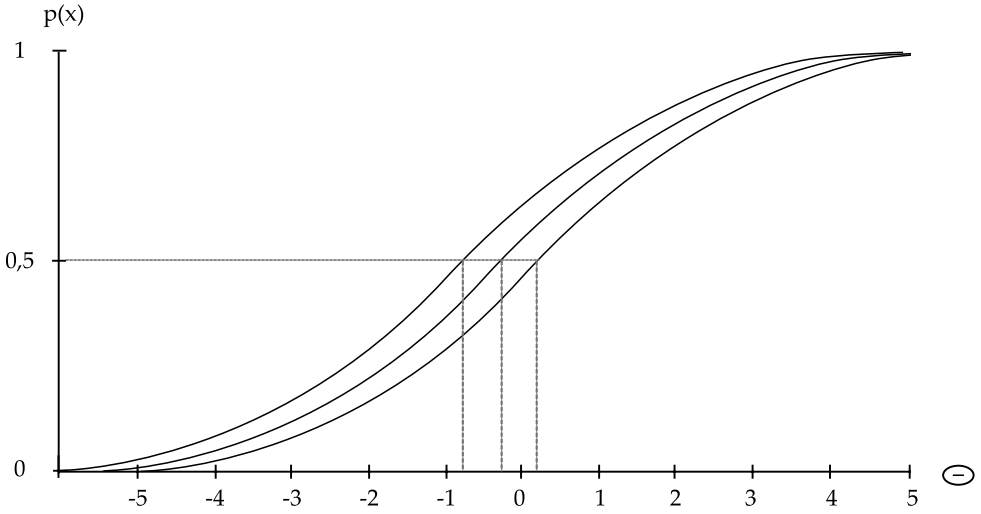


Abbildung 3.3: Parallel verschobene Itemcharakteristik-Kurven im Rasch-Modell

entwicklung dar, die der Prüfung der Itemqualität dient und damit zusätzliche Informationen für die Entscheidung bietet, inwiefern Items überarbeitet oder entfernt werden sollten. Die klassische Trennschärfe ist definiert als die Korrelation eines Items i mit dem Testergebnis t (Rost, 2004a):

$$\boxed{r_{it} = Korr(\hat{\theta}_v, x_{vi})}. \tag{3.4}$$

Die Werte von Itemtrennschärfen sollten zwischen $0,4$ und $0,7$ liegen, um als gut gelten zu können (Kelava & Moosbrugger, 2007).

Ein weiteres Merkmal des Rasch-Modells ist die *spezifische Objektivität*. Darunter wird in Bezug auf psychometrische Tests verstanden, dass die Differenzen der mit dem Rasch-Modell geschätzten Personenparameter von der Stichprobe der Items unabhängig sind, anhand derer diese Merkmale erfasst wurden. Dies bezieht sich darauf, dass die Items, die einen Test bilden, lediglich eine begrenzte Auswahl aus einem hypothetischen Item-Universum darstellen und das Testergebnis nicht allein etwas über die Fähigkeit zur Beantwortung dieser Items aussagen soll. Die Eigenschaftsmessung bezieht sich immer auf ein ganzes Itemuniversum und es sollte sich unabhängig von der Itemauswahl immer dasselbe Testergebnis zeigen. Analog zur Argumentation bezüglich der Personenparameter sind die Werte für die Differenzen der Itemparameter von der Stichprobe der Testpersonen unabhängig (Rost, 2004a). Die Eigenschaft der spezifischen Objektivität steht in enger Verbindung mit einer weiteren wichtigen Eigenschaft des Rasch-Modells, der *Separabilität*. Darunter ist zu verstehen, dass Personen- und Itemparameter getrennt voneinander geschätzt werden

können.

Im Rasch-Modell erfolgt die Schätzung der Personen- und Aufgabenparameter nach dem *Maximum-Likelihood-Prinzip*. Die Likelihood-Funktion beschreibt die Wahrscheinlichkeit der Daten (aller Daten der Testdatenmatrix) unter der Annahme, dass ein bestimmtes Modell gilt. Vereinfacht ausgedrückt, werden die Parameter so geschätzt, dass die Wahrscheinlichkeit der Daten unter der Bedingung des Testmodells möglichst groß wird (daher Maximum-Likelihood). Werden verschiedene Testmodelle miteinander verglichen, so ist das Modell vorzuziehen, unter dem dieselben Testdaten eine höhere Wahrscheinlichkeit besitzen und das sich hinsichtlich der Anzahl der geschätzten Parameter als sparsamer erweist (Rost, 2004a).

Die Eigenschaften des Rasch-Modells erlauben eine Schätzung der Parameter auf der Grundlage der Zeilen- und Spaltensummen der Rohdatenmatrix. Dies bezeichnet man als *suffiziente Statistiken*. Die Schwierigkeit eines Items wird demnach dadurch bestimmt, *wie viele* Personen es gelöst haben und nicht *welche*. Entsprechend wird die Fähigkeit einer Person dadurch bestimmt, *wie viele* Items sie gelöst hat und nicht *welche*.

Um die Personenparameter schätzen zu können, werden zunächst die Itemparameter geschätzt und danach fixiert. Zur Schätzung der zu Grunde liegenden Personenfähigkeit werden diese Werte zusammen mit den Antworten der Testpersonen auf die einzelnen Testitems benötigt. Die Antworten der Testpersonen auf die Items liegen dabei als Vektoren vor, die beim dichotomen Rasch-Modell aus den Ziffern 1 (Item gelöst) und 0 (Item nicht gelöst) bestehen. Die Schätzung der Personenfähigkeit erfolgt ebenso wie im Fall der Itemparameter nach dem Maximum-Likelihood-Prinzip.

Für die Schätzung der Personenparameter stehen unterschiedliche Schätzer zur Verfügung. Ein Auswahlkriterium für den Schätzer besteht darin, diejenige Funktion zur Berechnung zu wählen, deren Erwartungswert dem wahren Wert des Parameters möglichst nahe kommt. Dieses Kriterium erfüllt der *Weighted Likelihood Estimator (WLE)* besonders gut. Simulationsstudien zeigen, dass er im Vergleich zu anderen gebräuchlichen Schätzern die geringsten Verzerrungen aufweist. Darüber hinaus besitzt er die vorteilhafte Eigenschaft, dass er auch Messwerte für Personen liefert, die kein Item oder alle Items gelöst haben. Die nach der WLE-Methode geschätzten Parameter weisen die besten psychometrischen Eigenschaften auf, sie besitzen Intervallskalen-Niveau und stellen die besten Punktschätzer der individuellen Messwerte dar. Sie treffen also die Fähigkeitsausprägung für eine einzelne Person am besten (Rost, 2004a) und werden in dieser Arbeit daher zur Personenparameterschätzung herangezogen. Die WLEs werden anhand des Programms Conquest (Wu, 1997)

berechnet. Die Formel zur Schätzung des WLE sowie vertiefende Informationen zu seinen Eigenschaften finden sich bei Walter (2005).

Aufgrund der genannten wünschenswerten Eigenschaften wird das 1PL-Modell als günstiges Modell eingestuft (Moosbrugger, 2007). Nur wenn das angenommene Rasch-Modell gilt, so besteht auch die angenommene Beziehung zwischen erreichtem Summenscore der Personen und der latenten Variable. Lediglich in diesem Fall ist der Summenscore aussagekräftig hinsichtlich der zu Grunde liegenden Fähigkeitsausprägung einer Person (Rost, 2004a).

DAS MEHRDIMENSIONALE RASCH-MODELL

Das mehrdimensionale Rasch-Modell wird im Rahmen der Testentwicklung benötigt, da die zu messende Kompetenz den theoretischen Ausführungen zufolge als eindimensional angenommen wird, aber anhand dreier Fertigkeiten operationalisiert wurde. Somit besteht die Notwendigkeit zu prüfen, inwiefern sich die anhand des Tests erhobenen Daten angemessener durch ein eindimensionales oder ein dreidimensionales Modell erklären lassen. Auf die Modellprüfung wird später in diesem Kapitel eingegangen. An dieser Stelle wird zunächst das zur Schätzung der Parameter benötigte mehrdimensionale Modell beschrieben.

Das mehrdimensionale Rasch-Modell geht von mehreren latenten Personeneigenschaften aus. Gemäß Gleichung 3.5 (Lösungswahrscheinlichkeit für das Item i) formuliert es die Annahme, dass die Wahrscheinlichkeit einer Antwort einerseits durch eine gewichtete Summe der loglinearen Personen-Dimensions-Parameter vj und andererseits durch einen Itemparameter festgelegt ist. Die Gewichte q_ij stammen aus einer Designmatrix, in der festgelegt wird, mit welchem Gewicht eine Eigenschaftsdimension j an der Antwort auf ein Item i beteiligt ist, und zwar für alle Personen v in gleicher Weise. Insofern ist die Designmatrix vergleichbar mit der Ladungsmatrix von Faktorenanalysen, die angibt, wie hoch die Items auf einzelnen Faktoren laden.

$$p(x_{vi} = 1) = \frac{exp(\sum_{j=1}^{h} q_{ij}\theta_{vj} - \sigma_i)}{1 + exp(\sum_{j=1}^{h} q_{ij}\theta_{vj} - \sigma_i)} \tag{3.5}$$

In der dargestellten Gleichung indiziert v eine Person, i ein Item und j eine Dimension. q_{ij} steht für die Einträge in der Designmatrix, θ für die Personenfähigkeit und σ für die Itemschwierigkeit. Für jede Person ist in dem Modell ein eigener Parameter

vj vorgesehen.

Wang (1995) und Adams, Wilson und Wang (1997) beschreiben zwei Typen multidimensionaler Tests: er *between-item*-Test findet hier Anwendung. Es werden Gruppen von Items gebildet, die exklusiv unterschiedliche latente Variablen erfassen. Im Falle dieser Arbeit werden die Testitems gemäß der Fertigkeit zusammengefasst zu deren Messung sie konzipiert wurden. Es entstehen also drei Itemgruppen.

DAS PARTIAL-CREDIT-MODELL

Das *Partial-Credit-Modell* ist ein Rasch-Modell für ordinale Daten. Es wird an dieser Stelle deshalb eingeführt, da die Kriterien, die im Rahmen dieser Testentwicklung zur Validierung erhoben wurden, in Form eines über die zwei Kategorien des einfachen Rasch-Modells hinausgehenden Ratings erfasst wurden. Bei mehr als zwei Antwortkategorien reicht das einfache Rasch-Modell allerdings nicht mehr aus. Das Partial-Credit-Modell stellt eine Erweiterung des einfachen Rasch-Modells zum Zweck der Skalierung mehrfach abgestufter Antwortformate dar (Masters, 1982). Konkret wird dieses Modell zur Skalierung der ordinalen Items der Interessenskalen verwendet, die als Validitätskriterien für den Kompetenztest verwendet werden 3.2.1.

Im Fall des Partial-Credit-Modells werden die Wahrscheinlichkeiten, dass die Antwort einer Testperson in eine der Antwortkategorien fällt, als *Kategorienwahrscheinlichkeit* bezeichnet und in Beziehung zum latenten Merkmal gesetzt. Die aus den Kategorienwahrscheinlichkeiten resultierenden Funktionen werden als *Kategorienfunktion* bezeichnet. Es liegen dabei so viele Kategorienfunktionen vor wie es Antwortkategorien gibt. Für jeden Übergang einer Kategorie zu einer folgenden Kategorie kann in Abhängigkeit vom latenten Merkmal eine Schwellenwahrscheinlichkeit, also ein Schwellenparameter geschätzt werden. Die Schwellenwahrscheinlichkeit ist definiert als die Wahrscheinlichkeit, dass sich eine Person für die höhere von zwei Antwortkategorien entscheidet unter der Bedingung, dass die Antwort in einer der beiden Kategorien liegt.

Die logistische Funktion, die diese Schwellenwahrscheinlichkeiten in Abhängigkeit vom latenten Merkmal modelliert, sieht in der allgemeinsten Form nach Masters (1982) folgendermaßen aus:

$$p(X_{vi} = x_{vi}|\theta_v, \sigma_{ix}) = \frac{exp(x_{vi}\theta_v - \sigma_{ix})}{1 + \sum_{s=1}^{m} exp(s\theta_v - \sigma_{is})} \tag{3.6}$$

In dem Modell werden die einzelnen Schwellenwahrscheinlichkeiten (die Wahrscheinlichkeit, dass Person v bei Item i Schwelle x überschreitet) zusammengefasst, um die Kategorienwahrscheinlichkeit zu berechnen. θ_v stellt den Fähigkeitsparameter von Person v dar, σ_{ix} die Schwierigkeit, bei Item i Schwelle x zu überschreiten.

Die Skalierung der Interessensskalen auf die beschriebene Weise ermöglicht die Prüfung der externen Validität durch eine Korrelation dieser Skalen mit der anhand des dichotomen Rasch-Modells skalierten Kompetenz in Abschnitt 5.5.1.

3.1.3 KOMPETENZSTUFEN

Einen weiteren Punkt, der im Rahmen der test- und verfahrenstheoretischen Grundlagen zu beachten ist, stellt die Einteilung gemessener Kompetenzen in Kompetenzstufen dar.

Um Personen hinsichtlich ihrer Merkmalsausprägung unterscheiden und die einzelnen Ausprägungen inhaltlich beschreiben zu können, hat es sich als sinnvoll erwiesen, die kontinuierliche Skala der gemessenen Kompetenz in Niveau-Abschnitte zu unterteilen (Hartig & Klieme, 2006). Eine Einteilung in Niveaus ist hier notwendig, da eine inhaltliche Beschreibung jedes einzelnen Punktwertes in der Praxis nicht realisierbar ist (Beaton & Allen, 1992). Innerhalb der einzelnen Niveaus wird demnach keine weitere Differenzierung vorgenommen, sondern die Skalenabschnitte werden als Ganzes kriteriumsorientiert beschrieben. Hierzu bezieht man sich auf die konkreten Inhalte der Aufgaben des jeweiligen Abschnitts.

Die entscheidende Frage bei der Definition von Kompetenzniveaus ist jedoch, wo die Grenzen zwischen den Niveaus gezogen werden, welche Strategie also für die Festlegung der Schwellen gewählt wird. Hier sind unterschiedlich stark modellgeleitete Vorgehen denkbar.

Der am wenigsten modellgeleitete Ansatz besteht darin, die Grenzen relativ willkürlich in etwa gleichen Abständen zu setzen. Im Anschluss daran wird nach Aufgaben gesucht, deren Schwierigkeiten für die gesetzten Schwellen charakteristisch sind. Erst danach können die Skalenniveaus anhand der ausgewählten Aufgaben inhaltlich beschrieben werden.

Ein stark modellgeleitetes Vorgehen würde im Gegensatz dazu so aussehen, dass ein genaues Anforderungsprofil für jede Aufgabe erstellt wird, um die Schwierigkeiten der Aufgaben vorherzusagen. Diese Annahmen werden anhand der empirisch ermittelten Aufgabenschwierigkeiten in Regressionsanalysen überprüft, um diejenigen Merkmale auszuwählen, welche die stärkste Erklärungskraft für die Unterschiede in den Aufgabenschwierigkeiten haben. Aufgrund der Regressionsanalysen kann

im Vorfeld antizipiert werden, welche Schwierigkeiten Aufgaben mit bestimmten Anforderungskombinationen besitzen. Diese erwarteten Schwierigkeiten werden genutzt, um Abschnitte auf der Skala zu definieren. Die Kompetenzniveaus werden hier nicht mehr durch konkrete Aufgaben definiert, sondern durch generalisierbare Anforderungen (Hartig & Klieme, 2006). Ein solches Vorgehen erfordert, dass sich die postulierten Schwierigkeitsunterschiede zwischen Aufgaben auch nachweisen lassen. Idealerweise sollten sich die Items der verschiedenen Kompetenzstufen entlang des gemessenen Kontinuums in Gruppen anordnen lassen: Alle Items einer höheren Kompetenzstufe müssen dann allerdings auch tatsächlich schwerer sein als die Items der niedrigeren Kompetenzstufen. Dieses Vorgehen hat sich unter anderem im Naturwissenschaftstest der PISA-Untersuchung als nicht günstig erwiesen, da es immer wieder Aufgaben gab, die sich in einem anderen Schwierigkeitssegment befanden, als zunächst angenommen (Rost, 2004b).

Ein alternativer Weg der Bestimmung von Kompetenzstufen besteht darin, sie über die von den Testpersonen produzierten Aufgabenlösungen zu definieren. Dies setzt allerdings voraus, dass die Antworten nicht nur dichotom, sondern mehrstufig erfasst werden. In diesem Fall werden Aufgaben *nicht gelöst*, *teilweise gelöst* oder *vollständig gelöst*. Ein Schüler, der Aufgaben jeweils nur zum Teil lösen kann, würde dann in den mittleren Kompetenzbereich fallen. Eine solche Konzeption von Kompetenzstufen erfordert ein Testmodell, das mit ordinalen Itemantworten umgehen kann, wie z.B. das bereits beschriebene *Partial-Credit*-Modell.

In dieser Arbeit soll nach Abschluss der Testentwicklung das als pragmatisch einzustufende Vorgehen gewählt werden, das Fähigkeitskontinuum in etwa gleiche Abstände zu zerlegen und im Anschluss daran Items zu finden, die für die gesetzten Schwellen charakteristisch sind. Die Umsetzung sollte jedoch erst nach einer ausreichenden Validierung erfolgen.

3.2 GÜTEKRITERIEN UND IHRE PRÜFUNG

Das Ziel der hier dargestellten Testentwicklung besteht darin, ein Verfahren zu entwickeln, das misst, was es zu messen beansprucht, das genau misst und dessen Ergebnis unabhängig von der Person ist, die den Test durchführt, auswertet oder interpretiert. Um das Erreichen dieses Ziels sicher zu stellen, ist es notwendig, die sogenannten Hauptgütekriterien zu betrachten und zu prüfen. Dieses sind die Validität, die Reliabilität und die Objektivität.

3.2.1 VALIDIERUNG

Die Validität stellt das zentrale Gütekriterium für die Prüfung der Qualität eines Tests dar. Sie prüft, ob ein Test tatsächlich das Merkmal misst, was er zu messen beansprucht. Ist dies nicht der Fall, misst er also etwas anderes, so sind nicht nur die übrigen Haupt-Gütekriterien Reliabilität und Objektivität nicht mehr von Belang, der Test kann darüber hinaus auch nicht eingesetzt werden. Die Prüfung der Validität stellt kein triviales Unterfangen dar. Es ist das komplexeste und am schwierigsten zu bestimmende Gütekriterium (Hartig et al., 2007), da sich die Messung in der Regel auf ein nicht direkt beobachtbares Konstrukt bezieht, im Falle dieses Tests auf die *prozessbezogene naturwissenschaftliche Grundbildung*.

So müssen vor allem die Kriterien, die zur Prüfung der Validität herangezogen werden sollen, sorgfältig ausgewählt und theoretisch begründet werden. Die begründete Kriterienauswahl ist bereits im Abschnitt 2.6 erfolgt. Der folgende Abschnitt widmet sich der Beschreibung der Bedeutung und der gewählten Zugänge und Methoden der Validierung. Die Prüfung der Validität teilt sich in *interne* und *externe Validierung* (Lienert & Raatz, 1998; Rost, 2004a).

Im Rahmen der *internen Validierung* wird die Gültigkeit des zu Grunde liegenden Modells, in diesem Falle des einfachen Rasch-Modells geprüft. Sie umfasst die Prüfung des Infits[4], die Prüfung von Gruppenunterschieden, die Prüfung der Personenhomogenität (Analyse des Differential-Item-Functioning (DIF)) sowie die globale Modellprüfung zur Feststellung des angemessenen Testmodells.

Die *externe Validierung* setzt die gemessene Kompetenz mit Außenkriterien in Beziehung. Anhand des Leistungskonstrukts *Schulnoten* wird die konvergente Validität und anhand des Motivationskonstrukts *Interesse* die diskriminante Validität geprüft.

In Bezug auf die Validierung ist allgemein stets zu beachten, dass es sich um einen Prozess handelt, in dessen Verlauf sich die Kriterien, die zur Validierung herangezogen werden, immer weiter verfeinern oder gegen angemessenere Kriterien ausgetauscht werden. So wird im Rahmen dieser Testentwicklung absichtlich von der *Validierung* oder *Prüfung der Validität* als Tätigkeit und Prozess gesprochen und nicht von *der* Validität als abgeschlossenem Zustand.

4 Darunter wird die Prüfung der Itemhomogenität anhand eines Tests der Voraussetzung gleicher Trennschärfen verstanden, die für das einfache Rasch-Modell gelten muss.

INTERNE VALIDIERUNG

PRÜFUNG DES INFITS

Die Prüfung des Infit-Maßes (*Mean Squares*) der Items erfolgt anhand einer Prüfgröße T. Sie dient als Indikator für abweichende Trennschärfen, für zu große Differenzen zwischen beobachteten und gemäß dem angenommenen Modell zu erwartenden Werten für bestimmte Items. Genauer wird hier geprüft, wie sehr diese Abweichungen variieren im Vergleich dazu, wie sehr sie variieren dürften, wenn das angenommene Modell gilt. Bei der Betrachtung des T-Wertes werden im Allgemeinen Werte über $2,0$ und unter $-2,0$ (Wright, Mead & Bell, o. J.; R. M. Smith, 1995) als signifikant betrachtet. Zu beachten ist allerdings, dass diese Statistik abhängig von der Stichprobengröße ist, so dass große Stichproben leichter zu signifikanten Ergebnissen führen (Rost, 2004a; Wilson, 2005). Aus diesem Grund werden diese Grenzen zwar als Anhaltspunkt genommen, es wird allerdings zunächst allgemein ermittelt, welche Items stark abweichende T-Werte erkennen lassen. Werte < -2 können ein Anzeichen für besonders hohe Trennschärfen, Werte > 2 ein Anzeichen für zu geringe Trennschärfen sein. Da hoch trennscharfe Items unter anderem zu einer besseren Reliabilität führen als niedrig trennscharfe Items, sind signifikante Abweichungen im negativen Bereich eher tolerierbar als Abweichungen im positiven Bereich.

Da der Zusammenhang zwischen T-Wert und Trennschärfe nicht zwangsläufig besteht und um sicher klären zu können, ob eine nicht modellkonforme Abweichung der Trennschärfe vorliegt, schließt sich der Feststellung eines auffälligen T-Wertes zunächst eine Betrachtung der *Mean Squares* an. Optimale Werte des *Mean Squares* liegen nahe bei 1. In diesem Fall variieren die beobachteten Abweichungen wie erwartet, d.h. gemäß Gleichung 3.7 sind beobachtete und erwartete Abweichung gleich.

$$\boxed{MNSQ_i = \frac{beob.Abweichung}{erw.Abweichung}} \tag{3.7}$$

Abweichungen bedeuten, dass das angenommene Modell nicht zum Testverhalten passt. Werte > 1 sind hier ein Zeichen für eine unerwartet hohe Varianz der Abweichungen, die darauf hindeutet, dass das Item die zu Grunde liegende Kompetenz nicht ausreichend gut bzw. etwas anderes erfasst. Werte < 1 deuten auf eine unerwartet geringe Varianz hin, die dafür spricht, dass das Item in einem relativ kleinen Fähigkeitsbereich sehr gut diskriminiert (Wilson, 2005). Als Grenzen für akzeptable *Mean Squares* schlagen Adams und Khoo (1996) $0,75$ und $1,33$ vor.

Um sicher klären zu können, ob wirklich die Trennschärfe für die Abweichung eines Items verantwortlich ist, müssen in einem abschließenden Schritt die *Item-Characteristic-Kurven (ICC)* der jeweiligen Items geprüft werden, in der sich die Trennschärfe als Steigung der Kurve an ihrer steilsten Stelle ablesen lässt. Hier sollten die beobachteten *ICCs* den erwarteten *ICCs* sehr ähnlich sein. Es sind vor allem die Items kritisch zu beachten, deren beobachtete *ICC* eine deutlich geringere Steigung aufweisen als die erwartete *ICC*, da dies ein Zeichen für eine zu geringe Trennschärfe ist. Sind in diesem Sinne große Abweichungen zu erkennen, so kann die Schlussfolgerung gezogen werden, dass das jeweilige Item der in Abschnitt 3.1.2 beschriebenen Modellvoraussetzung gleicher Trennschärfen widerspricht. Die Konsequenzen dieser Feststellung können in der Eliminierung oder der Überarbeitung des Items liegen. Wilson (2003) zufolge sollte bei einer theoretisch fundierten Testentwicklung mit einem zu Grunde liegenden Rahmenkonzept zunächst die Überarbeitung der Items den Vorzug erhalten.

PRÜFUNG VON GRUPPENUNTERSCHIEDEN

Die Prüfung von Gruppenunterschieden wird durchgeführt, um Hypothesen zur Ausprägung der *prozessbezogenen naturwissenschaftlichen Grundbildung* in definierten Teilstichproben prüfen zu können. Zu beachten ist, dass die Ergebnisse dieser Prüfungen nur schwache Hinweise auf die Validität des Verfahrens liefern können, da die postulierten Gruppenunterschiede zwar eine notwendige, aber keine hinreichende Bedingung darstellen, um von Validität sprechen zu können.

Die Prüfung auf signifikante Testwertunterschiede zwischen Schülerinnen und Schülern erfolgt anhand eines einseitigen t-Tests, während die Testwertunterschiede zwischen den einzelnen Schulniveaus durch eine einfaktorielle Varianzanalyse mit Einzelvergleichen geprüft wird.

T-Tests werden allgemein herangezogen, wenn *zwei* abhängige oder unabhängige Gruppen bezüglich eines Unterschieds ihrer Mittelwerte eines intervallskalierten Merkmals untersucht werden. Im Fall des Mittelwertvergleichs zwischen Schülerinnen und Schülern hinsichtlich ihrer Testleistung kommt der t-Test zum Vergleich unabhängiger Stichproben zum Einsatz. Der Unterschied der Mittelwerte wird auf Signifikanz geprüft, indem der empirisch berechnete t-Wert mit dem kritischen t-Wert verglichen wird. Dieser richtet sich nach der Anzahl der Freiheitsgrade[5] und dem

5 Freiheitsgrade stellen die Anzahl der bei der Berechnung eines Kennwertes frei variierbaren Werte dar. Beispiel: Die Summe der Differenzen aller Werte von ihrem Mittelwert ergibt 0. Sind also von n Werten bereits n-1 Werte zufällig gewählt, so steht die Größe des letzten Wertes fest. Die Varianz,

Alpha-Niveau, das als Irrtumswahrscheinlichkeit eingeräumt wird (Bortz, 1999).

Anhand von Varianzanalyse ist es möglich Mittelwertunterschiede in $n \geq 2$ Gruppen auf Signifikanz zu prüfen. In Falle des Schulniveauvergleichs wird die einfaktorielle Varianzanalyse durchgeführt: Die Variable *prozessbezogene naturwissenschaftliche Grundbildung* wird in Abhängigkeit *einer* unabhängigen Variable (Schulniveau) untersucht. Diese ist dreifach gestuft, da sie aus Haupt-, Realschule und Gymnasium besteht. Die Varianzanalyse prüft, inwiefern die Stichprobenunterschiede (die Varianz der Stichprobe hinsichtlich der gemessenen Kompetenz) durch den systematischen und überzufälligen Einfluss der Variable Schulniveau erklärt werden kann. Zur Feststellung der Signifikanz wird der empirisch ermittelte F-Wert mit dem kritischen F-Wert verglichen, der sich ebenfalls nach der Anzahl der Freiheitsgrade und dem Alpha-Niveau richtet.

Um im Rahmen der Varianzanalyse nicht nur Aussagen darüber treffen zu können, *dass* zwischen den Schulniveaus signifikante Mittelwertunterschiede bestehen, sondern auch, *welche* der Mittelwertunterschiede sich signifikant unterscheiden, werden Einzelvergleiche, so genannte *Kontraste*, geprüft, die theoriegeleitet vor der Durchführung der Analyse festgelegt werden (vgl.Bortz, 1999).

Die Voraussetzungen für den Einsatz des t-Tests sind Normalverteilung sowie Varianzhomogenität der Daten. Er ist gegenüber Verletzungen dieser Voraussetzungen relativ robust. Allerdings nimmt die Wahrscheinlichkeit normalverteilter Daten mit steigender Stichprobengröße zu und die Verletzungen der Voraussetzungen sind insbesondere dann von geringem Einfluss, wenn die Stichproben entweder gleich groß oder aber die Varianzen gleich sind. Die Voraussetzungen für die Durchführung der Varianzanalyse bestehen in der Normalverteilung und der Varianzhomogenität der Fehlervarianzen. Sie ist gegenüber Verletzungen ihrer Voraussetzungen nicht so robust wie der t-Test. Eine Verletzung der Varianzhomogenität ist dann tolerierbar, wenn die Stichproben gleich groß und umfangreich genug sind (Bortz, 1999).

PRÜFUNG DER PERSONENHOMOGENITÄT: DIF-ANALYSEN

Als *Differential Item Functioning (DIF)* wird der Umstand bezeichnet, dass ein Item in unterschiedlichen Gruppen unterschiedliche statistische Eigenschaften zeigt, also für Personen einer bestimmten Gruppe schwerer ist als für Personen einer anderen Gruppe. Die Feststellung von DIFs stellt im Sinne der Item-Response-Theorie einen Hinweis auf Items dar, die etwas anderes messen als intendiert und ist damit eine

deren Formel Mittelwertsdifferenzen enthält, hat daher n-1 Freiheitsgrade. (Bortz, 1999)

Möglichkeit zur Prüfung der internen Validität. Die Analyse erfolgt anhand der nach Geschlecht und nach Schulniveau gebildeten Teilstichproben in drei Schritten:

- *Schritt 1*: Anhand von graphischen Darstellungen, auch *graphischer Modelltest* genannt (vgl.Rost, 2004a), werden die Itemschwierigkeiten von jeweils zwei Stichproben verglichen. Dazu werden die Itemschwierigkeiten einzeln für die zu vergleichenden Untergruppen geschätzt und in einem Koordinatensystem abgetragen. Jedes Item wird demnach durch die zwei in den Untergruppen vorgenommenen Schätzungen lokalisiert. Stimmen die Schätzwerte der beiden Gruppen perfekt überein, so müssten sie auf einer linearen Geraden liegen. Je unterschiedlicher die Schätzwerte des jeweiligen Items sind, desto stärker weichen die Koordinatenpunkte von der Geraden ab. Dies wird als erster Hinweis auf DIF gewertet.
- *Schritt 2*: Die visuelle Prüfung wird dadurch fortgeführt, dass für jede pro Untergruppe geschätzte Itemschwierigkeit ein 95%-Konfidenzintervall berechnet wird, in dem der Schätzwert unter Berücksichtigung einer Irrtumswahrscheinlichkeit von 5% liegt. Hier wird geprüft, inwiefern sich die beiden pro Item und Untergruppe berechneten Schätzwert-Intervalle überschneiden. Je mehr sie sich überschneiden, desto wahrscheinlicher ist es, dass bezüglich des betrachteten Items kein *DIF* besteht.
- *Schritt 3*: Um den Grad der Abweichung zweier Schätzwerte schließlich im Sinne einer Effektstärke beurteilen zu können, werden abschließend die Logit[6]-Differenzen der beiden Schwierigkeits-Schätzungen berechnet. Gemäß Wilson (2005) sind die berechneten Logit-Differenzen folgendermaßen zu bewerten:
  - Werte $< 0,43$: vernachlässigbare Differenz
  - Werte zwischen $0,43$ und $0,64$: mittelmäßige Differenz
  - Werte $> 0,64$: große Differenz.

GLOBALE MODELLPRÜFUNG

Die Globale Modellprüfung wird durchgeführt, um festzustellen, inwiefern sich die theoretisch angenommenen strukturellen Grundlagen der Kompetenz anhand der erhobenen Daten bestätigen lassen. Sie ist notwendig, da die anhand des Tests zu messende Kompetenz den theoretischen Ausführungen zufolge als eindimensional

6 Logits stellen die Maßeinheit probabilistischer Kennwerte dar. Mit Logit-Differenzen sind hier die Differenzen zwischen den pro Item und Stichprobe geschätzten Itemschwierigkeiten gemeint.

angenommen wird, aber anhand dreier Fertigkeiten operationalisiert wurde. Somit besteht die Notwendigkeit im Rahmen der Modellprüfung, das eindimensionale Modell, das die naturwissenschaftliche Grundbildung als eine einzige Dimension sieht, mit dem dreidimensionalen Modell zu vergleichen, das die drei Fähigkeitsbereiche *H*, *P* und *N* als einzelne Dimensionen annimmt. Um entscheiden zu können, welches theoretische Modell eine Datenlage am besten erklärt, gibt es unterschiedliche Modellselektionsmaße oder auch Informationskriterien, die dieser Entscheidung dienen. Es werden zwei Maße dargestellt, die im Falle dieser Testentwicklung von Bedeutung sind, der sogenannte CAIC (Consistent Akaike´s Information Criterion) und der BIC (Bayes Information Criterion).

$$\boxed{CAIC = -2lnL + (lnN) * n_p + n_p} \tag{3.8}$$

$$\boxed{BIC = -2lnL + lnN * n_p} \tag{3.9}$$

Wichtige Größen, die in die Berechnung der beiden Werte einfließen sind die *Likelihood L*, also die Wahrscheinlichkeit der Daten unter der Annahme, dass ein bestimmtes Modell gilt, die *Größe der Stichprobe N* und die *Anzahl der Parameter* n_p. Eine hohe Wahrscheinlichkeit führt zu einem besseren Wert im Modellgeltungsmaß. Eine hohe Anzahl an Parametern wirkt sich dagegen negativ auf das Modellgeltungsmaß aus. Am Ende wird das Modell zur Erklärung der Datenlage herangezogen, das mit einer hohen Wahrscheinlichkeit und auf möglichst einfache Weise (mit möglichst wenigen Parametern) die Datenlage zu erklären vermag. Da inferenzstatistische Tests anhand des CAIC und des BIC nicht möglich sind, ist letztendlich das Modell als angemessener zu bewerten, das den geringeren CAIC oder BIC besitzt. Man muss sich in der Prüfung der Modelle daher damit zufrieden geben, das relativ beste zu finden (Rost, 2004a).

EXTERNE VALIDIERUNG

Zur Prüfung der externen Validität wird die gemessene Testleistung in Beziehung zu externen Validitätskriterien gesetzt. Zu diesem Zweck wird die Testleistung mit verschiedenen Außenkriterien (Leistungs- und Motivationskriterien) korreliert. Dieser Teil der Validierung gliedert sich folgendermaßen:

1. Prüfung der *konvergenten Validität* durch Korrelation der Testergebnisse mit den Schulnoten der Schülerinnen und Schüler:
   - Prüfung der Korrelation zwischen gemessener Testleistung und Naturwissen-

schaftsnoten auf Signifikanz;

- Vergleich der Korrelation zwischen Testleistung und Naturwissenschaftsnoten mit der Korrelation zwischen Testleistung und nicht naturwissenschaftlichen Noten[7])

2. Prüfung der *diskriminanten Validität*:

   - Vergleich der Korrelation zwischen zwei Leistungskonstrukten mit der Korrelation zwischen einem Leistungs- und einem Motivationskonstrukt: Hier werden konkret die Korrelationen zwischen Testleistung und Schulnoten verglichen mit der Korrelation zwischen Testleistung und Fachinteresse, Testleistung und Interesse an naturwissenschaftlichen Aktivitäten sowie Testleistung und Interesse an naturwissenschaftlichen Tätigkeiten;
   - Vergleich der Korrelation zwischen Testleistung und inhaltlich näheren Motivationskonstrukten: Hier wird die Korrelation zwischen Testleistung und Interesse an naturwissenschaftlichen Tätigkeiten mit der Korrelation zwischen Testleistung und Interesse an naturwissenschaftlichen Aktivitäten verglichen.

Eine der genannten Variablen, die Schulnoten, liegt in Form ordinaler Daten vor. Vor dem Hintergrund, dass die Testleistung in Form intervallskalierter WLEs vorliegt, bedeutet dieser Umstand, dass Korrelationen von Daten unterschiedlichen Skalennniveaus berechnet werden müssen. Aus diesem Grund wird zur Berechnung der einseitigen bivariaten Korrelationen der nicht parametrische *Spearman-Korrelationskoeffizient* herangezogen. In gleicher Weise werden die Korrelationen zwischen Schulnoten und Interessensvariablen berechnet.

Die Interessensvariablen werden in Form von Ratings erhoben, die zunächst zu einer ordinalen Datenlage führen. Für die Weiterverarbeitung der Daten werden die einzelnen Fachinteressen zu übergeordneten Skalen zusammengefasst. Um dies zu verwirklichen werden sie als Items der Skalen *Sprachinteresse* (Deutsch und Englisch), *nicht naturwissenschaftliches Interesse* (Deutsch, Englisch, Mathematik), *mathematisch-naturwissenschaftliches Interesse* (Mathematik, Physik, Chemie, Biologie) und *naturwissenschaftliches Interesse* (Physik, Chemie, Biologie) aufgefasst und die Werte der Testpersonen hinsichtlich dieser neuen Skalen werden in Form von WLEs geschätzt. Auch die Skalen *Interesse an naturwissenschaftlichen Tätigkeiten* und *Interesse an naturwissenschaftsbezogenen Aktivitäten* werden anhand von WLEs neu skaliert. Somit können aufgrund des gleichen Skalenniveaus der Testleistung und der Interessensvaria-

7 Als nicht naturwissenschaftliche Noten gelten hier die Deutsch-, Englisch- und Mathenote.

blen die einseitigen, bivariaten Korrelationen anhand des *Pearson-Korrelationskoeffizienten* ausgedrückt werden.

Um Unterschiede zwischen Korrelationen auf Signifikanz prüfen zu können, die an einer Stichprobe ermittelt wurden und somit voneinander abhängen, wurde die *Fishers-Z-Transformation*

$$\boxed{z = \frac{\sqrt{(n-3)} \cdot (Z_{ab} - Z_{ac})}{\sqrt{(n - 2 \cdot CV_1)}}} \tag{3.10}$$

mit n=Stichprobenumfang, Z_{ab}, Z_{ac} = Fishers Z-Werte für die Korrelationen r_{ab} und r_{ac} herangezogen, wobei a für die Testleistung steht und b und c für zwei unterschiedliche Variablen, deren Korrelationen mit der Testleistung miteinander verglichen werden sollen. CV_1 kennzeichnet die Kovarianz der Korrelationsverteilungen von r_{ab} und r_{ac}. Der auf genannte Weise ermittelte z-Wert wird mit einem kritischen z-Wert verglichen, der in diesem Fall bei einem einseitigen Test und einem Alpha-Niveau von $0,05$ bei $1,645$ liegt (Bortz, 1999).

3.2.2 RELIABILITÄT

Neben der Überprüfung der Validität ist die Genauigkeit zu prüfen, mit der ein Merkmal durch den Test erfasst wird. Dies geschieht anhand der Feststellung der *Reliabilität*. Ein Test ist um so reliabler, je geringer der Anteil zufälliger Messfehler ist (Schermelleh-Engel & Werner, 2007). Bei der Bestimmung der Reliabilität ist es nicht wesentlich, ob der Test inhaltlich das Merkmal misst, das er messen soll. Dieser Punkt ist Gegenstand der bereits beschriebenen Validierung. Eine wichtige Voraussetzung für Reliabilität ist die Objektivität eines Tests. Eine hohe Reliabilität ist nur realisierbar, wenn Testdurchführung, -auswertung und -interpretation standardisiert sind. Diese Kontrolle der Messbedingungen wird im kommenden Abschnitt weiter ausgeführt.

Es werden in diesem Zuge sowohl das klassische als auch das probabilistische Reliabilitätskonzept ausgeführt. Der Logik der Testentwicklung folgend, sollte an dieser Stelle lediglich das probabilistische Konzept dargestellt werden. Das klassische Konzept und die klassischen Reliabilitätskennwerte werden hier zusätzlich herangezogen, um das neu entwickelte Verfahren mit den in Abschnitt 2.4 dargestellten Verfahren vergleichen zu können, die lediglich klassische Kennwerte angeben.

KLASSISCHES RELIABILITÄTSKONZEPT

Im Rahmen der klassischen Testtheorie wird Reliabilität definiert als Verhältnis zwischen der Varianz der wahren Werte und der Varianz der Messwerte:

$$\boxed{Rel(x) = \frac{Var(\tau)}{Var(x)}} \tag{3.11}$$

Dabei wird τ bezeichnet als Variable der wahren Werte und x als die Variable der gemessenen Werte. *Var(x)* setzt sich dabei zusammen aus der Varianz der wahren Werte und der Messfehlervarianz:

$$\boxed{Var(x) = Var(\tau) + Var(\varepsilon)} \tag{3.12}$$

Daraus ergibt Gleichung 3.13, die Folgendes deutlich macht: Je kleiner die Fehler bei der Messung, desto geringer ist die Messfehlervarianz $Var(\varepsilon)$ und desto höher fällt die Reliabilität aus.

$$\boxed{Rel(x) = \frac{Var(\tau)}{Var(\tau) + Var(\varepsilon)}} \tag{3.13}$$

Die damit theoretisch eindeutig definierte Reliabilität kann in der Praxis nicht exakt berechnet werden, da wahre Werte und Messfehler der Testpersonen nicht bekannt sind. Das Varianzverhältnis kann jedoch geschätzt werden und dabei Werte zwischen *0* und *1* annehmen.

Von den vier klassischen Arten der Reliabilitätsschätzung[8], wird hier nur die interne Konsistenz (Cronbachs Alpha) berichtet. Sie wurde ausgewählt, da die meisten Testverfahren diesen Reliabilitätskennwert angeben und damit eine Vergleichbarkeit ermöglicht wird. Weiterhin ist die Berechnung der internen Konsistenz deshalb vorteilhaft, weil der Test lediglich einmal durchgeführt werden muss und es damit nicht notwendig ist, ihn derselben Stichprobe ein zweites Mal (Retest-Reliabilität) vorzulegen oder ihn aufwendig in zwei gleichwertige Testhälften Paralleltest-Reliabilität) zu teilen.

INTERNE KONSISTENZ (CRONBACHS ALPHA)

Die Berechnung der internen Konsistenz ist nur bei Testverfahren sinnvoll, deren Items den theoretischen Vorstellungen zufolge das gleiche Merkmal messen. Aus den Zusammenhangsstrukturen der Items wird auf das sogenannte Cronbachs α,

8 Retest-, Paralleltest-, Split-Half-Reliabilität und internen Konsistenz (Cronbachs Alpha)

die interne Konsistenz als Kennwert der Reliabilität, geschlossen. Dabei fällt dieser Kennwert um so höher aus, je stärker die Testitems untereinander korrelieren. Der Kennwert wird folgendermaßen berechnet:

$$Rel(x) = \alpha = \frac{m}{m-1} \cdot \left(1 - \frac{\sum_{i=1}^{m} Var(x_i)}{Var(x)} \right) \tag{3.14}$$

Cronbachs Alpha beruht auf dem Gedanken, dass die Kovarianzen zwischen Testitems als wahre Varianz interpretiert werden können, die auf das erfasste Merkmal zurückzuführen ist (Schermelleh-Engel & Werner, 2007).

PROBABILISTISCHES RELIABILITÄTSKONZEPT

Anders als in der klassischen wird in der probabilistischen Testtheorie die Reliabilität nicht anhand der Rohdaten geschätzt, sondern es wird sich ihr über die Personenparameterschätzer genähert. Wie bereits im Abschnitt über Parameterschätzungen (vgl. 3.1.2) beschrieben, lassen sich Personenparameter nach unterschiedlichen Methoden schätzen. In Abhängigkeit von der Parameterschätzung lassen sich analog auch unterschiedliche Reliabilitätsschätzungen berechnen.

Für den im Rahmen dieser Testentwicklung berechneten WLE-Personenschätzer lässt sich zur Reliabilitätsschätzung die *Erwartungswert-* oder auch *Andrich-Methode* (Andrich, 1988) anwenden (vglWalter, 2005).

Neben dieser Methode steht für die Reliabilitätsschätzung anhand von WLEs auch die Methode von Rost (2004a) zur Verfügung, um eine asymptotisch erwartungstreue Schätzung zu erreichen. Die Entscheidung darüber, welche Methode hier Anwendung findet, kann anhand der Testlänge getroffen werden. Bei kurzen Tests (Itemanzahl < 20 Items) unterschätzt die Andrich-Methode die Reliabilität. Diese Einschränkung ist bei der Interpretation von Reliabilitätskennwerten, die auf diese Weise berechnet wurden, zu beachten. Ab einer Testlänge von 20 Items liefern jedoch beide Methoden erwartungstreue Schätzungen und sind demnach gleichermaßen empfehlenswert. Da der *Test zur Erfassung prozessbezogener naturwissenschaftlicher Grundbildung* 30 Items umfasst, ist demnach die Anwendung beider Methoden möglich.

Die für die Parameterschätzungen herangezogene Conquest-Software verwendet zur Schätzung der Reliabilität die Andrich-Methode. Da aufgrund der ausreichenden Itemanzahl die beschriebenen Voraussetzungen für eine erwartungstreue Schätzung

der Reliabilität gegeben sind, konnte diese Methode hier Anwendung finden.

EINSCHÄTZUNG DER RELIABILITÄT

Um die Höhe der Reliabilität einschätzen zu können, seien an dieser Stelle in Anlehnung an Schermelleh-Engel und Werner (2007) einige Hinweise gegeben. Zum einen ist die Beurteilung der Reliabilitätshöhe abhängig von der Art des Merkmals, welches durch den Test erfasst wird. Handelt es sich um Leistungsmessungen, so sollte die Reliabilität in Anlehnung an die Reliabilität von Intelligenztests Werte um $0,9$ erreichen, bei Persönlichkeitstests liegen die Reliabilitäten meist um $0,7$. Im Rahmen einer Individualdiagnostik muss die Messgenauigkeit größer sein als in der Kollektivdiagnostik, da Fehlerurteile und falsche Interventionsempfehlungen dringend zu vermeiden sind. Auch in der Kollektivdiagnostik ist Fehlervarianz störend, Gruppenunterschiede sind jedoch im Falle individueller messfehlerbehafteter Werte korrekt schätzbar. Eine weitere Entscheidungsgrundlage liefern die Einsatzbedingungen des Verfahrens. Tests, die im Rahmen eines Screenings und zur ökonomischen Einschätzung eingesetzt werden, also kürzer ausfallen als aufwendige Verfahren, für die mehr Zeit zur Verfügung steht, besitzen aufgrund der geringeren Itemanzahl naturgemäß eine geringere Reliabilität.

Der Test zur Erfassung prozessbezogener naturwissenschaftlicher Grundbildung ist als ökonomisches Verfahren für ein Gruppen-Screening angelegt. In diesem Fall bedeuten diese Ausführungen, dass eine Reliabilität zu erwarten sein sollte, die unter der von Intelligenz- und Individualdiagnostik, jedoch über der von Persönlichkeitstests liegt. Es sollten also Werte zwischen $0,7$ und $0,9$ erreicht werden.

3.2.3 OBJEKTIVITÄT

Wie bereits im Abschnitt über die Reliabilität erwähnt, stellt die *Objektivität* eines Verfahrens eine wichtige Voraussetzung für die Reliabilität dar. Je nach Phase der Testdurchführung, in der Störungen der Objektivität auftreten, werden *Durchführungs-, Auswertungs- und Interpretationsobjektivität* unterschieden. Das Gütekriterium der Objektivität drückt das Bestreben aus, die Testergebnisse auf die Unabhängigkeit von zufälligen oder systematischen Verhaltensvariationen des Untersuchers zu prüfen.

Die *Durchführungsobjektivität* betrifft den Grad der Unabhängigkeit von der den Test durchführenden Person. Verhaltensweisen des Untersuchers haben Einfluss auf das Verhalten der Testpersonen. Um eine maximale Durchführungsobjektivität sicherstellen zu können, sollte es festgelegte Instruktionen für die Testdurchführung

geben, an die sich jeder Testleiter genau zu halten hat. Dieser Weg wird im Rahmen dieser Testentwicklung beschritten. Die Testdurchführung ist durch standardisierte Instruktionen festgelegt, wird immer mit den gleichen einführenden Worten begonnen und es ist festgelegt, welche Informationen den Testpersonen vermittelt werden (vgl. Anhang A.1, S. 241). Auf diesem Weg wird die Durchführungsobjektivität sichergestellt.

Die *Auswertungsobjektivität* beschäftigt sich mit der numerischen oder kategorialen Auswertung von registriertem Testverhalten nach vorgegebenen Regeln. Im Falle eines Leistungstests im Multiple-Choice-Format, der - wie in diesem Fall - klar festgelegte richtige und falsche Antworten besitzt, ist diese Anforderung dadurch praktisch vollkommen verwirklicht.

Unter dem Begriff der *Interpretationsobjektivität* wird der Grad der Unabhängigkeit der Interpretation des Testergebnisses von der Personen verstanden, welche die Testergebnisse interpretiert. Sie ist dann gegeben, wenn unabhängig von der interpretierenden Person gleiche Testergebnisse unterschiedlicher Personen die gleichen Schlüsse gezogen werden. In Leistungstests ist die Interpretationsobjektivität dann perfekt gegeben, wenn sie normiert sind, wenn also jedes Testergebnis einem Leistungsniveau zugeordnet wird. Dieses Ziel wird auch in dieser Testentwicklung verfolgt.

3.3 VERFAHRENSTHEORETISCHE GRUNDLAGEN

Zum Abschluss der theoretischen Grundlagen wird der folgende Abschnitt auf verfahrenstheoretische Aspekte der Testentwicklung eingehen, die zum einen das Test- und Antwortformat der Items sowie die Aspekte betreffen, die bei der Zusammenstellung von Items zu einem Test zu beachten sind. Zum anderen verbindet der Abschnitt die theoretischen Aspekte der Testentwicklung mit der praktischen Arbeit der Itementwicklung, die in Kapitel 4, der Operationalisierung, beschrieben wird. Diese Absicht verfolgt insbesondere der letzte Teil des Abschnitts, der einen Überblick über den Testentwicklungsansatz gibt, den diese Arbeit verfolgt. Er dient als Zwischenfazit, indem er einerseits Rückschau auf die bereits abgeschlossenen theoretischen Teile der Testentwicklung hält und andererseits den Blick auf die praktische Itementwicklung richtet.

3.3.1 MULTIPLE-CHOICE-TEST- UND ANTWORTFORMAT

Gemäß Haladyna et al. (1994) existieren grob unterteilt zwei Testformate: eines, in dem die richtige Antwort aus mehreren Antwortalternativen ausgewählt werden

muss (*Selected Response*) und eines, in dem die Antwort selbst konstruiert werden muss (*Constructed Response*). Die Entscheidung darüber, welches Format in einer Testentwicklung zum Einsatz kommt, hängt laut diesen Autoren zum einen davon ab, ob der Test Wissen oder eine Fertigkeit bzw. das Produkt einer Fertigkeit erfasst. Klassischerweise würden Selected-Response-Formate für die Erfassung von Wissen zum Einsatz kommen, während Fertigkeiten vorwiegend durch Constructed-Response-Formate festgestellt werden sollten. Für komplexere mentale Prozesse wie das Schlussfolgern und Problemlösen können beide Formate eingesetzt werden.

Um die in Abschnitt 2.7.2 genannten Anforderungen erfüllen zu können, dass das angestrebte Testverfahren ökonomisch durchführbar, auswertbar und interpretierbar sein und komplexe mentale Prozesse erfassen soll, die dem Problemlösen ähnlich sind, fiel die Wahl des Testformats auf ein Multiple-Choice-Format, also auf ein Selected-Response-Format. Der Multiple-Choice-Aufgabentyp vereint die Vorteile des Richtig-Falsch- und des offenen Antwortformats. Es ist in Durchführung und Auswertung objektiv und die Ratewahrscheinlichkeit in der Beantwortung kann durch eine ausreichende Anzahl von Antwortalternativen vermindert werden (Lienert & Raatz, 1998). Ein Überblick über mögliche Multiple-Choice-Formate wurde bereits in Abschnitt 2.4.5 gegeben. An dieser Stelle wird konkret auf die Besonderheiten des Formats eingegangen, das der Itementwicklung als Gerüst dient. Es stellt eine spezielle Form des *konventionellen* Multiple-Choice-Formats dar und besteht aus einem Aufgabenstamm, der das Problem bzw. die Aufgabenstellung enthält, sowie aus vier Antwortmöglichkeiten. Was dieses Format vom konventionellen Format abhebt, ist die Entwicklung von kontextabhängigen Itemsets. Diese ermöglichen eine Erfassung komplexerer Denkvorgänge und von Fertigkeiten (im Sinne einer Anwendung von Wissen auf konkrete Probleme) und besitzen ein größeres Potential für effiziente Messungen in diesem Bereich als das konventionelle Format (Haladyna, Downing & Rodriguez, 2002). Ein Itemset besteht im Falle dieser Testentwicklung aus drei thematisch zusammengehörigen Szenarien, aus denen sich jeweils eine Aufgabenstellung ergibt. Der Itemstamm besteht hier demnach nicht - wie bei einfachen MC-Aufgaben sonst üblich - aus einem einzigen Satz bzw. einer einzigen Frage, sondern er führt in einen naturwissenschaftlichen Themenbereich ein. Wichtig ist, dass die einzelnen Aufgaben eines Sets thematisch zusammengehörig, aber dennoch voneinander unabhängig sind. Die richtige oder falsche Beantwortung einer Aufgabe darf keine Auswirkungen auf die Schwierigkeit der folgenden Aufgaben haben.

Ausgehend von dem Szenario folgt ein Antwortformat, das aus einer richtigen Antwort, der Lösung (wird kodiert mit *1*), und drei falschen Antworten, den Distrakto-

ren (werden kodiert mit *0*), besteht. Ihr Zweck besteht darin, die Testpersonen von der Lösung abzulenken. Die besondere Herausforderung in der Entwicklung der Distraktoren besteht darin, dass sie plausibel erscheinen müssen, um ihren Zweck erfüllen zu können. Dazu gehört unter anderem, dass sie in grammatischer Form, in Stil und Länge der Lösung ähneln müssen. Ein guter Distraktor muss Testpersonen mit geringer Kompetenz attraktiv erscheinen und für Testpersonen mit hoher Kompetenz als solcher erkennbar sein (Haladyna, 1994; Lienert & Raatz, 1998). Wichtig ist, dass in der Analyse der Items im Laufe der Testentwicklung auch die Antwortalternativen einer genauen Prüfung unterzogen werden. Die Distraktoren sollten bezüglich ihrer Auswahlwahrscheinlichkeit annähernd gleich sein. Abschnitt 4.3 wird detailliert und mit konkreten Beispielen auf die Itementwicklung und die Entwicklung der Distraktoren eingehen.

Eine wichtige Frage, die bezüglich der Gestaltung des Antwortformats geklärt werden muss, ist die nach der Anzahl der Distraktoren. Eine natürliche Beschränkung der Anzahl liegt in der Tatsache, dass es schwierig ist, qualitativ gleichwertige Distraktoren zu produzieren. Haladyna et al. (2002) fanden in Untersuchungen heraus, dass die Anzahl brauchbarer Distraktoren im Durchschnitt zwei nicht überschreitet und dass eine Anzahl von drei Distraktoren eine Art natürliches Limit darstellt. Forschungsergebnisse zu Vor- und Nachteilen von mehr oder weniger Antwortalternativen sind in ihren Aussagen nicht konsistent. Manche stellen eine Verminderung der Schwierigkeit fest (Rogers & Harley, 1999; Sidick, Barrett & Doverspike, 1994), andere einen Anstieg der Schwierigkeit (Crehan, Haladyna & Brewer, 1993). Einige stellen eine Verbesserung der Diskriminanz fest, andere können keine Veränderungen finden. Auch in der Feststellung einer erhöhten oder verminderten Reliabilität gehen die Forschungsergebnisse auseinander und oft wird hauptsächlich festgestellt, dass Multiple-Choice-Tests mit drei Antwortalternativen keine schlechteren Kennwerte aufweisen als diejenigen mit vier Alternativen. Klare Aussagen bezüglich einer optimalen Anzahl der Antwortalternativen sind also nicht möglich. Allerdings ist die maximale Anzahl zum einen durch die Anforderung begrenzt, dass die Distraktoren in oben beschriebenem Sinne gleichermaßen gut funktionieren müssen und zum anderen spielt eine Rolle, wie viele Distraktoren inhaltlich sinnvoll sind. Grundsätzlich schlagen Haladyna et al. (2002) vor, zunächst so viele Distraktoren wie möglich zu formulieren, da sich im Laufe der Testentwicklung nicht nur Items als Ganzes als unbrauchbar erweisen, sondern auch einzelne Distraktoren. Da also keine gesicherten Aussagen zur optimalen Distraktorenanzahl möglich sind, wurde die Entscheidung über das Format testspezifisch nach inhaltlichen Gesichtspunkten (vgl. Abschnitt

4.3.2) und auf der Grundlage der Ratewahrscheinlichkeit getroffen. Die Entscheidung fiel auf vier Antwortalternativen, also auf die Entwicklung von je drei Distraktoren. Bei vier Antwortalternativen liegt die Ratewahrscheinlichkeit bei lediglich 25%. Diese Anzahl erfüllt die Anforderung einer möglichst geringen Ratewahrscheinlichkeit und hält die Schwierigkeit in Grenzen, viele gleichwertig gute Distraktoren zu entwickeln.

3.3.2 LOGISCHE ABHÄNGIGKEIT, POSITIONS- UND REIHENFOLGEEFFEKTE

Neben der Frage, wie einzelne Aufgaben hinsichtlich des Itemstamms und der Antwortalternativen zusammengestellt sein sollten, muss auf einer weiteren Ebene der Testentwicklung entschieden werden, auf welche Weise die fertigen Testaufgaben zu einem zusammenhängenden Test kombiniert werden. Hier ist zu beachten, dass es Einflüsse gibt, die die Qualität der Items und damit des Tests herabsetzen können. Zu diesen Einflüssen gehören *logische Abhängigkeiten* einzelner Testitems in dem Sinne, dass die richtige oder falsche Beantwortung eines Testitems die Antwort auf ein folgendes Item erleichtert oder erschwert. Solcherlei Abhängigkeiten dürfen nicht bestehen und sind dringend zu vermeiden (Rost, 2004a).

Weiterhin gehören *Positions- und Reihenfolgeeffekte* in diese Gruppe ungünstiger Einflussfaktoren. Unter *Positionseffekten* werden Veränderungen der Schwierigkeit oder anderer Merkmale eines Items infolge seiner Platzierung im Test verstanden. Sie treten insbesondere bei Items am Testanfang (mangelndes Instruktionsverständnis oder Warmup-Prozesse) oder am Testende (Ermüdung, schwindende Testmotivation oder Zeitmangel) auf. Weiterhin kann dieser Effekt auch bei der Positionierung der richtigen und falschen Antworten innerhalb des Antwortformats einzelner Items auftreten.

Reihenfolgeeffekte bezeichnen die Beeinflussung der Itemantwort dadurch, welche anderen Items ihr vorangegangen sind. Bei einer Zusammenstellung der Items nach aufsteigender Schwierigkeit kann damit gerechnet werden, dass die schwierigeren Items aufgrund von Übungseinflüssen eventuell etwas leichter zu lösen sind. Weiterhin können thematisch weniger interessante Items zu einer Abnahme der Motivation führen, den Test weiter zu bearbeiten.

Für die Zusammenstellung von Items zu einem Test folgt aus diesen Überlegungen zum einen, dass Items keine logische Abhängigkeit aufweisen dürfen. Zum anderen sollte für einzelne Items oder zumindest für Itemsets innerhalb eines Tests eine Zufallsreihenfolge gewählt werden. Diese Form der Aufgabenzusammenstellung findet in der vorliegenden Testentwicklung Anwendung.

3.3.3 FEHLENDE WERTE: DEFINITION UND UMGANG

Fehlende Werte entstehen dadurch, dass Testitems unbeantwortet bleiben. Dafür kann es unterschiedliche Gründe geben, die wie folgt klassifiziert werden können. Zum einen können sie dadurch entstehen, dass den Testpersonen nicht alle Items vorgelegt wurden (*not administered*). Andere Gründe können gemäß Lord (1980) darin bestehen, dass Items am Ende eines Tests aus zeitlichen Gründen oder aufgrund einer Schwierigkeitszunahme der Items nicht beantwortet werden (*not reached*) oder dass sie aus Versehen oder aber absichtlich ausgelassen werden (*omitted*), weil sie zum Beispiel als zu schwer empfunden werden.

Um zu entscheiden, wie mit fehlenden Werten umgegangen werden soll, ist es wichtig, ihre Ursachen zu kennen oder aber begründet annehmen zu können. Der einfachste Fall fehlender Daten besteht dann, wenn Testaufgaben nicht beantwortet wurden, weil sie nicht vorgelegt wurden, also wie im Falle der PISA-Studien aufgrund eine Multimatrix-Testdesigns fehlen (*missing by design*). Sie sind geplant und stellen deshalb kein großes Problem dar, da die Testpersonen per Zufall unterschiedliche Testhefte erhalten und die fehlenden Werte nicht mit Personenmerkmalen zusammenhängen.

Wichtig für den Umgang mit fehlenden Werten ist weiterhin, ob sie zufälliger oder nicht zufälliger Natur sind. Beispielsweise ist zu untersuchen, ob systematische Häufungen fehlender Werte in bestimmten Unterstichproben oder im Hinblick auf bestimmte Items auftreten. Hier sind vor allem die Items zu beachten, die als *not reached* oder aber *omitted* klassifiziert werden.

Fehlende Werte des Typs *missing by design* und *not reached* können im Rahmen der vorliegenden Testentwicklung nicht auftreten, da allen Testpersonen alle Items vorgelegt werden und ihnen ausreichend Zeit zur Verfügung steht, um alle Items zu bearbeiten. Sehr wohl kann es zu fehlenden Werten des Typs *omitted* kommen. Zeigen sich hier Muster fehlender Werte in Abhängigkeit von Personen- oder Aufgabeneigenschaften, so sind diese fehlenden Werte als problematisch anzusehen und es muss entschieden werden, wie mit ihnen umzugehen ist.

Aus diesen Darstellungen folgt, dass in jeder Phase der Testentwicklung eine Analyse fehlender Daten durchgeführt werden muss, um systematische Häufungen feststellen zu können. Der Anteil fehlender Werte, der hier toleriert wird, orientiert sich an dem von Lüdtke et al. (2007) in einem etwas anderen Zusammenhang beschriebenen 5%-Kriterium. Bei diesem Anteil fehlender Werte gehen die Autoren davon aus, dass sich selbst ein strenger fallweiser Datenausschluss nicht negativ (im Sinne ungenauer Schätzungen) auf die statistischen Berechnungen auswirken würde.

Im Falle der hier vorliegenden Daten wurde zwar kein solcher fallweiser Ausschluss vorgenommen, das strenge 5%-Kriterium soll hier dennoch Anwendung finden, um besonders sensibel für die Häufung fehlender Werte zu sein.

Liegt der Anteil fehlender Werte unter 5%, ist ihr Anteil also gering, und häufen sie sich nicht systematisch hinsichtlich bestimmter Items oder am Ende des Testheftes, so wird im Falle dieser Testentwicklung das von Baker (2004) vorgeschlagene Vorgehen gewählt, diese Werte als falsche Antworten zu betrachten. Diesem Vorgehen liegt die Annahme zu Grunde, dass Testpersonen die entsprechenden Aufgaben nicht aus Versehen ausgelassen haben, sondern sie nicht beantworten konnten. Die Annahme ist deshalb plausibel, da die Testpersonen zu Beginn der Testung instruiert wurden, dass alle Items beantwortet werden müssen und am Ende der Testung dazu aufgefordert wurden, zu prüfen, ob dies auch der Fall ist (Lord, 1980). Ein solches Vorgehen ist allerdings lediglich bei Leistungstests (z.B. im Unterschied zu Persönlichkeitstests) möglich. Problematisch ist, dass in diesem Fall vermutlich die Personenfähigkeit unterschätzt wird, jedoch erscheint dieses Vorgehen unter den genannten Bedingungen insbesondere dann nachvollziehbar, wenn die Anzahl fehlender Werte mit der Personenfähigkeit verbunden ist, wenn also Personen mit geringerer Personenfähigkeit einen höheren Anteil fehlender Werte zeigen.

3.3.4 TESTENTWICKLUNGSANSATZ

Der im Folgenden dargestellte Testentwicklungsansatz fungiert als Rückschau auf die theoretischen Grundlagen der Arbeit und als Überleitung auf das Operationalisierungskapitel, das die praktische Umsetzung dieser Grundlagen in Items beschreibt. Er ist angelehnt an den Testentwicklungsansatz von Wilson (2005) und an das Vorgehen, das der Itementwicklung im Rahmen der PISA-Studie zu Grunde lag (OECD, 2005). Zur vertiefenden Lektüre seien weiterhin Lienert und Raatz (1998) empfohlen.

Als wichtige Bestandteile des Ansatzes werden zwei spezielle Methoden beschrieben, die für die Testentwicklung von besonderer Bedeutung waren, da sie wichtige Informationen zur Überarbeitung der Testitems lieferten: das *Expertenpanel* (in der PISA-Itementwicklung auch *Cognitive Walk-Through* genannt) und die *Cognitive-Lab-Interviews*.

Das dargestellte Prozedere, das von Wilson (2005) als Konstruktmodellierung bezeichnet wird, kann als exemplarische und aktuelle Herangehensweise an eine Testentwicklung bezeichnet werden. Sie besteht aus den vier so genannten Building Blocks der Testentwicklung, die hier übersetzt wurden mit *Inhalt und Aufbau des Konstrukts (Construct Map)*, *Item-Design (Items Design)*, *Ergebnisraum (Outcome Space)* und *Testmo-*

dell (Measurement Model).

- *Inhalt und Aufbau des Konstrukts*

  Bevor die Testentwicklung beginnen kann, muss zunächst das zu messende Konstrukt benannt sowie inhaltlich und strukturell beschrieben werden. Lienert und Raatz (1998) bezeichnen diesen Schritt als *Merkmalsanalyse*. Dafür ist zunächst ein umfassendes Studium vorliegenden Literatur bzw. bestehender Testverfahren notwendig. Die Beschreibung des Konstrukts besteht im einfachsten Fall in Form eines Kontinuums, auf dem Testpersonen später verortet werden. Bereits zu diesem frühen Zeitpunkt sollten die Abstufungen möglicher Konstruktausprägungen inhaltlich beschrieben werden. Mit der Beschreibung des Konstrukts wird nicht zuletzt auch der Zweck festgelegt, dem das zu entwickelnde Instrument dienen wird. Die zentrale Idee einer möglichst genauen Konstruktdefinition besteht darin, den Fokus auf die essentiellen Merkmale des Konstrukts zu lenken und damit die spätere Entwicklung von Testitems zu erleichtern. Die Beschreibung von Inhalt und Aufbau des Konstrukts erfolgte im Falle dieser Arbeit in Abschnitt 2.2, der die *prozessbezogenen naturwissenschaftlichen Grundbildung* definiert.

- *Item-Design*

  In einem folgenden Schritt ist zu klären, in welcher Form sich das theoretische Konstrukt manifestiert, also sichtbar wird. Diese Überlegung stellt zunächst nur eine Ahnung oder eine auf bestehenden Theorien fußende Annahme dar, die sich im Verlauf der Itementwicklung konkretisiert. Hier liefert oft erst die praktische Arbeit an der Formulierung der Items notwendige Hinweise für weitere theoretische Überlegungen bezüglich des Konstrukts. Wichtig ist, dass die theoretische und praktische Ebene, also die Ebene des latenten Konstrukts und die Ebene des messbaren Verhaltens, das sich in der Beantwortung der Testitems manifestiert, unterschieden werden. Letztendlich werden die Items zwar als Realisierungen des Konstrukts betrachtet, aber während des Entwicklungsprozesses ist es unerlässlich, immer wieder beide Blickwinkel einzunehmen und sowohl vom Konstrukt aus auf die Items als auch von den Items aus auf das Konstrukt zu schauen. Auf diese Weise kommt es auf beiden Seiten zu qualitativen Verbesserungen und Verfeinerungen.

  Was die Entscheidung für ein bestimmtes Itemformat angeht, so spielen neben den bereits genannten inhaltlichen Überlegungen, auf welche Weise sich die Ausprägungen eines latenten Konstrukts manifestieren und messen lassen könnten, auch testtheoretische Überlegungen sowie die Testabsicht eine Rolle. Grundsätzlich sind

von geschlossenen Antwortformaten (Multiple- oder Forced-Choice, Ratings etc.) bis hin zu offenen Antwortformaten alle Varianten denkbar. Als Entscheidungsgrundlage dienen hier die Testökonomie, sowie die Validität, Reliabilität und der Objektivität des angestrebten Verfahrens (vgl. Abschnitte 2.7.2 und 3.2). Zusammen mit Abschnitt 3.3.1 erklären sie, aus welchem Grund die Items dieses Tests im Multiple-Choice-Format angelegt sind.

- *Ergebnisraum*

  Für den Schluss von einem manifesten Verhalten der Testperson auf das zu Grunde liegende Konstrukt muss zunächst eine Basis gelegt werden, um die Auswertung zu ermöglichen. Hier muss festgelegt werden, welche Aspekte der Itemantwort kategorisiert und gezählt werden. Diesen Bereich bezeichnet Wilson (2005) als Ergebnisraum. Manchmal werden den Antworten bereits per Konvention Kategorien und entsprechende Zahlen zugeordnet. Bei einfachen Multiple-Choice-Items beispielsweise ist der Ergebnisraum durch die Antwort des Probanden direkt festgelegt: Das Item wird richtig, falsch oder aber gar nicht beantwortet. Hier wird den falschen Antworten die *0* und den richtigen Antworten die *1* zugeordnet. Die Antworten in Ratingformaten besitzen zahlenmäßig feste Abstände (z.B. von 0-4, wobei größere Zahlen für eine stärkere Ausprägung stehen). Auch hier ist der Ergebnisraum klar definiert. Die Antworten auf offene Fragen jedoch bilden eine Ausnahme. Sie müssen zunächst nach einem bestimmten Schema inhaltlich interpretiert und kategorisiert werden, bevor ihnen Zahlen zugewiesen werden können. Manchmal sind die Kategorien also das Endprodukt des Ergebnisraumes und manchmal müssen ihnen erst Werte zugeordnet werden, die dann in unterschiedlicher Weise weiterverarbeitet werden.

  Auch bei Multiple-Choice-Aufgaben finden im Vorfeld der Testung Überlegungen bezüglich inhaltlich unterschiedlich zu kategorisierender Antworten statt. Anders als im offenen Antwortformat werden diese jedoch genutzt, um Antwortalternativen für das geschlossene Antwortformat auszuwählen oder zu formulieren.

  Im Falle dieser Arbeit wurde der Ergebnisraum durch die Festlegung des Antwortformats und durch die Kodierung richtiger Antworten mit einer *1* und falscher Antworten mit einer *0* in Abschnitt 3.3.1 festgelegt.

- *Testmodell*

  Ein weiterer Schritt, um vom Testverhalten auf die Konstruktausprägung zu schließen, besteht darin, eine Verbindung zwischen konkreten Werten und dem Kon-

strukt herzustellen. Dies wird im Rahmen des Testmodells geleistet, das in diesem Falle besser als Interpretationsmodell bezeichnet werden sollte. Das Modell übersetzt die mit Werten belegten Antworten der Testpersonen in eine Einschätzung der Personenfähigkeit hinsichtlich der latenten Kompetenz. Ob diese Schlüsse im Rahmen der klassischen oder der probabilistischen Testtheorie erfolgen, ist dabei unerheblich. Wichtig ist nur, dass der Testentwicklung ein solches Testmodell zu Grunde liegt und die Testwerte vor dem Hintergrund dieses Modells analysiert werden. Abschnitt 3.1.2 beschreibt, dass die Testentwicklung im Rahmen dieser Arbeit vor dem Hintergrund des Rasch-Modells erfolgt und welche Bedeutung diese Entscheidung für die Schätzung der entsprechenden Parameter und deren Interpretation hat.

Unter Berücksichtigung der Konstruktmodellierung und der Einplanung eines gewissen Itemausschusses ist ein adäquat großer Itempool zu entwickeln. Dieser Entwicklung widmet sich das folgende Operationalisierungskapitel (vgl. Kapitel 4). Es beschreibt, auf welche Weise die theoretischen Grundlagen in die Entwicklung von Items umgesetzt wurden und zu welchen Zeitpunkten der Itempool Gegenstand bestimmter Prüfungen wurde. Anhand solcher Prüfungen werden Items identifiziert, die nicht in das Testmodell passen, also nicht zur Messung der modellierten Kompetenz taugen und somit entweder eliminiert oder überarbeitet werden müssen. Die Wirkung dieser Maßnahmen wird jeweils in der Pilotierung, im Feld- und im Haupttest geprüft.

Auskunft über die Qualität der Items geben hier vor allem statistische Kennwerte sowie die Ergebnisse des *Expertenpanels*. Eine weitere Methode, die über das Expertenpanel hinaus wertvolle Informationen liefern kann, sind Interviews mit Vertretern der Zielgruppe, sogenannte *Cognitive Lab Interviews*. Die folgenden Exkurse werden sich diesen beiden Methoden widmen.

EXKURS: EXPERTENPANEL

Der Begriff *Expertenpanel* bezeichnet wiederholte Treffen unterschiedlicher Expertinnen und Experten, bei denen es darum geht, Testitems kritisch zu beurteilen und Ansätze zu ihrer Verbesserung zu finden. Ein Vorschlag, wie ein solches Panel durchgeführt werden könnte, findet sich bei Wilson (2005, S. 59 ff.). Er bezeichnet dieses Panel dort als *Itempanel*, an dem sowohl Vertreter der zukünftigen Zielgruppe des Tests als auch eine Expertengruppe beteiligt sind. Im Rahmen dieser Arbeit wurde das Itempanel in ein *Expertenpanel* und in *Cognitiv-Lab-Interviews* mit der Zielgruppe aufgeteilt.

Das *Expertenpanel* sollte aus einer heterogenen Gruppe von Personen bestehen, die zur Entwicklung und Überarbeitung der Items beitragen können. Dazu können je nach Inhaltsbereich der Testitems bestimmte Fachwissenschaftlerinnen und Fachwissenschaftler, Lehrerinnen und Lehrer, Fachdidaktikerinnen und Fachdidaktiker, Psychologinnen und Psychologen oder allgemein Personen gehören, die Erfahrung in der Entwicklung von Testaufgaben haben. Das Panel sollte je nach Entwicklungsphase des Tests unterschiedliche Schwerpunkte besitzen. Zu Beginn der Test- und Itementwicklung sollte der Schwerpunkt der Itemprüfung auf der fachlichen Richtigkeit und der Zuordnung zu bestimmten Teilen eines Rahmenkonzepts liegen. Zu einem späteren Zeitpunkt (nach Pilotierung und Feldtests) liegt der Schwerpunkt auf der Erarbeitung von Gründen für die mangelnde Qualität bestimmter Items und Antwortalternativen und auf Möglichkeiten zur Verbesserung dieser Items. Je nach Entwicklungsphase stehen also alle Testitems (zu Beginn der Prüfung) oder nur noch ein Teil der Testitems (diejenigen, die qualitativ mangelhaft sind) auf dem Prüfstand.

Eine wichtige Voraussetzung für ein Expertenpanel besteht darin, den Expertinnen und Experten im Vorfeld des eigentlichen Treffens alle zur Beurteilung der Items notwendigen Informationen zur Verfügung zu stellen. Dazu gehören:

- Das Rahmenkonzept des Tests und alle notwendigen Hintergrundinformationen
- Die Items selbst, ihre Einordnung in das Rahmenkonzept des Tests und ihre Beziehung zum Konstrukt
- Itemkennwerte, wie Trennschärfe und Schwierigkeit (falls bereits eine Pilotierung oder Feldtestung stattgefunden hat)
- Hinweise dazu, in welcher Form die Itembeurteilung zu verfassen ist

Vor der Durchführung des Expertenpanels sollte sich der Testentwickler kritisch mit den Items auseinandersetzen und eigene Vermutungen bezüglich der mangelnden Qualität von Items oder Antwortalternativen formulieren. Das Treffen des Expertenpanels sollte so ablaufen, dass der Test systematisch Item für Item analysiert und diskutiert wird. Die Meinungen der Expertinnen und Experten werden festgehalten, um später optimal für die Überarbeitung der Items genutzt werden zu können. Nach der Überarbeitung sollten weitere Treffen stattfinden, um die Adäquatheit der Itemüberarbeitung zu prüfen. Im Optimalfall sollten weitere Treffen und Diskussionen über Items zusätzlich durch Informationen aus weiteren Testungen gestützt werden, um entsprechende Entscheidungsgrundlagen zu besitzen.

Ein weiteres Feld, in dem die Expertenkommentare Anwendung finden, stellen die *Cognitive-Lab-Interviews* dar, die im folgenden Abschnitt beschrieben werden. Hier können die Expertenkommentare dazu dienen, vertiefend auf die Schüleräußerungen einzugehen, die in diesem Rahmen erhoben werden.

Exkurs: Cognitive-Lab-Interviews

Cognitive-Lab-Interviews stellen neben dem Expertenpanel eine weitere Informationsquelle für die Überarbeitung von Testitems dar. Diese Methode konzentriert sich auf Vertreter der Zielgruppe des Tests. Um möglichst unmittelbar zu erfahren, inwiefern Items wie angestrebt funktionieren, werden die Testpersonen im Rahmen der Cognitive-Lab-Interviews gebeten, während der Beantwortung von Testaufgaben laut zu denken. Dabei sind sie aufgefordert, alle Denk- und Lösungsprozesse sowie alle Verständnisschwierigkeiten zu äußern.

Zum ersten Mal wurden die Prozeduren, die heute in Cognitive Labs durchgeführt werden, von Karl Duncker (1945) als *think aloud verbalizations* bezeichnet und als Weg beschrieben, die Gedankenentwicklung von Probanden zu erfassen. Später griffen Ericsson und Simon (1980, 1993) diese Methode als Zugang zur Untersuchung kognitiver Prozesse auf. Da alle kognitiven Prozesse durch das Kurzzeitgedächtnis gingen, könnten Personen alle bewussten Gedanken zu dem Zeitpunkt berichten, in dem sie verarbeitet würden. Ericsson und Simon (1980, 1993) sahen die kognitiven Prozesse, die nötig sind, um Dinge zu verbalisieren als Unterform der kognitiven Prozesse, die Verhalten oder Handlungen generieren. Sie betrachteten die Verbalisierung von Gedanken während ihrer gleichzeitigen Verarbeitung im Kurzzeitgedächtnis als vorteilhaft, da sie uninterpretiert wiedergegeben würden (Ericsson & Simon, 1993). Allerdings kann die Gleichzeitigkeit des Problemlösens und der Verbalisierung Personen auch kognitiv überfordern (Branch, 2000). Um hier einen Ausgleich zu schaffen, gibt es die Möglichkeit der retrospektiven Betrachtung des Problemlöseprozesses, die zusätzlich im Anschluss an das prozessbegleitende Laut-Denken angestellt werden kann. In diesem Fall wird daher ein zweistufiges Vorgehen vorgeschlagen (Snijkers, 2002; Johnstone, Bottsford-Miller & Thompson, 2006):

1. Testpersonen verbalisieren ihre Denkprozesse. Alle Äußerungen werden erfasst (möglichst gestützt durch Tonbandaufnahmen), und die Testpersonen werden während ihrer Ausführungen nicht unterbrochen. Entstehen längere Pausen, so werden sie aufgefordert, weiterhin laut zu denken.

2. Ist der Prozess des lauten Denkens abgeschlossen, so besteht die zweite Stufe dar-

in, Anschlussfragen zu stellen. Diese Methode stellt keine primäre Datenquelle dar, sondern dient vorwiegend der Präzisierung der unter Punkt 1 erhobenen Daten.

Auf der zweiten Stufe können gegebenenfalls bereits gesammelte Expertenkommentare genutzt werden, um vertiefend auf Äußerungen der Testpersonen einzugehen. Hier ist wichtig, gezielte Nachfragen erst dann zu stellen, wenn die Testpersonen ihre Kommentare abgegeben haben, um suggestive Einflüsse zu vermeiden.

Die Stärke der Cognitive-Lab-Interviews liegt darin, unmittelbar zu erfahren, wie Vertreter der Zielgruppe sich konkret der Lösung von Testitems nähern und ob sie diejenigen Strategien und Lösungsprozesse anwenden, welche als notwendig angenommen wurden. Es können konkrete Hinweise gesammelt werden, ob Items funktionieren oder aus welchen Gründen sie dies nicht tun (Paulsen, Best, Levine, Milne & Ferrara, 1999). Paulsen et al. (1999) nennen folgende Probleme, die im wesentlichen dazu führen können, dass Items nicht funktionieren und die zur Kategorisierung der im Rahmen der Cognitive-Lab-Interviews identifizierten Probleme genutzt werden können:

- Die Lösung wird ungenügend durch den Iteminhalt gestützt.
- Es wird eine alternative *richtige* Antwort durch den Iteminhalt gestützt.
- Die Lösung des Items ist möglich, ohne dass die beabsichtigten Strategien Anwendung finden.
- Sprachliche Ausdrücke sind unklar, mehrdeutig oder unbekannt.

Die Stichprobengröße im Rahmen von Cognitive-Lab-Interviews ist sehr begrenzt. Dies ist durch die Aufwendigkeit der Methode begründet, da meist Einzel- oder Kleingruppeninterviews durchgeführt werden (OECD, 2005). Nielsen (1993) fand heraus, dass eine Gruppe von fünf Personen ausreichen würde, um 75% der Probleme zu erkennen, die eine Gruppe von zwanzig Personen erkennt. Bezieht man Überlegungen zum Aufwand dieses Verfahrens mit ein, so spricht dieses Ergebnis dafür, dass bereits kleine Stichprobengrößen einen großen Ertrag im Hinblick auf die qualitative Verbesserung von Testitems haben können. Die Entscheidung für eine maximale Stichprobengröße ist also eine Frage des ökonomischen Vorgehens und der personellen Ressourcen einer Testentwicklung. Wichtig für diese Entscheidung ist auch die Tatsache, dass Cognitive-Lab-Interviews keinen Ersatz für die statistischen Verfahren zur Feststellung von Itemdefiziten darstellen, sondern vielmehr als eine wertvolle Ergänzung dienen.

4 Operationalisierung

Der Bereich der Operationalisierung beschreibt die Umsetzung der vorgegebenen Kompetenzstruktur in Aufgaben, anhand derer die *prozessbezogene naturwissenschaftliche Grundbildung* von Schülerinnen und Schülern der neunten Klasse gemessen werden kann. Ausgehend von der übergeordneten Ebene des Messmodells und der allgemeinen Teststruktur wird auf die Itemstruktur und die Entwicklung des Itemstamms eingegangen, bevor die Entwicklung der Antwortalternativen folgt. Abschließend wird der Prozess der Itemüberarbeitung und -prüfung von der Pilotierung über den Feldtest bis hin zum Haupttest, sowie die Umsetzung der Validierung beschrieben. Im Zuge der Aufgabenüberarbeitung wird unter anderem auf die Stichproben eingegangen, welche die für die Überarbeitung der Items notwendigen Daten lieferten.

4.1 Strukturelle Grundlagen der Itementwicklung

Die Grundlagen der Testentwicklung werden auf unterschiedlichen Ebenen gelegt: zunächst inhaltlich in Form einer möglichst exakten Beschreibung des zu messenden Merkmals, dann strukturell in Form des Testmodells und des Messmodells und schließlich als Vorform der praktischen Itementwicklung in Gestalt der Ausformulierung von Itemmustern.

Nachdem die zu messende Kompetenz bereits in Abschnitt 2.2 definiert und das angewandte Testmodell in Abschnitt 3.1.2 ausgeführt wurde, wird an dieser Stelle auf einen weiteren grundlegenden Ausgangspunkt, das in Abschnitt 4.1.1 dargestellte Messmodell, eingegangen. Auf diese Weise wird die Entwicklung der Items transparent gestaltet und die Bedeutung des Messmodells im Rahmen der Itementwicklung verdeutlicht. Jedes Abweichen von dieser theoretisch begründeten Struktur hätte zur Folge, dass Items die zu messenden Fertigkeiten nicht mehr adäquat repräsentieren würden, also von den Messwerten nicht auf die zu Grunde liegende Kompetenz geschlossen werden könnte.

4.1.1 MESSMODELL

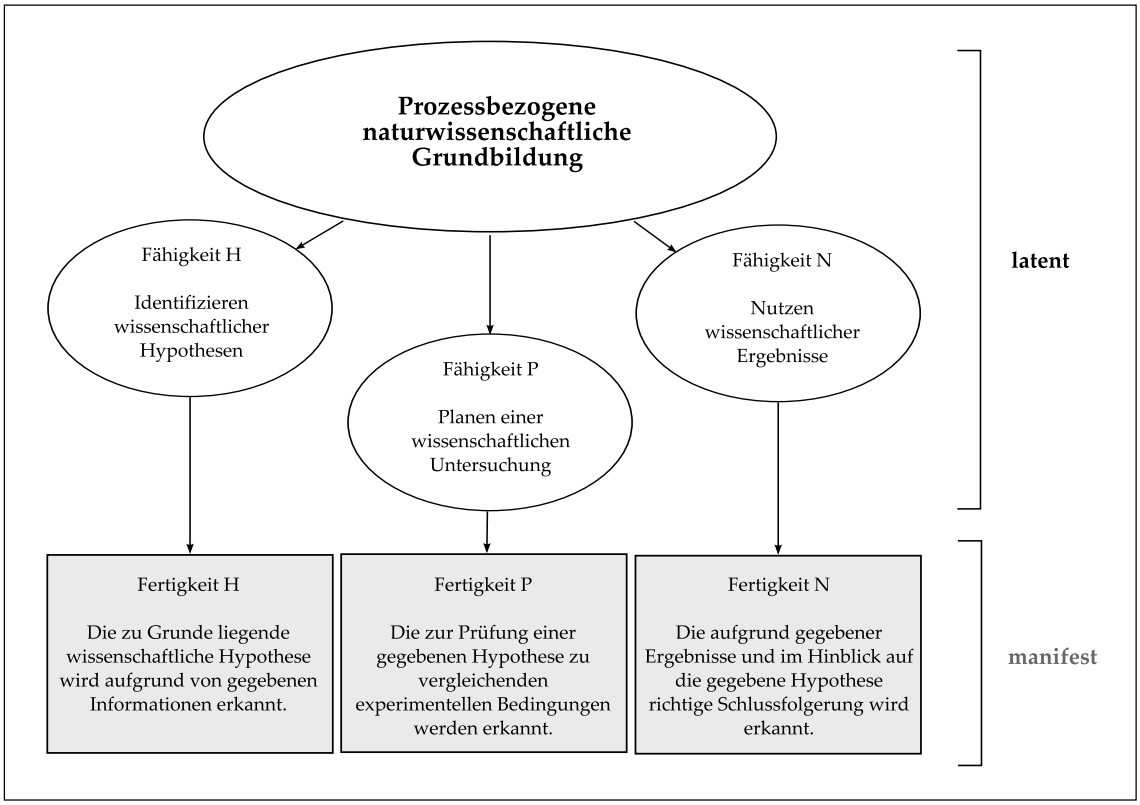


Abbildung 4.1: Messmodell

Abbildung 4.1 stellt das der Testentwicklung zu Grunde liegende Messmodell dar. In Anlehnung an die Darstellungskonventionen in Pfaddiagrammen im Rahmen von Strukturgleichungsmodellen finden sich die latenten, also keiner Messung zugänglichen, Kompetenzen und Fähigkeiten in den Ovalen. Die manifesten, und damit messbaren Fertigkeiten, werden dagegen in Vierecken dargestellt. Zur visuellen Unterstützung wurden sie in der Abbildung grau unterlegt, da sie in der Folge die Grundlage der Itementwicklung bilden werden.

Die Pfeile, die im Messmodell von der zu Grunde liegenden *prozessbezogenen naturwissenschaftlichen Grundbildung* ausgehen, geben an, dass die drei Fähigkeiten *H*, *P* und *N* repräsentativ für diese Kompetenz stehen. Als Repräsentanten der *prozessbezogenen naturwissenschaftlichen Grundbildung* wird für die Fähigkeiten ein hoher korrelativer Zusammenhang angenommen, der allein durch den gemeinsamen Ursprung in der Kompetenz erklärt wird.

Operationalisiert werden die drei Fähigkeiten durch die ihnen untergeordneten Fertigkeiten *H*, *P* und *N*, die ebenfalls durch Pfeile mit den einzelnen Fähigkeiten

verbunden sind. Die Pfeile stellen dar, dass die Fähigkeiten anhand der genannten Fertigkeiten messbar werden. Auch hier wird, ebenso wie bei den latenten Fähigkeiten, ein hoher korrelativer Zusammenhang der Fertigkeiten erwartet. Jede Fertigkeit wird im Modell anhand eines kurzen Satzes definiert. Da die Definitionen die Basis der Itementwicklung bilden, wurden die zu messenden Fertigkeiten möglichst konkret formuliert. Die drei Fertigkeiten *H*, *P* und *N* bilden in Kombination mit dem grundsätzlichen Konstruktionsprinzip, welches in Abschnitt 4.3 beschrieben wird, das Gerüst der Itementwicklung. Auf der Basis dieses Gerüsts werden die Inhalte des Stimulus-Materials zu Aufgaben. Alle Ideen, die im Rahmen der Itementwicklung entstehen, müssen zu den Definitionen der Fertigkeiten und zu den Konstruktionsprinzipien passen. Die Definitionen und Prinzipien leiten wiederum ihrerseits die Entwicklung von Ideen.

4.1.2 TESTSTRUKTUR

Unter der Teststruktur wird die Zusammenstellung der Items innerhalb eines Szenarios und die Verteilung der Szenarien über den gesamten Test hinweg verstanden. Dieser Abschnitt gibt Auskunft über die Gesamtzahl der Testitems und über Anzahl und Reihenfolge der Items innerhalb eines Itemsets, sowie über den gesamten Test hinweg. Die Teststruktur ist durch folgende Faktoren bedingt: Sie ergibt sich zum einen aus dem Messmodell, das drei zu messende Fertigkeiten vorgibt, die jeweils anhand spezieller Items erfasst werden müssen. Zum anderen ergibt sie sich daraus, dass eine ausreichend große Grundgesamtheit an Items vorhanden sein muss, um auch nach der im Testentwicklungsverlauf auftretenden Entfernung qualitativ unzureichender Items noch eine statistisch angemessene Anzahl zur Messung der Kompetenz zurückzubehalten. Ein weiteres Strukturelement besteht darin, dass jedes Itemset je eine Aufgabe zur Messung der jeweiligen Kompetenz enthält, es also pro Itemset drei Aufgaben gibt. Diese Struktur wird durch Abbildung 4.2 noch einmal verdeutlicht.

Die Informationen, die in den Itemstämmen der Sets präsentiert werden, sind möglichst einfach und logisch nachvollziehbar dargestellt. In diesem Sinne sind auch Bilder, Graphiken und Tabellen in den Text eingebettet. Sie sollen ein optimales Verständnis aller notwendigen Informationen ermöglichen. Die genannten Darstellungsarten dienen der Vereinfachung und der Vermeidung von überflüssigem Text und sind so entwickelt, dass sie den Schwierigkeitsgrad der Aufgabe nicht erhöhen. In Bezug auf die Themenbereiche, die die Items abdecken, wurde versucht, in möglichst ausgewogener Form biologische, physikalische und chemische Aspekte zu berück-

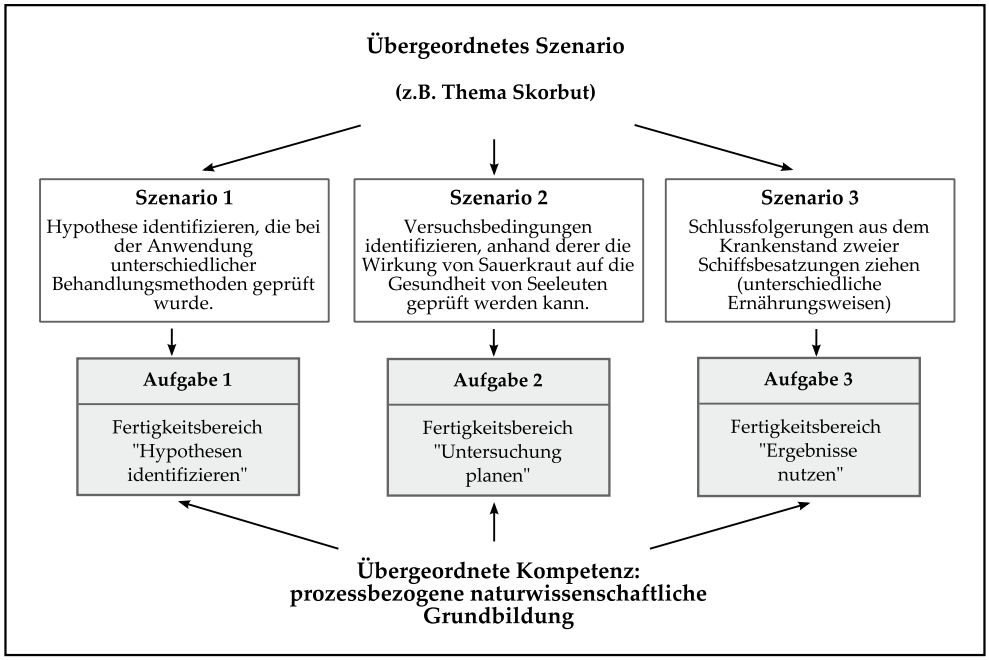


Abbildung 4.2: Itemset und Aufgaben als Verbindung zwischen theoretischer Kompetenz und Szenario

sichtigen, um keiner Testperson inhaltliche Vorteile zu verschaffen.

Jedes Itemset entstammt einem bestimmten Inhaltsbereich (z.B. Gesundheit) und wird durch ein bestimmtes Thema umschlossen (z.B. Skorbut). Dieses Thema wird zunächst als übergeordnetes Szenario allgemein eingeleitet, um alle Testpersonen auf den gleichen Wissensstand zu bringen. Im Anschluss folgen die eigentlichen Item-Szenarien, die den Testpersonen alle notwendigen Hintergrundinformationen und die Aufgabenstellung liefern. Wie in Abbildung 4.2 zu erkennen ist, enthält jedes Set drei Aufgaben.

Weiterhin ist zu erkennen, dass die einzelnen Aufgaben trotz des übergeordneten Szenarios unabhängig voneinander konstruiert sind. Ermöglicht wird dies durch eine jeweilige Spezialisierung des Themas auf bestimmte Bereiche. Diese Bereiche sind durch die theoretische Konzeption der *prozessbezogenen naturwissenschaftlichen Grundbildung* definiert. Sie erweitern das übergeordnete Szenario in Richtung der drei zu messenden Fertigkeiten. Für die Entscheidung, mit welcher Itemanzahl die Testentwicklung beginnen sollte, waren Überlegungen zur Testlänge, also zur notwendigen Aufgabenanzahl notwendig. Da keine allgemeingültigen Aussagen zur Itemanzahl im Rahmen von Tests dieser Art möglich sind, mussten im Rahmen der Entscheidung über die Testlänge unterschiedliche Punkte berücksichtigt werden. Zum einen

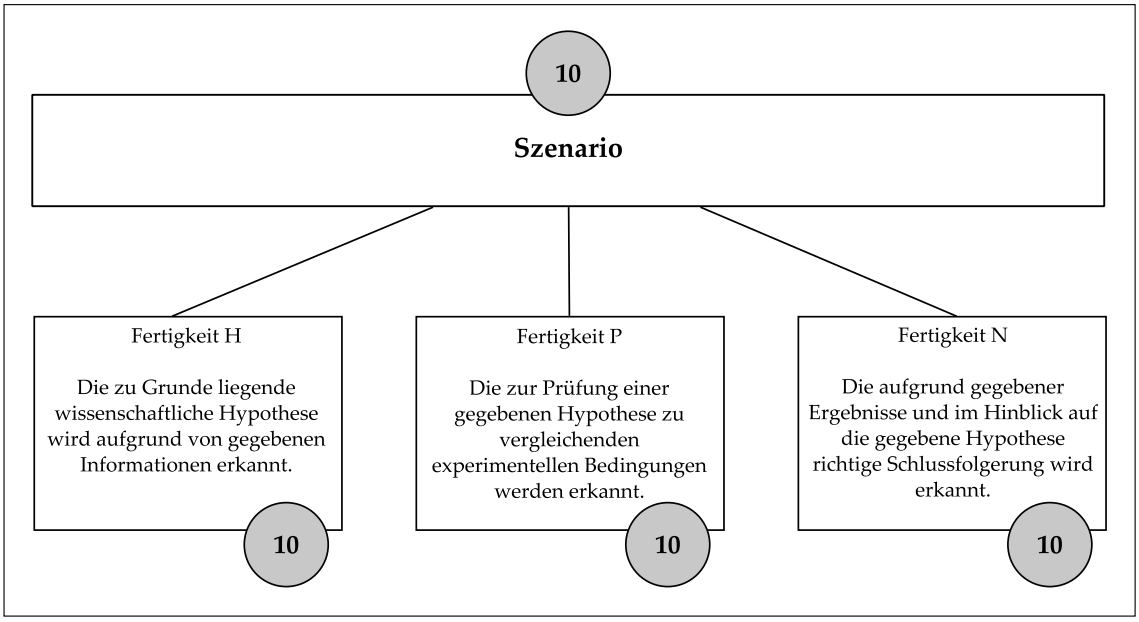


Abbildung 4.3: Itemset

war es zur Messung der Fertigkeiten notwendig, ausreichend viele Items zu entwickeln, um eine reliable Skala bilden zu können und damit eine hohe Messgenauigkeit zu erreichen. Zum anderen musste eine zu große Testlänge vermieden werden, um nicht Ermüdung, Konzentrationseinbußen oder Lern- und Übungseffekte in den Testpersonen hervorzurufen. Weiterhin musste im Laufe der Itementwicklung noch mit einer gewissen Ausfallquote von Items gerechnet werden. In Anlehnung an den PISA-Feldtest 1999 (Prenzel, Rost, Senkbeil & Klopp, 2001), der doppelt so viele Aufgaben umfasste als letztlich benötigt wurden, und in Anlehnung an Wilson (2005), der mit einer durchschnittlichen Ausfallquote von 50% der produzierten Items im Rahmen der Testentwicklung rechnet, wurde auch hier zunächst die doppelte Itemanzahl entwickelt. Weniger als fünf Items sollten pro zu messender Fertigkeit am Ende der Testentwicklung nicht zur Verfügung stehen, um noch eine möglichst reliable Messung zu erreichen. Auf Ebene der Itemsets bedeutete dies, dass mindestens fünf Itemsets mit je einem Item pro zu messender Fertigkeit entwickelt werden mussten. Unter Berücksichtigung der Ausfallquote wurden also zunächst zehn Itemsets mit je zehn Items pro Fertigkeit entwickelt (4.3). Dies ergab ein 30 Items umfassendes erstes Testheft, das ausreichend Spielraum ließ, um Items mangelnder Qualität zu entfernen. Es ermöglicht trotz dieses Umfangs eine Bearbeitung des Itempools innerhalb einer Doppelstunde, ohne dass bei den Schülerinnen und Schülern zu starke Konzentrationseinbußen zu befürchten sind.

Die Anordnung der Items ist bedingt durch die Variation der Itemset-Reihenfolge in jedem Testheft unterschiedlich. Die Reihenfolge wurde für jedes Testheft individuell per Zufall festgelegt. Auf diese Weise können Reihenfolge- und Positionseffekte (s. Abschnitt 3.3.2) ausgeschlossen werden. Innerhalb der Itemsets wurden die Aufgaben nicht rotiert. Die Items der Fertigkeiten *H*, *P* und *N* wurden so verteilt, dass jede Fertigkeit gleich häufig an jeder der drei möglichen Positionen innerhalb eines Sets gemessen wird. Die folgende Tabelle 4.1 verdeutlicht diesen Aufbau anhand einer beispielhaften Verteilung der Fertigkeitsmessungen über ein gesamtes Testheft.

Da jede der Fertigkeiten insgesamt zehn Items umfasst, ist es nicht möglich, diese absolut gleich auf die drei Positionen zu verteilen. Der Einfluss dieser kleinen Unregelmäßigkeit ist unbedeutend.

Tabelle 4.1: Beispielhafte Verteilung der Fertigkeitsmessungen über die Itemsets

Itemset/ Szenario	Fertigkeit	Position 1	Position 2	Position 3
Klima	H	X		
	P			X
	N		X	
Solarzelle	H		X	
	P			X
	N	X		
Hurrikans	H		X	
	P	X		
	N			X
Brot backen	H		X	
	P			X
	N	X		
Rost	H	X		
	P		X	
	N			X
Skorbut	H			X
	P	X		
	N		X	
Schimmelpilze	H	X		
	P		X	
	N			X
Stichlinge	H			X
	P	X		
	N		X	
Regenbogenforelle	H			X
	P		X	
	N	X		
kleine Teilchen	H			X
	P		X	
	N	X		

4.2 ZIELGRUPPE UND EXPERTENGRUPPE DER TESTENTWICKLUNG

Nachdem die strukturellen Fundamente zur Itementwicklung gelegt wurden, folgt in diesem Abschnitt zum einen die Beschreibung der Auswahl und die Charakterisierung der Stichproben, welche die Daten für die Entwicklung der Testitems geliefert haben. Bei der Auswahl war es wichtig zu beachten, dass sie die gleichen Merkmale besitzt, wie die in Abschnitt 2.7.1 definierte Zielgruppe des Tests. Aufgrund der theoretischen Grundlagen und der Definition der Zielgruppe kann die Auswahl der Stichprobe im Rahmen der Testentwicklung sich folglich nur auf Schülerinnen und Schüler der neunten Klasse beziehen.

Zum anderen wird ein Überblick über die Expertengruppen gegeben, deren Einschätzungen zur Weiterentwicklung der Items beitrugen. Im Gegensatz zur Auswahl der Schülerstichproben unterschied sich die Expertengruppe zwischen dem Feldtest und dem Haupttest in Abhängigkeit von den Zielen, die mit den Expertenurteilen verfolgt wurden.

4.2.1 BESCHREIBUNG DER STICHPROBE

Bei der Zielgruppe des Tests und damit auch bei der Stichprobe der Testentwicklung handelt es sich um Jugendliche der neunten Klasse. Sie bezieht sich gleichermaßen auf Schülerinnen und Schüler der Haupt- und Realschule und des Gymnasiums. In Bezug auf das Alter bedeutet dies, dass die Testpersonen fünfzehn bis sechzehn Jahre alt sind. Die Entscheidung für die neunte Klassenstufe folgte zum einen aus den theoretischen Ausführungen bezüglich der Entwicklung der zu messenden Fertigkeiten. Zum anderen stellen Jugendliche der neunten Klasse eine Altersgruppe dar, in der bereits ein gewisses Grundwissen im Bereich der Naturwissenschaften vorhanden ist und dieses auch verbalisiert werden kann. Weiterhin handelt es sich um die höchste Altersstufe, die es noch ermöglicht, Schülerinnen und Schüler aus Haupt-, Realschule und Gymnasium zu erfassen. Durch die Einbeziehung aller drei Schulformen wird gewährleistet, dass eine gewisse Leistungsvarianz vorliegt (statistische Begründung, s. Kapitel 3).

4.2.2 AUSWAHL DER STICHPROBE

Die Auswahl der Versuchspersonen erfolgte nicht zufällig. Zum einen wurden aus Gründen einer ökonomischen Testdurchführung schleswig-holsteinische Schulen ausgewählt. Zum anderen wird die Zufälligkeit der Auswahl dadurch eingeschränkt,

dass die Teilnahme der jeweiligen Schulen an der Testentwicklung von der Zustimmung der jeweiligen Schulleiterin bzw. des jeweiligen Schulleiters abhängig war. Die Teilnahme der Schülerinnen und Schüler hing weiterhin von der Zustimmung ihrer Eltern ab. Die Schulen entstammten zum größten Teil dem ländlichen Raum. Teilnehmende Schulen stellten jeweils alle neunten Klassen zur Verfügung.

Letztendlich ist trotz der nicht zufälligen Stichprobenauswahl nicht mit negativen Folgen für die Testentwicklung zu rechnen, da die Auswahl der Stichproben im Rahmen der Pilotierung und im Rahmen des Feld- und Haupttests in gleicher Weise erfolgte und es sich in allen Fällen um Schulen des ländlichen schleswig-holsteinischen Raums handelte. So wurde gewährleistet, dass die Stichproben ähnlich zusammengesetzt und vergleichbar sind.

4.2.3 BESCHREIBUNG DER EXPERTENGRUPPE

Die Expertengruppe bestand in allen Stadien der Testentwicklung aus einem interdisziplinären Team. Die erste Aufgabenversion wurde noch vor der Pilotierung von einem Team aus fünf Wissenschaftlerinnen und Wissenschaftlern beurteilt, das sich aus einer Diplom-Biologin, einem Biologielehrer, einem Diplom-Physiker, einer Physiklehrerin und einem Diplom-Psychologen zusammensetzte. Eine weitere Expertenrunde folgte zwischen Feld- und Haupttest. Hier setzte sich die Gruppe heterogener zusammen. Sie bestand aus zwei Naturwissenschaftlern (Physiklehrerin und Chemielehrer), zwei Studenten (Fachrichtungen Physik/Mathematik und Psychologie) und zwei Nicht-Wissenschaftlerinnen (Architektin und Buchhändlerin). Die Gründe für diese unterschiedlichen Zusammensetzungen werden im folgenden Abschnitt erläutert.

4.2.4 AUSWAHL DER EXPERTENGRUPPE

Die Zusammensetzung der jeweiligen Expertenrunden ist durch die Ziele zu begründen, die in den einzelnen Stadien der Itemüberarbeitung verfolgt wurden.

Das Ziel der ersten Itemüberarbeitung direkt nach der Erstellung der Items bestand darin, diese in Bezug auf ihre sachliche Richtigkeit zu prüfen. Die Expertengruppe hatte die Aufgabe zu entscheiden, ob die Inhalte für fünfzehn bis sechzehnjährige Schülerinnen und Schüler verständlich, in der Schwierigkeit angemessen und interessant sind. Um diese Punkte beurteilen zu können, war es wichtig, einerseits Fachwissenschaftlerinnen und Fachwissenschaftler und andererseits Lehrerinnen und Lehrer einzubeziehen, die aufgrund ihrer Lehrtätigkeit gewisse Erfahrungswerte in die

Aufgabenbeurteilung einfließen lassen konnten. So befanden sich in dieser ersten Expertengruppe eine Physiklehrerin und ein Biologielehrer, sowie zwei Personen (Diplom-Biologin, und -Physiker), die durch ihre Tätigkeit an außerschulischen Lernorten Erfahrungen in der Vermittlung naturwissenschaftlicher Inhalte an Kinder und Jugendliche gesammelt haben. Die fünfte Person, der Diplom-Psychologe, beurteilte die gleichen Punkte wie die übrigen Expertinnen und Experten und lieferte darüber hinaus wichtige theoretische Hinweise zur Überarbeitung der Teststruktur und des Testaufbaus.

Das Ziel der zweiten Itemüberarbeitung nach der statistischen Auswertung des Feldtests bestand darin, möglichst viele Hypothesen zu generieren, welche die qualitativen Unzulänglichkeiten einzelner Items oder Antwortalternativen erklären konnten. In diesem Stadium kam es also nicht mehr darauf an, die fachliche Korrektheit der Items und ihre Angemessenheit im Hinblick auf das Alter und die Interessen der Zielgruppe zu beurteilen. Aus diesem Grund wurde die Zusammensetzung der Expertengruppe hier heterogener gewählt (Wilson, 2005). Zwei der sechs Expertinnen und Experten waren erneut Fachwissenschaftler (Biologie– und Chemielehrer, Physik- und Mathematiklehrerin), zwei der Expertinnen und Experten hatten einen wissenschaftlichen Hintergrund, ohne schon zu sehr in ihrer Fachwissenschaft verhaftet zu sein (Lehramtsstudent Physik/Mathematik, Psychologiestudentin) und die übrigen zwei Expertinnen (Architektin, Buchhändlerin) zeichneten sich durch Interesse an Naturwissenschaften aus, ohne jedoch beruflich eine Verbindung zur Wissenschaft zu haben. Die unterschiedlichen fachlichen Hintergründe und der unterschiedlich stark ausgeprägte Expertenstatus gewährleisteten unterschiedliche Sichtweisen, die zu einer Vielzahl von Vorschlägen für die Überarbeitung aller unzureichenden Items führten.

4.3 ITEMENTWICKLUNG

Im Rahmen der Itementwicklung ist eine Vielzahl von Überlegungen notwendig, die zum Teil bereits in den vergangenen Abschnitten angestellt wurden und an dieser Stelle in der Beschreibung des Stimulusmaterials ihre Fortsetzung finden. Weiterhin stehen die Beschreibung des grundsätzlichen Konstruktionsprinzips und der Struktur der Items im Fokus dieses Abschnitts. Konstruktionsprinzip und Itemstruktur werden dabei anhand von Beispielen verdeutlicht. Die Gewinnung der Antwortalternativen beschließt diesen Abschnitt.

4.3.1 STIMULUS-MATERIAL

Das Stimulus-Material umfasste Bilder, Graphiken, Tabellen und Texte, die anhand der im Folgenden beschriebenen Konzepte und Strukturen in Aufgaben umgesetzt wurden. Um geeignetes Stimulusmaterial zu finden, war es wichtig, zunächst eine Auswahl in Frage kommender Materialien zusammenzutragen. Die Entscheidung über die Eignung wurde danach getroffen, ob das Material bzw. die Themen für Schülerinnen und Schüler der neunten Klasse interessant und ansprechend sind.

Das Konzept und die für die Testentwicklung vorgegebenen Strukturen leiteten die Auswahl des Materials. Auf dieser Grundlage wurde entschieden, ob vorliegendes Material in die vorgesehene Form gebracht werden konnte.

QUELLE DES MATERIALS

Die Inspiration und das Material für die Entwicklung der Items stammen aus drei unterschiedlichen Arten von Quellen: aus wissenschaftlichen Zeitschriften (Spektrum der Wissenschaft, Geo, Geolino), naturwissenschaftlichen Lehrbüchern (Bresler, Kuck, Lichtenberger & Pollmann, 2006; Halldis et al., 2005; G. Gräber et al., 2006) sowie aus dem Internet. Graphiken und Abbildungen basieren – so weit es möglich war – auf realen wissenschaftlichen Daten. Wenn es nötig war, wurden passende Graphiken erzeugt. Die wissenschaftlichen Themen und Daten entstammen zum kleineren Teil den oben genannten wissenschaftlichen Zeitschriften und Büchern. Der wesentlich größere Teil stammt aus im Internet veröffentlichten wissenschaftlichen Artikeln. Es war für die Entwicklung der Items wichtig, authentische Forschungsinhalte zu verarbeiten und die Themen dadurch interessant und glaubwürdig zu gestalten. Um auch einige leichtere Items zu generieren, wurden Szenarien gewählt, die nicht vorrangig im Forschungsbereich anzusiedeln sind, sondern bei denen es sich um die Übertragung experimenteller Anordnungen auf Alltagsgegenstände und -bereiche handelt. Ein Beispiel hierfür ist die Prüfung der für das Brotbacken wesentlichen Bestandteile.

AUSWAHL

Die Recherche des Stimulusmaterials wurde zu Beginn thematisch offen gehalten. Es wurde zunächst in Zeitschriften, Büchern und dem Internet nach interessanten naturwissenschaftlichen Themen und Materialien gesucht, die in die vorgegebene Itemstruktur gebracht werden konnten. Um die naturwissenschaftlichen Fächer gleichberechtigt zu behandeln, wurden physikalische, biologische und chemische Aspekte möglichst in gleicher Häufigkeit einbezogen. Dies wurde zum einen deshalb ange-

strebt, weil ein Test zur Messung naturwissenschaftlicher Grundbildung erfordert, dass Aspekte aller Naturwissenschaften vertreten sind. Zum anderen war die gleichberechtigte Berücksichtigung wichtig, um den unterschiedlichen Interessen der späteren Testpersonen besser entsprechen zu können, ohne denjenigen einen Vorteil oder Nachteil zu verschaffen, die sich in einem der Bereiche besonders gut oder schlecht auskennen. Zum Abschluss der Recherche des Stimulusmaterials konnten die Inhalte folgendermaßen gruppiert werden (s. Tabelle 4.2):

Tabelle 4.2: Übersicht über Inhaltsbereiche und zugehörige Aufgaben

Inhaltsbereich	**Aufgaben**
Aktuelles	Klima, Solarzelle, Hurrikans
Alltag	Brot backen, Rost
Gesundheit	Skorbut, Schimmelpilze
Natur	Stichline, Regenbogenforelle, kleine Teilchen

Diese zehn inhaltlichen Bereiche wurden in Aufgaben-Szenarien umgesetzt, das gefundene Material wurde in die bereits beschriebene Itemstruktur gebracht und die Aufgaben entsprechend der beschriebenen Teststruktur angeordnet.

4.3.2 GRUNDSÄTZLICHES KONSTRUKTIONSPRINZIP

Teil des grundsätzlichen Konstruktionsprinzips ist zum einen die in Abschnitt 3.3 erläuterte Entscheidung über das Testformat: Es handelt sich um ein Multiple-Choice-Format mit vier Antwortmöglichkeiten, wobei jeweils nur eine der Antwortmöglichkeiten richtig ist und die übrigen drei als Distraktoren fungieren. Jede Aufgabe wird durch den Itemstamm eingeleitet, in dem zunächst die grundlegenden Informationen vermittelt werden, um den Testpersonen das Aufgabenszenario zu eröffnen, aus dem sich die Aufgabenstellung ergibt. Der Itemstamm hat die wichtige Aufgabe, alle Testpersonen auf den gleichen Wissensstand zu bringen. Dies ist insbesondere deshalb wichtig, da die Wahrscheinlichkeit, Hypothesen als unplausibel zu betrachten steigt, je weniger domänspezifisches Wissen die Testpersonen besitzen (Klahr et al., 1993). Indem man ihnen allen in gleichem Maße genau dieses Wissen vermittelt, verringert

sich die Fehlerquelle, korrekte Hypothesen fälschlicherweise zu verwerfen, nur weil die Testpersonen nicht genügend Informationen besitzen.

Ausgehend vom Itemstamm folgt anschließend das beschriebene Antwortformat mit drei Distraktoren und der richtigen Antwort.

Ein weiteres wichtiges Konstruktionsprinzip besteht darin, die Vermittlung von Informationen im Itemstamm so kurz wie möglich zu halten und auf Informationen und Darstellungen zu verzichten, die nicht unbedingt notwendig sind. *Nicht notwendig* bedeutet in diesem Fall, dass sie nichts zum Verständnis, zur Veranschaulichung oder zur Informationsvermittlung beitragen. Diese Entscheidung ist dadurch begründet, dass auch Testpersonen mit Leseschwäche in der Lage sein sollen, die Aufgabeninhalte vollständig erfassen und bearbeiten zu können.

Weitere wichtige Prinzipien, die bei der Itementwicklung beachtet werden sollten, sind (vgl. auch Haladyna et al. (1994; 2002), PISA Technical Report (2005):

- Die Aufgaben müssen eindeutig einem der drei beschriebenen Fertigkeitsbereiche zuzuordnen sein.
- Die Aufgaben müssen dem Entwicklungsstand von Schülerinnen und Schülern der neunten Klasse angepasst sein.
- Die Aufgaben sollten kohärent, eindeutig und klar sein.
- Die Aufgaben sollten in sich geschlossen sein.
- Trickaufgaben sollten vermieden werden.
- Die Aufgaben sollten mit der Essenz des Stimulusmaterials verbunden sein und umgekehrt.
- Der Aufbau der Aufgaben sollte einheitlich sein.
- Es dürfen keine Abhängigkeiten zwischen Aufgaben bestehen (die richtige oder falsche Beantwortung einer Aufgabe darf die Beantwortung einer folgenden Aufgabe nicht erleichtern oder erschweren).
- Die Aufgaben sollten authentische Inhalte enthalten.
- Die Aufgaben sollten fachlich korrekte Inhalte enthalten.

Ausgehend von diesen grundsätzlichen und für alle Items geltenden Konstruktionsprinzipien ergaben sich in der Itementwicklung je nach Fertigkeitsbereich leicht unterschiedliche Strukturen des Itemstamms. Diese Unterschiede sind durch die Fertigkeitsdefinitionen inhaltlich begründet und stellen keine Abweichung vom Messmodell dar. Alle Items einer Fertigkeit sind jeweils durch den gleichen charakteristischen Aufbau gekennzeichnet.

Itemstrukturen und -beispiele

Im Folgenden werden die Konstruktionsprinzipien der Items für die einzelnen Fertigkeitsbereiche beschrieben. Die jeweilige Itemstruktur und die Unterschiede der Itemstrukturen werden anhand eines Beispiels veranschaulicht.

Fertigkeit H: „Die zu Grunde liegende wissenschaftliche Hypothese wird aufgrund von gegebenen Informationen erkannt."

Die Testpersonen zeigen durch die Beantwortung der Items dieses Bereichs, ob sie verstehen, dass Wissenschaftlerinnen und Wissenschaftler ihre Hypothesen nicht aufgrund einer Laune aufstellen, sondern sie aufbauend auf das in der Wissenschaftsgemeinschaft vorhandene Wissen formulieren. Für die Entwicklung entsprechender Items ist es wichtig, den Testpersonen die Wissensbasis der Wissenschaftlerinnen und Wissenschaftler zu vermitteln und ihnen gleichzeitig Informationen darüber zu geben, wie die Untersuchung aussieht, welche sie durchgeführt haben. Es fehlt also zwischen dem theoretischen Hintergrundwissen und der tatsächlichen Durchführung der Versuche noch die Hypothese, die diese beiden Teile verbindet. Dieses Teilstück sollen die Testpersonen identifizieren. Zur Messung der Fertigkeit *H* wurde der Itemstamm so konstruiert, dass die Testpersonen zunächst anhand eines möglichst kurz gehaltenen Textes in die Thematik eingeführt werden. Bei den Itemstämmen dieses Bereichs wurde bis auf zwei Ausnahmen auf jegliche bildliche Darstellung oder graphische Abbildung verzichtet. Die Ausnahmen bilden die Aufgaben *Kleine Lebewesen, kleine Teilchen I* und *Brot backen I*. In beiden Fällen wurden Darstellungen zur Unterstützung des Textes benötigt, um den Testpersonen eine bessere Vorstellung der Sachverhalte zu ermöglichen. Direkt im Anschluss an den Einführungstext folgt jeweils die Frage, welcher Vermutung die Wissenschaftlerinnen und Wissenschaftler nachgehen bzw. nachgegangen sind. Zur Auswahl erhalten die Testpersonen vier Antwortmöglichkeiten in Form von vier möglichen Hypothesen, von denen nur eine richtig ist. Folgende Arten von Distraktoren werden eingesetzt, um die Testpersonen

Regenbogenforelle I Wissenschaftlerinnen haben herausgefunden, dass das Verhalten von Regenborgenforellen veränderbar ist. In einem Versuch haben sie Forellen zunächst durch einen Verhaltenstest in besonders aggressive und wenig aggressive Forellen eingeteilt. Dies stellten die Wissenschaftlerinnen in dem Test dadurch fest, dass sie beobachteten, wie viel Zeit vergeht, bis sich eine Forelle einem unbekannten Objekt nähert: verging wenig Zeit ⟶ Forelle als aggressiv eingestuft verging viel Zeit ⟶ Forelle als wenig aggressiv eingestuft Im Anschluss an die Einteilung in diese zwei Gruppen ließen die Wissenschaftlerinnen je eine aggressive und eine wenig aggressive Forelle gegeneinander kämpfen. Danach stellten sie erneut durch den Verhaltenstest den Grad der Aggressivität fest.	Beschreibung der Wissensbasis der Wissenschaftler/innen + Darstellung des durchgeführten Versuchs	Aufgabenstamm
Welcher Vermutung gehen die Wissenschaftlerinnen mit ihrer Studie nach?	Aufgabe	
a) Regenbogenforellen kämpfen nur miteinander, wenn es sich um zwei aggressive Exemplare handelt. b) Aggressive Regenbogenforellen werden aggressiver, wenn sie gegen einen Artgenossen im Kampf verloren haben. c) Wenig aggressive Regenbogenforellen werden aggressiver, wenn sie einen Kampf gewonnen haben. d) Regenbogenforellen ändern ihr Verhalten nach einem Kampf, je nachdem, ob sie gewonnen oder verloren haben.	Antwortmöglichkeiten	

Abbildung 4.4: Itembeispiel Fertigkeit H

von der richtigen Lösung abzulenken:

- Der Distraktor enthält Informationen, die nicht mit den im Text genannten Informationen übereinstimmen bzw. gar nicht im Text genannt wurden.
- Der Distraktor verdreht dargestellte Informationen ins Gegenteil: die Hypothese behauptet das Gegenteil von dem, was aufgrund der dargestellten Wissensbasis logischerweise gefordert werden müsste.
- Der Distraktor gibt eine Hypothese vor, die grundsätzlich richtig und auch wissenschaftlich untersuchbar sein könnte, die aber im Falle des dargestellten Versuchs nicht gemeint ist.

- Der Distraktor benutzt Informationen aus dem Text, die lediglich Zusatzinformationen darstellen und die für die Formulierung der Hypothese nicht relevant sind.

Die richtige Beantwortung erfordert das Lesen des Textes, da die Testpersonen die Wissensbasis der in der Aufgabe genannten Wissenschaftlerinnen und Wissenschaftler erfassen müssen. Zur Lösung der Aufgabe reicht das einfache Erfassen der Informationen jedoch nicht aus. Es ist also zwar eine notwendige aber keine hinreichende Bedingung zur Lösung. Je nachdem, für welche Antwort die Testperson sich entscheidet, wird deutlich, ob sie den Text gar nicht gelesen hat, lediglich Textinformationen verarbeitet hat oder sich tatsächlich bewusst ist, dass wissenschaftliche Hypothesen aus einer Wissensbasis heraus formuliert werden. Abbildung 4.4 verdeutlicht die beschriebene Itemstruktur noch einmal.

Fertigkeit P: „Die zur Prüfung einer gegebenen Hypothese zu vergleichenden experimentellen Bedingungen werden erkannt."

Durch die Beantwortung von Items dieses Bereichs zeigen die Testpersonen, ob sie in der Lage sind, aus einer Auswahl experimenteller Bedingungen diejenigen zu erkennen, die zur Prüfung einer vorgegebenen Hypothese verglichen werden müssen. Dazu müssen sie zum einen in der Lage sein, die Hypothese in experimentelle Bedingungen umzusetzen. Zum anderen müssen sie wissen, dass es in der Gestaltung experimenteller Bedingungen wichtig ist, die zu untersuchende Variable zu variieren und die übrigen Bedingungen bzw. Variablen konstant zu halten.

Zur Messung der Fertigkeit *P* wurde der Itemstamm wie schon bei der Fertigkeit *H* so konstruiert, dass die Testpersonen anhand eines kurzen Textes in die Thematik eingeführt werden. Zusätzlich bekommen sie bereits einige grundsätzliche Informationen zur Planung des Versuchs. Die entscheidenden Schritte der experimentellen Planung müssen die Testpersonen anhand einer Entscheidung für den Vergleich bestimmter experimenteller Bedingungen leisten.

Drei der zehn Items zur Verdeutlichung der dargestellten Inhalte enthalten eine erläuternde Abbildung im Itemstamm. Im Anschluss an die Einleitung in das Szenario stellen alle Items vier experimentelle Bedingungen dar, gefolgt von der Frage, welche zwei der vier zur Auswahl vorgegebenen Bedingungen die Wissenschaftlerinnen und Wissenschaftler zur Prüfung ihrer Vermutung vergleichen müssen. Die Darstellung der experimentellen Bedingungen erfolgt bei allen Items in bildlicher, graphischer oder tabellarischer Form. Diese Art der Darstellung vermeidet eine umständliche Erklärung der Bedingungen und verringert den Leseaufwand. Um die Frage nach den

Regenbogenforelle II

Um die Vermutung zu prüfen, dass Forellen schneller wachsen, wenn sie ein spezielles Zucht-futter (Pellets) bekommen, führen Wissenschaftler einen Versuch durch. Dazu werden zwei Forellengruppen verglichen, die unterschiedlich ernährt wurden.

Forellengruppe 1: Nahrung besteht aus Pellets, einem speziellen Zuchtfutter, das aus gepresstem Fischmehl, Blutmehl, Fischöl und einem meist pflanzlichen Bindemittel besteht.

Forellengruppe 2: Nahrung besteht aus dem natürlichen Futter, das den Forellen normalerweise zur Verfügung steht und das aus Insektenlarven, Würmern und kleinen Fischen besteht. .

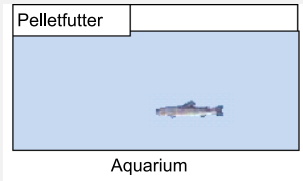
Beschreibung der Wissensbasis der Wissenschaftler/innen

+

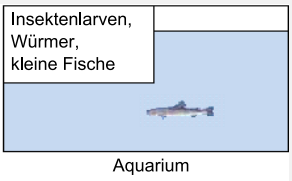
Darstellung des Versuchs

Welche beiden Versuchsbedingungen müssen die Wissenschaftler vergleichen, um ihre Vermutung zu prüfen?

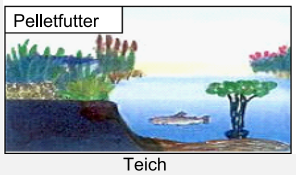
1)



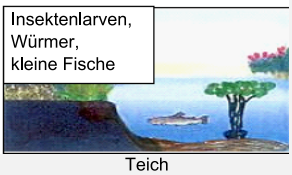
2)



3)



4)



Aufgabe

Aufgabenstamm

a) 1 und 2
b) 1 und 3
c) 2 und 4
d) 3 und 4

Antwortalternativen

Abbildung 4.5: Itembeispiel Fertigkeit P

beiden zu vergleichenden experimentellen Bedingungen beantworten zu können, erhalten die Testpersonen in diesem Bereich vier unterschiedliche Kombinationen von experimentellen Bedingungen, die zur Prüfung der Vermutung verglichen werden sollen. Folgende Kombinationen werden als Distraktoren eingesetzt:

- Vergleich zweier Bedingungen mit je zwei Merkmalen XY (mit je zwei Ausprägungen 1 und 2), wobei Y das zu variierende Merkmal ist.
  - Lösung: Vergleich X_1Y_1 und X_1Y_2
  - Distraktor 1: Vergleich X_1Y_1 und X_2Y_1

  - Distraktor 2: Vergleich X_1Y_2 und X_2Y_2
  - Distraktor 3: Vergleich X_2Y_1 und X_2Y_2

- Vergleich zweier Bedingungen mit je zwei Merkmalen X (mit drei Ausprägungen) und Y (mit zwei Ausprägungen), wobei Y das zu variierende Merkmal ist.
  - Lösung: Vergleich X_1Y_1 und X_1Y_2
  - Distraktor 1: Vergleich X_1Y_1 und X_2Y_2
  - Distraktor 2: Vergleich X_1Y_2 und X_2Y_2
  - Distraktor 3: Vergleich X_3Y_1 und X_2Y_2

- Vergleich zweier Bedingungen mit je drei Merkmalen XYZ (mit je zwei Ausprägungen), wobei Z das zu variierende Merkmal ist.
  - Lösung: Vergleich $X_1Y_1Z_1$ und $X_1Y_1Z_2$
  - Distraktor 1: Vergleich $X_1Y_1Z_1$ und $X_2Y_2Z_2$
  - Distraktor 2: Vergleich $X_1Y_1Z_1$ und $X_1Y_2Z_1$
  - Distraktor 3: Vergleich $X_1Y_1Z_2$ mit $X_2Y_2Z_2$

Auch im Bereich dieser Fertigkeit ist das Lesen des Textes notwendig, aber nicht hinreichend, um die Items zu lösen. Eine Lösung erfordert das Umsetzen der Hintergrundinformationen in Kombination mit der Fertigkeit, die experimentellen Bedingungen so zu wählen, dass die zu untersuchende Variable variiert und die übrigen Variablen konstant gehalten werden. Abbildung 4.5 dient der Verdeutlichung dieser Ausführungen.

FERTIGKEIT N: „DIE AUFGRUND GEGEBENER ERGEBNISSE UND IM HINBLICK AUF DIE GEGEBENE HYPOTHESE RICHTIGE SCHLUSSFOLGERUNG WIRD ERKANNT.“

Durch die Beantwortung von Items dieses Bereichs zeigen die Testpersonen, ob sie in der Lage sind, aus einer Auswahl möglicher Schlussfolgerungen diejenige zu erkennen, die das Ergebnis des vorgegebenen Experiments richtig interpretiert und die es gleichzeitig in eine korrekte Verbindung mit der vorgegebenen Hypothese bringt. Die Testperson muss also wissen, dass für eine Entscheidung hinsichtlich des Beibehaltens oder Verwerfens einer Hypothese das experimentelle Ergebnis auf die zuvor aufgestellte Hypothese zurückbezogen werden muss.

Zur Messung der Fertigkeit *N* wurde der Itemstamm wie bei den anderen Fertigkeiten so konstruiert, dass die Testpersonen anhand eines kurzen Textes in die Thematik eingeführt werden. Desweiteren bekommen sie Informationen über die Vermutung, welcher die Wissenschaftlerinnen und Wissenschaftler nachgegangen sind, sowie über die Ergebnisse des Experiments, welches die Wissenschaftlerinnen und Wissenschaftler durchgeführt haben. Gefolgt werden diese Informationen von der Frage, welche Schlussfolgerung aus den Ergebnissen im Hinblick auf die Hypothese gezogen werden kann. Die Testpersonen müssen also die Verbindung zwischen dem Ergebnis des Experiments und der gegebenen Hypothese herstellen.

Ebenso wie bei der Itementwicklung des Fertigkeitsbereichs *P*, in dem die Darstellung der experimentellen Bedingungen möglichst kurz in Form von Graphiken u.Ä. erfolgte, war es bei der Entwicklung der Items dieses Bereichs wichtig, die Ergebnisse der durchgeführten Versuche möglichst kurz und ohne weiteren Text darzustellen. Um die Ergebnisse der Versuche kurz und textfrei darzustellen, wurden Graphiken zur Veranschaulichung eingesetzt. Bei sieben der zehn Items wurde die Darstellungsform von Balkendiagrammen gewählt, da diese Art der Darstellung für Schülerinnen und Schüler am leichtesten zu interpretieren ist (vgl. Abschnitt 2.5). Im Falle der übrigen Items erforderte die Art der Informationen eine Darstellung in Form von Kurvendiagrammen. Wie Abbildung 4.6 zeigt, erhalten die Testpersonen in diesem Bereich zur Lösung der Aufgabe vier mögliche Schlussfolgerungen. Die Distraktoren sehen strukturell folgendermaßen aus:

- Der Distraktor enthält eine richtige Schlussfolgerung in Bezug auf die Hypothese, aber eine falsche Interpretation der Graphik.
- Der Distraktor enthält eine falsche Schlussfolgerung trotz richtiger Interpretation der Graphik.
- Der Distraktor enthält eine falsche Schlussfolgerung und eine falsche Interpretation der Graphik.
- Der Distraktor enthält eine unwissenschaftliche Schlussfolgerung, die sich nicht nach den dargestellten Ergebnissen richtet, sondern beispielsweise nach dem Gefühl der Wissenschaftlerinnen und Wissenschaftler.
- Der Distraktor enthält den Vorschlag, dass die Graphik keine Aussage bezüglich des Beibehaltens oder Verwerfens der Hypothese zulässt.

  Bemerkung: die beiden zuletzt genannten Distraktoren treten nicht gemeinsam auf.

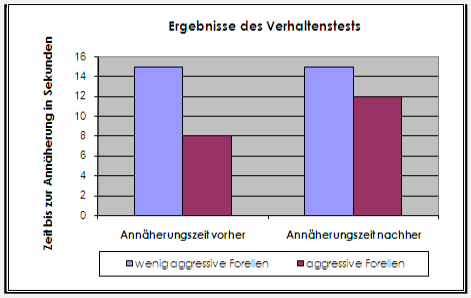
Regenbogenforelle III In einem Versuch zum Verhalten von Forellen ließen Wissenschaftlerinnen Forellen andere Forellen beim Kämpfen beobachten. Dabei beobachteten wenig aggressive das Verhalten von aggressiven Forellen und aggressive das Verhalten von wenig aggressiven Forellen. wenig aggressive Forelle —beobachtet→ aggressive Forelle aggressive Forelle —beobachtet→ wenig aggressive Forelle Aufgrund vorheriger Beobachtungen haben die Wissenschaftlerinnen die Vermutung, dass aggressive Forellen durch die Beobachtung wenig aggressiver Artgenossen an Aggressivität verlieren. Das Verhalten wenig aggressiver durch die Beobachtung aggressiver Forellen nicht ändern. Folgende Abbildung zeigt das Ergebnis, das die Wissenschaftlerinnen festhalten konnten:	Beschreibung der Wissensbasis der Wissenschaftler/ innen + Darstellung des durchgeführten Versuchs	Aufgabenstamm
	Darstellung der Ergebnisse	
Was bedeuten die dargestellten Ergebnisse für die Vermutung der Wissenschaftlerinnen?	Aufgabe	
a) Die Vermutung hat sich nicht bestätigt, da es logisch ist, dass aggressive Forellen ihre Aggressivität nicht verlieren.. b) Die Vermutung hat sich bestätigt, da sich sowohl die aggressiven als auch die wenig aggressiven Forellen wie vorhergesagt verhalten haben. c) Die Vermutung hat sich bestätigt, da sich das Verhalten der wenig aggressiven Forellen nicht geändert hat. d) Die Vermutung hat sich nicht bestätigt, da sich das Verhalten der wenig aggressiven Forellen nicht geändert hat.	Antwortalternativen	

Abbildung 4.6: Itembeispiel Fertigkeit N

Im Bereich dieser Fertigkeit kommt zum Lesen des Textes noch das erfolgreiche Interpretieren der Graphiken als notwendige Bedingung hinzu, um die Items zu lösen. Um dem Leseverständnis und dem Graphikverständnis nicht zu viel Gewicht zukommen zu lassen, wurden die Texte so kurz und die Graphiken so einfach wie möglich gehalten. Hinreichend werden die Bedingungen zur Lösung erst, wenn die Testpersonen die Fertigkeit mitbringen, das Ergebnis des Experiments in richtiger Schlussfolgerung mit der Hypothese zu verbinden.

GENERIERUNG DER ANTWORTALTERNATIVEN

Die Generierung der Antwortalternativen stellte den abschließenden Punkt in der Entwicklung der Testitems dar, bevor sie anhand von Expertenurteilen und Ergebnissen erster Stichproben getestet und überarbeitet wurden. Ebenso wie bei der Entwicklung des Itemstamms war es bei der Entwicklung der Antwortalternativen wichtig, Ausdrücke und Formulierungen zu wählen, die den Testpersonen vertraut sind. Die Lösung eines Items durfte nicht dadurch unmöglich werden, dass Distraktoren attraktiver erscheinen und mit größerer Wahrscheinlichkeit gewählt werden als die richtige Antwortalternative.

Ein wirkungsvolles Verfahren zur Findung möglicher Antwortalternativen stellen *Cognitive-Lab-Interviews* dar, die bereits in Abschnitt 3.3.4 näher beschrieben wurden. Die Items werden den Schülerinnen und Schülern im Rahmen dieses Verfahrens im offenen Antwortformat vorgelegt. Sie erhalten die Aufgabe, bei der Bearbeitung der Items laut zu denken und dann eigene Antworten auf die Fragen zu formulieren. Aus dem Pool der formulierten Antworten werden später die notwendigen Formulierungen für die Entwicklung der Antwortalternativen übernommen. Es ist allerdings notwendig, bereits im Vorfeld Vorstellungen über die Struktur der Antwortalternativen zu entwickeln. Somit erfolgte die Generierung der Antwortalternativen durch eine Kombination aus vorformulierter Struktur und den Informationen und Ausdrücken, die im Rahmen der *Cognitive-Lab-Interviews* gewonnen wurden.

Aufgrund der zeitlichen Begrenzung konnten im Falle dieser Testentwicklung die *Cognitive-Lab-Interviews* zur Gewinnung der Antwortalternativen lediglich in eingeschränkter Form durchgeführt werden. Es wurde daher ein eher pragmatisches Vorgehen gewählt. Die Struktur und die Formulierungen der Antwortalternativen wurden bereits im Vorfeld festgelegt. Die Aufgaben wurden anschließend mit den vorformulierten Antwortalternativen in *Cognitive-Lab-Interviews* mit zwei Schülerinnen getestet. In Einzelsitzungen bekamen sie die Aufgabe, die Testitems zu bearbeiten und ihre Gedanken dabei zu verbalisieren. Das Ergebnis der Interviews umfasste sowohl

Kommentare zum Itemstamm als auch zu den Antwortalternativen. Im Folgenden ist ein Beispiel eines Kommentars aufgeführt, der sowohl zu einer Überarbeitung des Itemstamms als auch zu einer genaueren Betrachtung der Antwortalternativen geführt hat. Um einen besseren Überblick zu gewährleisten, stellt Abbildung 4.7 das Item zunächst in der Form dar, in der es den Schülerinnen vorgelegt wurde.

Bis ins 18. Jahrhundert hinein war Skorbut die häufigste Todesursache bei Seeleuten. Sie waren oft lange unterwegs und ernährten sich während dieser Zeit hauptsächlich von Pökelfleisch und Schiffszwieback. Nach Zahnfleischbluten und Zahnausfall bekamen die Skorbut-Kranken hohes Fieber und starben schließlich. Erst der britische Schiffsarzt James Lind untersuchte 1754 in einer für damalige Verhältnisse sehr modernen Studie die Ursache für Skorbut.

James Lind nahm zwölf Seeleute in seine Studie auf, bei denen die Krankheit gleich weit fortgeschritten war. Er behandelte seine Patienten in einem speziellen Raum unter Deck mit sechs unterschiedlichen Behandlungsmethoden. Unter anderem hatte er gehört, dass die Besatzung des Schiffskapitäns Lord Anson die Krankheit auf einer Insel durch das Essen von Orangen überwunden hatten.
Für die Untersuchung ordnete Lind jeder Behandlungsmethode zwei Personen zu. Folgende unterschiedliche Behandlungen wandte Lind an:

	Behandlung	Anzahl der Patienten
1	Apfelwein	2
2	Mineralstoffhaltiges Wasser	2
3	Essig	2
4	Meerwasser	2
5	Zitrusfrüchte (Orangen und Zitronen)	2
6	Medizin mit Wasser und Honig vermischt	2

Welcher Vermutung ist James Lind in seiner Studie im Jahre 1754 nachgegangen?

a) Skorbut entsteht deshalb, weil die Ernährung auf See so schlecht ist.
b) Skorbut entsteht deshalb, weil die Seeleute so lange auf See unterwegs sind.
c) Skorbut kann durch die Verabreichung von Zitrusfrüchten geheilt werden.
d) Skorbut wird am besten durch die Behandlung mit Apfelwein behandelt.

Abbildung 4.7: Itembeispiel *Skorbut*

Die folgenden Kommentare äußerte die Realschülerin bei Bearbeitung dieses Items:

Schülerin: Ich habe *a* genommen, *c* kann auch stimmen. Ich wusste schon vorher, woran es liegt [dass die Seeleute an Skorbut erkranken]. Also wenn sie sich schon vorher

vernünftig ernährt hätten, wären sie nicht krank geworden, also ist es, weil die Ernährung so schlecht ist.
Testautorin: Sind sie wirklich dieser Vermutung nachgegangen? Eigentlich soll *c* die richtige Antwort sein. (Die Schülerin hatte auch zunächst Antwort *c* ausgewählt)
Schülerin: Dachte ich auch zuerst.
Testautorin: Was war dein erster Lösungsansatz?
Schülerin: Na ja, eigentlich überlegt er [James Lind] ja, wie man sie [die Seeleute] heilen kann.
Testautorin: Warum bist du auf *a* umgeschwenkt?
Schülerin: Vielleicht, weil das auch richtig ist.

Am Beispiel dieser Kommentare ist zu erkennen, dass sich die Schülerin letztendlich für die Antwortalternative entschieden hat, die zwar grundsätzlich keine falsche Aussage trifft, die aber nicht dem entspricht, was James Lind untersuchen wollte. Hier wird deutlich, dass der Distraktor *a* zu nahe am Alltagswissen der Schülerin liegt. Sie hat schon einmal gehört, dass Skorbut unter Seeleuten damals durch die schlechte Ernährung auf den Schiffen entstand und hat diese Antwortalternative gewählt, weil die Aussage an sich richtig ist. Die Aussage stellt allerdings nicht die richtige Antwort auf die Frage dar, welcher Vermutung James Lind nachgegangen ist. Die Schülerin wurde durch eine zu ihrem Vorwissen passende Antwort von der Lösung abgelenkt. Sie hatte sich – das zeigt das Gespräch nur in Ansätzen, wurde aber bei der Beobachtung der Testperson handschriftlich festgehalten – zunächst für die richtige Antwort *c* entschieden. Ihr Kommentar zeigt, dass sie die Hintergrundinformationen eigentlich richtig verarbeitet hat. An dieser Stelle wird deutlich, dass der Distraktor *a* überarbeitet werden muss. Er darf nicht zu nah am Alltagswissen der Schülerinnen und Schüler liegen, da er sonst mit zu hoher Wahrscheinlichkeit gewählt wird. Als Folge dieser Informationen wurde der Distraktor *a* folgendermaßen umformuliert: *„Skorbut entsteht deshalb, weil sich die Seeleute von Schiffszwieback ernähren."*

Sicherlich ist an dieser Stelle zu bedenken, dass die Informationen von lediglich zwei Schülerinnen nur einen kleinen Einblick geben können. Dennoch lieferten sie erste wertvolle Informationen für die Überarbeitung der Items und Antwortalternativen. Eine weitere Überarbeitung der Items fand anhand einer zweiten Runde von *Cognitive-Lab-Interviews* statt, die im Anschluss an den Feldtest durchgeführt wurde. Dieser Feldtest lieferte die ersten statistischen Kennwerte, die Aussagen darüber zu-

ließen, welche Antwortalternativen bereits sehr gut funktionierten und welche noch einer Überarbeitung bedurften.

4.4 PRÜFUNG UND WEITERENTWICKLUNG DER ITEMS

Die Items des *Tests zur Erfassung prozessbezogener naturwissenschaftlicher Grundbildung* durchliefen drei Überarbeitungsphasen bis zu der Endversion, die anhand des Haupttests geprüft wurde. Abbildung 4.8 gibt Auskunft darüber, welche Stationen die Itementwicklung umfasste. Der folgende Abschnitt beschreibt, welche Informationsquellen genutzt wurden und welcher Art die Informationen waren, die die Basis zur Überarbeitung der Testitems bildeten. Dabei wird auf die in der Abbildung aufgeführten Testphasen eingegangen. Das *Expertenpanel* stellt eine der wichtigen Informationsquellen zur Überarbeitung der Testitems dar. Die Expertengruppen setzten sich dabei je nach Stand der Testentwicklung unterschiedlich zusammen, waren aber in allen Fällen im Rahmen einer bestimmten Aufgabenstellung dazu aufgefordert, ihr Urteil direkt und in offenem Antwortformat auf jedes einzelne Aufgabenblatt einzutragen. Die Kommentare wurden zunächst pro Aufgabe gesammelt.

Ein weiteres wichtiges Verfahren stellten die *Cognitive-Lab-Interviews* mit Schülerinnen und Schülern dar. Sie wurden gebeten, die Testaufgaben zu bearbeiten und dabei alles zu verbalisieren, was ihnen dabei durch den Kopf ging. Angefangen von Verständnisproblemen bis hin zu Lösungsansätzen und Lösungswegen, welche die Schülerinnen und Schüler in Betracht zogen, durften sie alles äußern, ohne dass ihre Kommentare eine Bewertung erfuhren. Die Schülerkommentare wurden ebenfalls pro Aufgabe gesammelt. Welche Kommentare tatsächlich in die Änderung einer Aufgabe einflossen, wurde jeweils durch eine kombinierte Betrachtung der *Expertenurteile* und *Cognitive-Lab-Interviews* entschieden. Auf welche Weise diese Interviews eingesetzt wurden, wird in den Abschnitten zur Präpilotphase und zum Feldtest beschrieben.

Zusätzlich zu den beiden genannten Quellen wurden die bereits in Abschnitt 3.1.2 beschriebenen, statistischen Kennwerte herangezogen. In der Pilotphase hatten diese Kennwerte aufgrund der zu kleinen Stichprobe einen eher geringfügigen Einfluss auf die Überarbeitung der Items. Dies änderte sich im Rahmen des Feldtests. Hier wurde zunächst aufgrund der statistischen Kennwerte festgestellt, welche Items der Überarbeitung bedurften. Basierend auf diesen Informationen wurden *Expertenpanel* und *Cognitive-Lab-Interviews* eingesetzt, um Hypothesen im Hinblick auf mögliche Gründe für die ungenügenden Kennwerte zu generieren. Auf der Grundlage aller

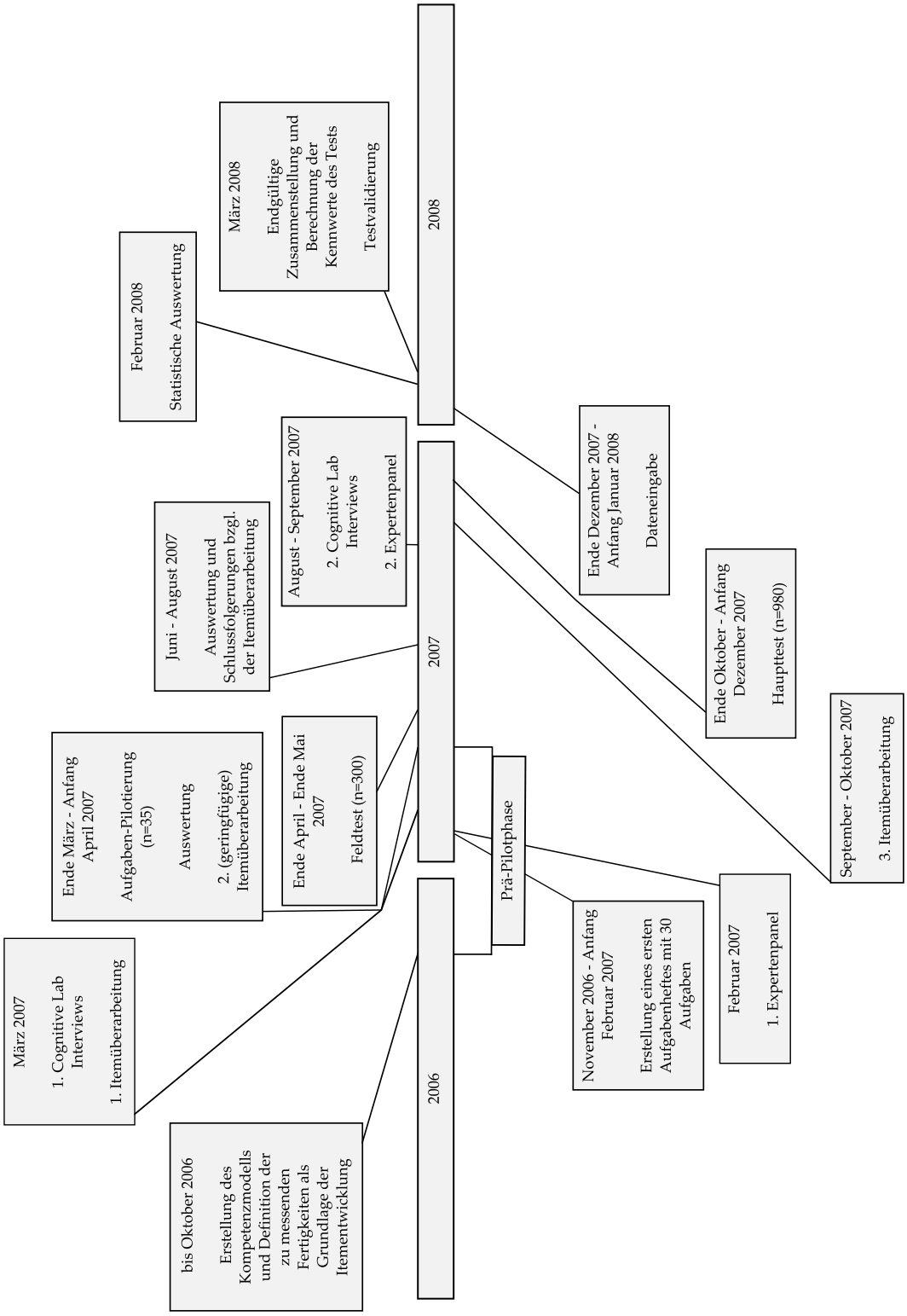


Abbildung 4.8: Zeitleiste der Testentwicklung

Informationen wurde abschließend entschieden, in welcher Weise Items überarbeitet werden sollten. Die gleichen Kennwerte wurden im Haupttest herangezogen, um die nicht funktionierenden Items bzw. Antwortalternativen anzuzeigen. Hier bestand die Konsequenz jedoch darin, die Items, deren Kennwerte trotz Überarbeitung noch immer ungenügende Kennwerte aufwiesen, endgültig aus dem Testheft zu entfernen.

Die folgenden Abschnitte gehen auf die einzelnen Testphasen und auf die Beschreibung der genannten Verfahren zur Item-Überarbeitung ein.

4.4.1 PRÄPILOTPHASE

Zunächst war vor der Entwicklung der ersten Aufgaben wichtig, ein Modell der zu messenden Kompetenz zu erstellen und die zu messenden Fertigkeiten zu definieren. Nach dieser grundlegenden theoretischen Arbeit wurde ein erstes Aufgabenheft mit 30 Aufgaben erstellt. Im Anschluss daran fand in der Präpilotphase ein erstes *Expertenpanel* und die ersten *Cognitive-Lab-Interviews* statt, deren Informationen in eine erste Überarbeitung der Aufgaben mündete.

EINGESETZTE VERFAHREN

Im Rahmen des ersten *Expertenpanels* wurde die sachliche Richtigkeit der Items geprüft und entschieden, ob die Inhalte für Schülerinnen und Schüler der neunten Klasse verständlich, in der Schwierigkeit angemessen und thematisch interessant sind. Nachdem die Expertinnen und Experten individuell ihre Bemerkungen zu den einzelnen Items verfasst hatten, wurde eine gemeinsame Expertenrunde durchgeführt. Die Diskussion der Expertinnen und Experten wurde in Stichpunkten festgehalten und zusätzlich per Tonband aufgezeichnet, um einen Verlust von Informationen zu vermeiden. Die Anregungen zur Überarbeitung der Items wurden zunächst pro Aufgabe gesammelt. Die Entscheidung, welche dieser Anregungen zu einer Veränderung der Items führen sollten, wurde erst nach den *Cognitive-Lab-Interviews* mit den Schülerinnen der Zielgruppe getroffen.

Die *Cognitive-Lab-Interviews* wurden mit zwei Schülerinnen, einer Realschülerin (15 Jahre) und einer Hauptschülerin (14 Jahre), einzeln durchgeführt. Die beiden Schülerinnen bearbeiteten alle Testitems und erhielten die Aufgabe, ihre Denkprozesse zu verbalisieren und angefangen von Verständnisproblemen bis hin zu Lösungsansätzen und -wegen alles zu begründen, was zur Wahl ihrer jeweiligen Antwort geführt hatte. Ihre Kommentare wurden zum einen handschriftlich in Stichpunkten festgehalten, zum anderen wurden sie per Tonband aufgenommen. Wenn sich die Schü-

lerinnen nicht weiter zum Lösungsprozess äußerten, wurden Nachfragen dazu gestellt, wie sie zu ihrer Antwort kamen und wie sie den Schwierigkeitsgrad der Aufgabe einschätzten. Außerdem wurden Anmerkungen der Expertinnen und Experten dazu verwandt, um spezielle Nachfragen, beispielsweise zum Verständnis bestimmter Ausdrücke, zu stellen. Wichtig waren bei der Sammlung der Kommentare insbesondere diejenigen Punkte, welche die Schülerinnen nicht verstanden, die zu Missverständnissen und zur Wahl einer falschen Antwortalternative führten. Diese Fälle führten ausnahmslos zu einer Überarbeitung der Items. Deuteten die Kommentare der Schülerinnen jedoch darauf hin, dass fehlende Kompetenzen im Bereich der *prozessbezogenen naturwissenschaftlichen Grundbildung* für die falsche Beantwortung verantwortlich waren (beispielsweise durch eine wissenschaftlich falsche Begründung), so wurden keine Änderungen vorgenommen. Es wurde in diesen Fällen lediglich in Betracht gezogen, dass der Schwierigkeitsgrad des Items zu hoch sein könnte. Alle Eindrücke, die im Zusammenhang mit der Schwierigkeit der Items zu tun hatten, wurden später anhand der statistischen Kennwerte verifiziert.

In einem dritten Schritt wurden die Kommentare der Expertinnen und Experten und der Schülerinnen Aufgabe für Aufgabe zusammengeführt. Es wurden alle vorliegenden Informationen betrachtet und danach entschieden, in welcher Weise die Items überarbeitet werden sollten. Sachlich falsche Ausdrücke und Inhalte führten ebenso ausnahmslos zu einer Überarbeitung wie missverständliche oder den Schülerinnen fremde Ausdrücke. Ebenfalls führten übereinstimmende Kommentare von Expertinnen und Experten und Schülerinnen, beispielsweise im Hinblick auf zu lange Texte oder komplizierte und unverständliche Ausdrücke in jedem Fall zu einer Überarbeitung. Auf diese Weise wurden alle Items betrachtet und überarbeitet, bevor sie abschließend in einem Testheft zusammengestellt und im Rahmen der Pilotierung einer ersten Stichprobe vorgelegt wurden.

4.4.2 PILOTIERUNG

Die Pilotierung hatte vorrangig das Ziel, die spätere Testung der Items in Feld- und Haupttest im Vorfeld zu simulieren. Es sollte zum einen festgestellt werden, ob der angesetzte zeitliche Rahmen einer Doppelstunde für die Bearbeitung der Aufgaben ausreichend ist. Desweiteren diente sie zur Überprüfung der einleitenden Worte und Erklärungen. Aus allen gesammelten Kommentaren ergaben sich noch einige geringfügige Veränderungen des Aufgabenheftes, insbesondere im Bereich der Einleitung.

MERKMALE DER STICHPROBE

Für die Pilotierung der Aufgaben wurde eine Gesamtschule ausgewählt, um die Testung möglichst ökonomisch durchzuführen und von Beginn an ein relativ breites Leistungsniveau in den Stichproben zu sichern. Auf diese Weise konnte von einem der späteren Teststichprobe ähnlichen Leistungsniveau ausgegangen werden.

Die Pilotierungsstichprobe bestand aus zwei neunten Klassen: Die eine Klasse umfasste Schülerinnen und Schüler mit Hauptschul- und unterem Realschulniveau, die andere Klasse Schülerinnen und Schüler mit oberem Realschul- und Gymnasialniveau. Die Stichprobe bestand aus 35 Schülerinnen und Schülern. Es wurden keine weiteren persönlichen Informationen oder Variablen im Rahmen der Pilotierung erfasst, da zu diesem Zeitpunkt der Testentwicklung weniger die statistischen Auswertungen als vielmehr die Testung der Durchführung und der allgemeinen Abläufe im Vordergrund standen.

EINGESETZTE VERFAHREN

Die Verfahren, die im Rahmen der Pilotierung zum Einsatz kamen, bestanden hauptsächlich in Beobachtungen der Schülerinnen und Schüler und Aufzeichnungen der Schülerkommentare. Statistische Verfahren wurden in dieser Phase lediglich dazu genutzt, um einen ersten Eindruck der Itemschwierigkeiten und Trennschärfen zu gewinnen. Diese Berechnungen konnten nur als erste Indikatoren gelten, da die Stichprobe mit 35 Testpersonen zu klein war und sich tiefergehende Analysen dadurch verbaten.

Während der Pilotierung wurde den Schülerinnen und Schülern anhand eines strukturierten Leitfadens zunächst der Zweck der Erhebung erklärt. Im Anschluss an diese Einführung wurden sie gefragt, inwiefern es Verständnisschwierigkeiten oder Unklarheiten gab. War dies der Fall, so wurden diese notiert und leichter verständliche Erklärungen zur Beschreibung des Untersuchungsziels gefunden.

Während der Bearbeitung der Testaufgaben wurden die Schülerinnen und Schüler beobachtet, und es wurden alle Reaktionen, die sich in irgendeiner Form auf das Aufgabenheft bezogen, notiert. Die Schülerinnen und Schüler äußerten sich hauptsächlich zum Umfang des Aufgabenheftes und zum großen Leseaufwand. Diese Kommentare führten zum einen dazu, für die zukünftigen Untersuchungen Formulierungen zu finden, die derartigen Kommentaren entgegenwirken konnten. Zum anderen führten sie dazu, dass in der zweiten Überarbeitung der Aufgaben noch einmal versucht wurde, Aufgabentexte zu kürzen.

Insgesamt sollten sich die Schülerinnen und Schüler offen zu allem äußern, was ihnen während der Bearbeitung der Aufgaben gefiel oder missfiel, was sie interessant oder uninteressant und was sie verständlich oder unverständlich fanden. Die Kommentare wurden in zwei Kategorien *„bei der Überarbeitung des Aufgabenhefts beachten“* und *„bei der Überarbeitung des Aufgabenhefts nicht beachten“* eingeteilt. Dementsprechend kam es zu einer zweiten, geringfügigen Überarbeitung des Aufgabenheftes, bevor die Items im Rahmen des Feldtests einer größeren Stichprobe vorgelegt wurden.

4.4.3 FELDTEST

Der Feldtest fand von April bis Mai 2007 statt. Er hatte das Ziel, anhand einer ausreichend großen Stichprobe einen ersten Überblick über die relevanten statistischen Kennwerte zu ermöglichen. Die Berechnungen der Kennwerte dienten dazu, Items zu identifizieren, die durch unzureichende oder nicht homogene Trennschärfen oder aber problematische Antwortalternativen auffielen. Diesem Schritt folgte eine inhaltliche Betrachtung der unzureichenden Items und einer Hypothesenbildung hinsichtlich der Gründe für ihr ungenügendes Funktionieren.

Im Folgenden wird ausgeführt, auf welche Weise die statistischen Kennwerte, ein weiteres *Expertenpanel* und weitere *Cognitive-Lab-Interviews* zur Weiterentwicklung der Items genutzt wurden. Zunächst wird die Stichprobe charakterisiert, die die Items im Rahmen des Feldtests bearbeitete, bevor anschließend genauer auf die Umsetzung der einzelnen Verfahren eingegangen wird.

MERKMALE DER STICHPROBE

Die Stichprobe des Feldtests bestand zu Beginn aus 402 Schülerinnen und Schülern. In einem ersten Schritt wurden zunächst diejenigen Testpersonen aus der Stichprobe entfernt, deren Antwortverhalten an einer ernsthaften Bearbeitung zweifeln ließen (z.B. durch ständiges Ankreuzen mehrerer Antworten) oder aber die mehr als zehn Items unbeantwortet gelassen haben. Dabei kann es viele Gründe für das Nicht-Beantworten der Aufgaben geben, die von fehlender Ernsthaftigkeit bei der Bearbeitung bis hin zu Leseschwierigkeiten oder Verständnisschwierigkeiten reichen können. Da in diesen Fällen die genauen Gründe nicht verifizierbar sind und da es wichtig ist, möglichst viele komplette Datensätze zur Berechnung der Itemparameter zur Verfügung zu haben, wurden 13 Datensätze aufgrund der genannten Kriterien aus dem Datenpool entfernt. Auf diese Weise blieben schließlich 389 Datensätze üb-

rig. Aus diesen 389 Datensätzen wurden nach dem Zufallsprinzip 105 Datensätze pro Schulform ausgewählt, um die Gesamtstichprobe für die Ermittlung der statistischen Kennwerte zu bilden. Für die probabilistischen Analysen wären ungleiche Stichprobengrößen unbedeutend, doch um für die Berechnung klassischer statistischer Kennwerte möglichst günstige Voraussetzungen zu schaffen, wurden gleiche Stichprobengrößen angestrebt. Die Stichprobe des Feldtests bestand letztendlich aus insgesamt 315 Schülerinnen und Schüler, je 105 aus Haupt-, Realschule und Gymnasium. Diese Schülerinnen und Schüler verteilten sich auf insgesamt 14 neunte Klassen (5 Hauptschul-, 4 Realschul- und 5 Gymnasialklassen).

Abgesehen vom Schulniveau wurden keine persönlichen Daten der Schülerinnen und Schüler erhoben. Erst im Rahmen des Haupttests wurden weitere Variablen erhoben, um erste Validierungsschritte vornehmen zu können.

EINGESETZTE VERFAHREN

Anhand der Daten aus dem Feldtest wurden zunächst die bereits in Abschnitt 3.1.2 beschriebenen statistischen Kennwerte berechnet. An dieser Stelle soll allerdings noch keine tiefer gehende, statistische Diskussion eröffnet, sondern vielmehr verdeutlicht werden, welche Rolle die statistischen Kennwerte in der Weiterentwicklung der Items gespielt haben. Ungenügende Kennwerte führten nicht zu einer Entfernung, sondern lediglich zu einer Überarbeitung der betreffenden Items. Um am Ende des Entwicklungsprozesses eine ausreichend große Anzahl an Items zu erhalten, wurden sie erst dann aus dem Aufgabenheft entfernt, wenn sie auch nach einer weiteren Überarbeitung im Haupttest ungenügende Werte zeigten.

Um die zu überarbeitenden Items zu identifizieren, wurden Itemtrennschärfen sowie Itemschwierigkeiten berechnet und mit festgelegten Qualitätskriterien verglichen. Weiterhin wurden die Antwortalternativen daraufhin untersucht, ob die meisten Antworten der Stichprobe auf die richtige Antwortalternative entfallen und die übrigen Antworten sich relativ gleichmäßig auf die Distraktoren verteilen. Die richtige Alternative sollte von Personen mit der höchsten geschätzten Fähigkeit gewählt werden. Wählten diese Personen einen der Distraktoren bzw. wählten Personen mit geringeren Fähigkeitswerten eher die Lösung als diejenigen mit hohen Fähigkeitswerten, so wurde dies als Zeichen dafür gesehen, dass das Item oder zumindest die Antwortalternativen einer Überarbeitung bedurften. Auf diese Weise wurden die statistischen Kennwerte als Detektoren für nicht funktionierende Items und Antwortalternativen genutzt.

Die statistischen Ergebnisse des Feldtests dienten als Grundlage, um Ideen und

Hypothesen im Hinblick auf die Gründe zu sammeln, die dazu geführt haben, dass Items in ihrer Gesamtheit oder aber einzelne Antwortalternativen der Items nicht funktionierten. Diese Hypothesenbildung erfolgte zunächst durch die Testautorin selbst. In weiteren Schritten wurden erneut ein Expertenpanel und *Cognitive-Lab-Interviews* mit Schülerinnen und Schülern als Informationsquellen herangezogen, um Ansätze für die Verbesserung der Items zu finden.

Die Zusammensetzung und das Vorgehen des zweiten *Expertenpanels* unterschied sich deutlich vom ersten Expertenpanel. Wie bereits in Abschnitt 4.2.4 ausgeführt wurde, setzte sich die Expertengruppe in dieser zweiten Runde heterogener zusammen, um möglichst viele kreative Ideen zu sammeln, welche die unzureichende Qualität einzelner Items oder Antwortalternativen erklären konnten. Anders als im ersten *Expertenpanel* wurden den Expertinnen und Experten hier Informationen über die Items vorgelegt. Sie bekamen eine Liste der Items, die sich durch unzureichende statistische Kennwerte auszeichneten. Die Liste enthielt neben allen relevanten statistischen Kennwerten eine kurze Beschreibung der Itemeigenschaften (z.B. *„Item ist zu leicht“*, *„Personen mit der höchsten Fähigkeit wählen die falsche Antwortalternative a“*).

Die Aufgabe bestand darin, Gründe für das ungenügende Funktionieren der Items bzw. ihrer Antwortalternativen zu finden. Das Verfahren der Informationssammlung erfolgte dabei wie beim ersten Panel. Im Anschluss an diese Phase der Ideengenerierung folgten drei Treffen mit Expertinnen und Experten: ein Treffen mit der Lehrerin und dem Lehrer, ein Treffen mit den beiden Studenten und ein Treffen mit der Architektin und der Buchhändlerin. In diesen Treffen wurden die Kommentare zu den einzelnen Aufgaben detailliert erörtert. Alle Kommentare wurden gesammelt und später mit den Ergebnissen der *Cognitive-Lab-Interviews* zusammengeführt.

Die Cognitive-Lab-Interviews wurden im Gegensatz zu den ersten Interviews dieser Art mit einer Gruppe, bestehend aus zwei Realschülerinnen und einem Realschülern, durchgeführt. Diese bearbeiteten individuell die Aufgaben. Im Anschluss an die Bearbeitung einer jeden Aufgabe wurden die Schülerinnen und Schüler nacheinander gefragt, wie sie zu ihrer Antwort gelangt sind und ob Verständnisschwierigkeiten oder andere Schwierigkeiten bei der Lösung der Aufgabe aufgetreten sind. Die Nachfragen hatten den Zweck, sicherzustellen, dass alle für eine Itemüberarbeitung notwendigen Informationen gesammelt werden konnten. Nachdem die drei Testpersonen sich geäußert hatten, folgte eine offene Diskussion zur Aufgabe, in der sie sich über ihre differierenden Antworten oder Meinungen austauschen konnten. War diese Diskussion beendet, so wurde den Schülerinnen und dem Schüler die richtige Lösung genannt. Sie wurden gebeten, sich dahingehend zu äußern, inwiefern sie die Lösung

für logisch hielten oder was sie daran eventuell nicht verstanden haben bzw. warum sie sich aber dennoch anders entschieden hätten. Alle Schülerkommentare wurden in Stichpunkten notiert und zusätzlich per Tonband festgehalten. Die Prozedur wurde für jedes nicht funktionierende Item wiederholt.

Wie schon in der ersten Sammlung von Experten- und Schülerkommentaren, wurden auch in diesem Fall alle gesammelten Aussagen pro Aufgabe betrachtet, um im Anschluss zu entscheiden, auf welche Weise die Aufgaben und Antwortalternativen überarbeitet werden sollten. Durch die Kombination von Expertenideen und Schülerkommentaren wurde die Wahrscheinlichkeit einer tatsächlichen Verbesserung der Items erhöht.

4.4.4 HAUPTTEST

Nach der im Anschluss an den Feldtest durchgeführten dritten Itemüberarbeitung, folgte von Oktober bis Dezember 2007 der Haupttest. Er stellte im Rahmen dieser Dissertation die abschließende Prüfung der einzelnen Items und des Tests als Ganzes dar und hatte weiterhin die Aufgabe, die ersten Validierungsschritte einzuleiten.

MERKMALE DER STICHPROBE

Die Stichprobe, die für den Haupttest herangezogen wurde, bestand zunächst aus 992 Schülerinnen und Schülern (w=493, m=499), die sich folgendermaßen auf die drei Schulformen aufteilten: 312 Hauptschülerinnen und Hauptschüler (w=155, m=157), 387 Realschülerinnen und Realschüler (w=187, m=200) und 293 Gymnasiastinnen und Gymnasiasten (w=151, m=142). Die Aufteilung auf Schulklassen sah folgendermaßen aus: 17 Hauptschulklassen (aus 7 Schulen), 15 Realschulklassen (aus 4 Schulen) und 14 Gymnasialklassen (aus 3 Schulen).

Bevor die eigentlichen Berechnungen und Verfahren, die im anschließenden Abschnitt beschrieben werden, zum Einsatz kamen, wurde die Stichprobe zunächst um Datensätze gekürzt, die nicht die nötige Vollständigkeit aufwiesen (fehlende Werte $n \geq 10$) oder bei denen Zweifel an einer gewissenhaften Bearbeitung aufkamen (z.B. ständiges Ankreuzen mehrerer Antworten).

Nach Bereinigung blieben 967 Datensätze übrig, aus denen nach dem Zufallsprinzip 250 Datensätze pro Schulform ausgewählt wurden, um die Gesamtstichprobe zu bilden. Auch hier wurden - wie bereits beim Feldtest geschehen - gleiche Stichprobengrößen in den Schulformen angestrebt, um für die Berechnung klassischer statistischer Kennwerte möglichst günstige Voraussetzungen zu schaffen.

Am Ende setzte sich die Haupttest-Stichprobe aus 750 Schülerinnen und Schülern (w=375, m=373, fehlende Angabe=2) zusammen, 250 Schülerinnen und Schüler pro Schulform.

EINGESETZTE VERFAHREN UND ERHOBENE VARIABLEN

Im Haupttest wurden die gleichen statistischen Kriterien herangezogen wie bereits im Feldtest. Anders als im Feldtest führte die Ermittlung der statistischen Kennwerte jedoch *nicht* zu einer weiteren Überarbeitung der Items. Vielmehr wurden Items, die zu diesem Zeitpunkt und trotz der vorangegangenen Überarbeitungen noch immer keine verbesserten Werte zeigten, nun endgültig aus dem Testheft entfernt. Erst im Anschluss an die Entfernung der Items wurde das Aufgabenheft endgültig zusammengestellt und mit der Validierung begonnen.

An dieser Stelle kamen die bereits in Abschnitt 3.2.1 genannten Verfahren zur Validierung zum Einsatz. Im Falle der *internen Validierung* erfolgte also die Prüfung des Infits, die Überprüfung von Gruppenunterschieden, DIF-Analysen sowie die globale Modellprüfung. Im Falle der externen Validierung wurden die erhobenen Testleistungen mit externen Leistungs- und Motivationskriterien in Beziehung gesetzt, deren Operationalisierung im folgenden Abschnitt beschrieben wird.

Mit dieser Beschreibung werden das Operationalisierungskapitel und die Darstellung der Testentwicklungsphasen abgeschlossen. Die folgenden Kapitel beschäftigen sich mit den statistischen Ergebnissen der einzelnen Entwicklungsstufen und werden anhand von Zahlen verdeutlichen, wodurch sich gute und problematische Items auszeichneten und welche Entwicklung das Testverfahren im Verlauf seiner Entwicklung konkret genommen hat.

4.5 VALIDIERUNG

Wie bereits im Theorieteil der Arbeit begründet, können folgende Kriterien zur Prüfung der externen Validität herangezogen werden: als Leistungskriterium sind dies die *Schulnoten*, als Motivationskriterien das *Fachinteresse*, das *Interesse an naturwissenschaftsbezogenen Aktivitäten* und das *Interesse an naturwissenschaftlichen Tätigkeiten*. Die genannten Variablen wurden jeweils vor der Bearbeitung der Testaufgaben erhoben. Dies geschah, um Einflüsse des Tests, insbesondere auf die Beantwortung der Fragen aus dem Bereich der Beschäftigung mit naturwissenschaftlichen Themen und Inhalten und der Interessensskala, zu vermeiden. Im Folgenden werden die Skalen zur Erfassung der Kriterien dargestellt.

SKALA „INTERESSE AN NATURWISSENSCHAFTSBEZOGENEN AKTIVITÄTEN"

Zur Messung des *Interesses an naturwissenschaftsbezogenen Aktivitäten* wurde die Skala *Science-Activities* der nationalen PISA-Erhebung 2006 (OECD, 2007) herangezogen, die sechs Items umfasst. Unter der Fragestellung *„Wie oft machst du folgende Dinge?"* hatten die befragten Personen ein vierfach abgestuftes Antwortformat zur Verfügung, um die genannten Tätigkeiten mit *sehr oft*, *regelmäßig*, *manchmal* oder *nie oder fast nie* zu bewerten. Die Testpersonen mussten sich bei der Einschätzung der Häufigkeit der genannten Tätigkeiten für jeweils eine der Abstufungen entscheiden (s. Tab. 4.3).

Tabelle 4.3: Skala *Interesse an naturwissenschaftsbezogenen Aktivitäten*

Wie oft machst du folgende Dinge?

	sehr oft	regelmäßig	manchmal	nie oder fast nie
1) Fernsehsendungen über Naturwissenschaften anschauen				
2) Bücher über naturwissenschaftliche Themen ausleihen oder kaufen				
3) Internetseiten zu naturwissenschaftlichen Themen besuchen				
4) Radiosendungen über Fortschritte in den Naturwissenschaften anhören				
5) Naturwissenschaftliche Zeitschriften oder Artikel in der Zeitung lesen				
6) Eine Naturwissenschafts-AG besuchen				

SKALA „INTERESSE AN NATURWISSENSCHAFTLICHEN TÄTIGKEITEN"

Die Erfassung des Interesses der Schülerinnen und Schüler an naturwissenschaftlichen Tätigkeiten erfolgte anhand von Items, die einer Skala der IPN-Interessenstudie Physik (Hoffmann, L. & P., 1998) entnommen und adaptiert wurden (s. Tab. 4.4). Bei der Skala handelt es sich um die Einschätzung des eigenen Interesses an *Tätigkeiten, die auch im Physikunterricht vorkommen*. Die Anpassung der Items bestand zum einen darin, dass die Wörter Physik, physikalisch usw. durch Naturwissenschaften, naturwissenschaftlich usw. ersetzt wurden. Zum anderen wurden vier der sechs ausgewählten Items leicht verkürzt, da sie in der ursprünglichen Form für die Messungen im Rahmen dieser Arbeit nicht adäquat waren. Zusätzlich zu diesen Änderungen

Tabelle 4.4: Skala *Interesse an naturwissenschaftlichen Tätigkeiten*

Wie groß ist dein Interesse an folgenden Dingen?

	Mein Interesse ist...				
	sehr groß	groß	mittel	gering	sehr gering
1) einen Versuch aufbauen.....					
2) einen Versuch selber durchführen, Messungen machen.....					
3) sich ausdenken, wie man eine bestimmte Vermutung durch einen Versuch prüfen könnte.....					
4) etwas berechnen, den Ausgang eines Versuchs exakt vorhersagen.....					
5) mit anderen über neue naturwissenschaftliche Erkenntnisse diskutieren.....					
6) den Wert oder Nutzen einer neuen naturwissenschaftlichen Erkenntnis beurteilen.....					

wurde nur ein Teil des ursprünglichen Antwortformats der Skala genutzt. Die Skala sah vor, dass die Testpersonen sich auf einer fünffach gestuften Rating-Skala zunächst entscheiden, wie groß ihr Interesse an der jeweiligen Tätigkeit ist (Abstufung von *sehr groß* bis *sehr gering*) und danach auf einer weiteren fünffach gestuften Rating-Skala beurteilen, wie häufig diese Tätigkeit im Unterricht vorkommt (Abstufung von *sehr oft* bis *nie*). Da für die Erhebungen innerhalb dieser Arbeit unterrichtliches Geschehen nicht von Bedeutung ist und um die Erhebung dieser Variablen klar zu gestalten, wurde hier auf die Beurteilung der Vorkommenshäufigkeit verzichtet. Die ursprüngliche Skala findet sich zum Vergleich im Anhang dieser Arbeit (s. Anhang, Abschnitt B).

In der modifizierten Skala bestand die Aufgabe der Testpersonen darin, auf die Frage *„Wie viel Interesse hast du an folgenden naturwissenschaftlichen Dingen?“* die dargestellten Items mit *„Mein Interesse ist...“ sehr groß*, *groß*, *mittel*, *gering* oder *sehr gering* zu bewerten. Auch hier mussten sich die Testpersonen jeweils für eine der Abstufungen entscheiden.

NATURWISSENSCHAFTLICHES FACHINTERESSE UND SCHULNOTEN IN NATURWISSENSCHAFTLICHEN FÄCHERN

Als weitere Variablen zur Prüfung der externen Validität wurden das *Fachinteresse* und die *Schulnoten* erhoben. Das *Fachinteresse* wurde dabei wie in Tabelle 4.5 dargestellt erfragt: Die fünffache Abstufung der Ratingskala wurde analog zur Beurteilung

Tabelle 4.5: *Erhebung des Fachinteresses*

Bewerte bitte folgende Schulfächer mit einer Zahl zwischen 1 und 5. Es geht darum auszudrücken, wie groß dein Interesse an den einzelnen Fächern ist:

1= sehr groß / 2= groß / 3= mittelmäßig / 4= eher geringer / 5= sehr gering.

Fach	Deutsch	Englisch	Mathe	Physik	Chemie	Biologie
Bewertung						

des *Interesses an naturwissenschaftlichen Tätigkeiten*, gewählt. Die Bewertung der Fächer in Form von Zahlen ist ans Schulnotensystem angelehnt (auch wenn die Sechs in der Beurteilung fehlt) und sollte den Schülerinnen und Schülern eine Bewertung erleichtern. Eine *1* steht für eine positive, eine *5* für eine negative Bewertung.

Im Anschluss an die Bewertung ihres Interesses an den genannten Schulfächern wurden die Schülerinnen und Schüler gebeten, die Noten des letzten Halbjahreszeugnisses in den entsprechenden Fächern anzugeben. Diese Angaben erfolgten in einer neuen Tabelle in ähnlichem Format. Um Missverständnisse zu vermeiden und um zu verhindern, dass die Schülerinnen und Schüler den Unterschied zwischen den beiden Tabellen nicht erkennen, wurde im Manual der Testdurchführung festgelegt, dass die Testleiterin diese beiden Punkte gemeinsam mit den Schülerinnen und Schülern durchgeht. Tabelle 4.6 stellt das Format der Notenabfrage dar. Im Falle von ge-

Tabelle 4.6: Erhebung der *Schulnoten*

Gib bitte für die folgenden Schulfächer an, welche Note du im letzten Zeugnis hattest:

Fach	Deutsch	Englisch	Mathe	Physik	Chemie	Biologie
Zeugnis-Note						

meinsam unterrichteten naturwissenschaftlichen Schulfächern waren die Schülerinnen und Schüler aufgefordert, den zusammen unterrichteten Naturwissenschaften die gleiche Note zu geben. Die Testleiterin notierte diese Besonderheit als allgemeine Anmerkung im Rahmen der Testdurchführung.

5 Darstellung der Ergebnisse

Der Ergebnisteil wird das im Operationalisierungskapitel skizzierte Vorgehen der Aufgabenentwicklung mit Datenmaterial unterlegen. Beginnend bei der Präpilotphase werden über Pilotierung und Feldtest bis hin zum Haupttest die Schritte der Aufgabenentwicklung in einer Form dokumentiert, die anhand verbaler und statistischer Daten verdeutlicht, welche Items sich als unzureichend erwiesen haben und demzufolge überarbeitet werden mussten und welche Items sich durch gute Eigenschaften auszeichnen. Die Ergebnisse der einzelnen Testphasen werden zeigen, inwiefern die Überarbeitung der Items gelungen ist und welche Items bzw. Itemsets letztendlich aus dem Test entfernt werden mussten.

Im Anschluss an diese Betrachtungen wird die Testendform hinsichtlich ihrer Reliabilität und Validität geprüft.

5.1 Ergebnisse der Präpilotphase

In der Präpilotphase wurde das Testverfahren anhand eines ersten *Expertenpanels* und erster *Cognitive-Lab-Interviews* geprüft. Die hier erhobenen Daten liegen in verbaler Form vor und sind demzufolge weniger geeignet, tabellarisch dargestellt zu werden als Zahlenmaterial. Aus diesem Grund wird im Rahmen der Ergebnisdarstellung der Präpilotierung ein Aufgabenbeispiel gegeben, um zu verdeutlichen, wie die jeweiligen Kommentare der Expertinnen und Experten sowie der Schülerinnen und Schüler in die Aufgabenüberarbeitung eingeflossen sind. In Kombination mit den Eindrücken der Pilotierung führten diese Kommentare zu einer Überarbeitung der Items für den Feldtest.

5.1.1 Auswertung der Expertenurteile

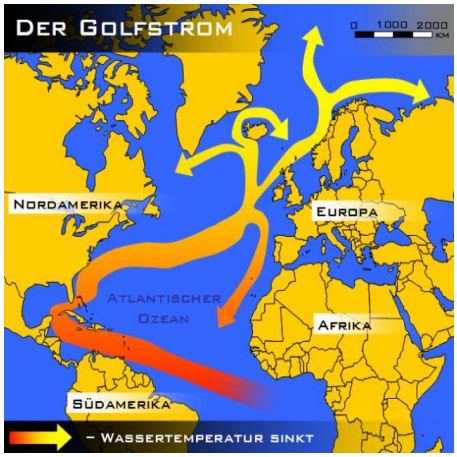
Die Expertengruppe bestand in der Präpilotphase aus einer Physiklehrerin, einem Biologielehrer, einer Diplom-Biologin, einem Diplom-Physiker sowie einem Diplom-Psychologen. Der folgende Abschnitt stellt beispielhafte Expertenurteile dar, die hinsichtlich des in Abbildung 5.1 dargestellten Items *Klima III (Kli3)* getroffen wurden. Die Kommentare wurden in Ergänzung des Beurteilungsleitfadens (s. Abschnitt C.1,

S. 249) geäußert. Die Expertinnen und Experten wurden gebeten, diesen Leitfaden für jedes Item zu bearbeiten und gegebenenfalls durch eigene Kommentare zu ergänzen. Die Beurteilung der Aufgabe gemäß Leitfaden fällt folgendermaßen aus: Zwei der

Klima III

Gletscher schmelzen, Permafrostböden tauen auf, Jahreszeiten verschieben sich, der Meeresspiegel steigt. Das Klima ändert sich, sagen die Forscher. Doch wie lässt sich diese Erkenntnis belegen? Sind 20 Grad warme Tage im Oktober ein Ausweis des Wandels, oder ein zufälliger Ausschlag? Woran lässt sich der Klimawandel festmachen?

Meerwasser sinkt ab, wenn es besonders salzreich und / oder besonders kalt (hohe Dichte) ist. Es steigt auf, wenn es besonders salzarm und / oder warm ist (geringere Dichte). Es gibt eine Vielzahl an Dingen, die die Dichte des Wassers beeinflussen, unter anderem geschieht dies durch Verdunstung (Dichte des Wassers nimmt zu) oder aber durch den Zufluss von Süßwasser (Dichte des Wassers nimmt ab). Von diesen Prozessen hängen unter anderem Meeresströme wie der Golfstrom ab. Forscher beobachten besonders das Abschmelzen der Polkappen und Gletscher kritisch.



Quelle: Wikipedia

Welcher Vermutung gehen die Wissenschaftler nach, wenn sie das Abschmelzen von Polkappen und Gletschern im Zusammenhang mit den Meeresströmungen beobachten?

a) Das Abschmelzen von Polkappen und Gletschern ändert den Salzhaushalt des Meerwassers und führt zu einer Veränderung der Meeresströmung.
b) Die Erwärmung des Klimas und der Meere führ dazu, dass Meeresströmungen wie der Golfstrom sich ändern.
c) Durch Verdunstung ändert sich der Salzgehalt des Meerwassers und damit die Meeresströmung.
d) Meerwasser sinkt ab, wenn es besonders kalt oder salzreich ist und es steigt auf, wenn es warm oder salzarm ist.

Abbildung 5.1: Erste Version der Aufgabe *Klima III*

drei hier zitierten Expertinnen und Experten schätzen die Aufgaben-Schwierigkeit als angemessen und die Aufgabe als verständlich formuliert ein. Alle Expertinnen und Experten halten die Aufgabe für interessant und beurteilen sie als sachlich richtig. Folgende Kommentare wurden durch die Expertinnen und Experten ergänzt:

- Kommentar Experte 1 (Physiker):

  - Das Wort *Permafrostböden* könnte für die Schülerinnen und Schüler nicht verständlich sein.
  - Die Klammern im Text sollten durch Umformulierung der Sätze entfernt werden.
  - Distraktor d erscheint zu attraktiv.
- Kommentar Expertin 2 (Biologin):
  - Das Wort *Permafrostböden* ist zu schwierig.
  - Der Text enthält zu viele Ausdrücke in Klammern.
  - Die Aufgabe ist zu schwer.
- Kommentar Expertin 3 (Physiklehrerin):
  - Aufgabe eventuell ohne den Begriff *Dichte* formulieren, damit physikalisches Grundwissen weniger von Bedeutung ist.
  - Die Aufgabe ist zu lang.

Es ist zu erkennen, dass die Anmerkungen der Expertinnen und Experten recht ähnlich ausfallen. Übergreifend wird das Wort *Permafrostböden* als für die Schülerinnen und Schüler nicht verständlich eingeschätzt, und die Klammern im Text werden im Hinblick auf den Lesefluss als störend empfunden. Die Aufgabe wird als zu lang, aber im Wesentlichen in der Schwierigkeit angemessen bewertet.

Alle Expertinnen und Experten trafen ihre Einschätzungen unabhängig voneinander und besprachen ihre Kommentare im ersten *Expertenpanel*. Danach wurden die jeweiligen Einschätzungen pro Aufgabe gepoolt. Das Gleiche geschah mit den Kommentaren der Schülerinnen im *Cognitive-Lab-Interview*, die im Folgenden dargestellt werden.

5.1.2 AUSWERTUNG DER COGNITIVE-LAB-INTERVIEWS

Die *Cognitive-Lab-Interviews* wurden mit zwei Schülerinnen, einer Realschülerin (15 Jahre) und einer Hauptschülerin (14 Jahre), einzeln durchgeführt. Den beiden Schülerinnen standen anders als den Expertinnen und Experten kein Leitfaden für die Bearbeitung der Aufgaben zur Verfügung. Sie waren lediglich dazu aufgefordert, im Sinne der Laut-Denken-Methode alle Gedanken zu äußern, die ihnen während der Bearbeitung der Aufgaben in den Sinn kamen. Wurden während der Aufgabenbearbeitung wenige oder gar keine Denkprozesse verbalisiert, wurden ermunternde Fragen gestellt. Die beiden standardisierten Nachfragen bezogen sich darauf, inwiefern

sie Inhalte und Abbildungen verstanden hatten und wie schwer auf einer Skala von 1-5 sie die Aufgabe einschätzten.

Bezüglich der Aufgabe *Klima III (Kli3)* merken die Schülerinnen Folgendes an:

- Kommentar Schülerin 1 (Hauptschülerin, 14 Jahre):
  - *Permafrostböden* - was ist Perma?
  - *Polkappen* ist auf Nord- und Südpol bezogen?
  - Die Klammern im Text stören.
  - Ich finde die Aufgabe eher mittelmäßig schwer.

- Kommentar Schülerin 2 (Realschülerin, 15 Jahre):
  - Die Klammern stören den Lesefluss.
  - Ansonsten ist der Text gut verständlich.
  - Die Graphik ist gut zu erkennen.
  - Ich finde die Aufgabe eher leicht.

Schüler- und Expertenkommentaren fallen sehr ähnlich aus. Das Wort *Permafrostböden* wird hier tatsächlich als problematisch und die Klammern im Text werden von den Schülerinnen ebenfalls als störend empfunden. Die Aufgabe wird als maximal mittelschwer eingeschätzt. Die Kommentare der Expertinnen und Experten sowie der Schülerinnen mündeten in folgenden Überarbeitungen: Das Wort *Permafrostböden* wurde durch *Dauerfrostböden* ersetzt, der Text wurde verkürzt und die Klammern wurden entfernt. Weiterhin wurde die Abbildung entfernt, weil sie keine notwendigen oder Verständnis fördernden Informationen enthält. Die allgemeinen einleitenden Informationen wurden zur weiteren Vereinfachung der Aufgabe von dem eigentlichen Aufgabenstamm getrennt und zur allgemeinen Einleitung des Itemsets *Klima* umfunktioniert. Abbildung 5.2 zeigt die neu bebilderte Einleitung des Sets. Auf diese Weise wurde auch im Fall aller anderen Itemsets verfahren. Die allgemeinen Informationen wurden aus den Aufgabenstämmen entfernt und in Form eines Informationsdeckblattes für die einzelnen Sets zusammengestellt. Abbildung 5.3 veranschaulicht die überarbeitete Aufgabe.

Klima

Quelle: ©Georges Bott / PIXELIO; www.pixelio.de

Gletscher schmelzen, Dauerfrostböden tauen auf, Jahreszeiten verschieben sich, der Meeresspiegel steigt. Das Klima ändert sich, sagen die Forscher. Doch wie lässt sich diese Erkenntnis belegen? Sind 20 Grad warme Tage im Oktober ein Beweis des Wandels, oder ein zufälliger Ausschlag? Woran lässt sich der Klimawandel festmachen?

Abbildung 5.2: Einleitung des Aufgabensets *Klima*

Klima III

Meerwasser sinkt ab, wenn es besonders salzreich und/oder besonders kalt ist und damit eine hohe Dichte hat. Es steigt auf, wenn es besonders salzarm und/oder warm ist und damit eine geringere Dichte hat. Eine Vielzahl von Vorgängen beeinflusst die Dichte des Wassers:

· Verdunstung erhöht die Dichte.
· Zufluss von Süßwasser vermindert die Dichte.

Von diesen Vorgängen hängen unter anderem Meeresströmungen wie der Golfstrom ab. In diesem Zusammenhang beobachten Forscher besonders das Abschmelzen der Polkappen und Gletscher kritisch.

Welcher Vermutung gehen die Forscher nach?

a) Durch Verdunstung ändert sich der Salzgehalt des Meerwassers und damit die Meeresströmung.
b) Die Erwärmung des Klimas und der Meere führ dazu, dass Meeresströmungen wie der Golfstrom sich ändern.
c) Das Abschmelzen von Polkappen und Gletschern ändert den Salzgehalt des Meerwassers und führt zu einer Veränderung der Meeresströmungen.
d) Meerwasser sinkt ab, wenn es besonders kalt oder salzreich ist und es steigt auf, wenn es warm oder salzarm ist.

Abbildung 5.3: Überarbeitete Version der Aufgabe *Klima III*

5.2 ERGEBNISSE DER PILOTIERUNG

Die Pilotierung diente in erster Linie der Überprüfung, inwiefern die veranschlagte Testzeit realistisch ist und inwiefern sowohl die Instruktionen der Testdurchführung als auch die Aufgaben verständlich sind. Die Kommentare und Fragen der Schülerinnen und Schüler wurden während der Testung schriftlich festgehalten und flossen in die bereits beispielhaft beschriebene Auswertung der übrigen verbalen Daten ein. Die ersten empirischen Ergebnisse, die im Folgenden präsentiert werden, müssen aufgrund der kleinen Stichprobe mit Vorsicht betrachtet werden und können lediglich als erste Hinweise auf qualitativ unzureichende Items gelten.

5.2.1 PILOTIERUNGSSTICHPROBE

Zur Pilotierung der Daten wurde eine kleine Stichprobe benötigt, die sich bezüglich des Leistungsniveaus ähnlich zusammensetzen sollte wie die Stichproben des späteren Feld- und Haupttests. Es sollten also Schülerinnen und Schüler der Haupt-, Realschule- und des Gymnasiums gleichermaßen vertreten sein.

Die Pilotierungsstichprobe bestand aus 35 Schülerinnen und Schülern, die zwei neunten Klassen einer schleswig-holsteinischen Gesamtschule entstammten. Die eine Klasse umfasste Schülerinnen und Schüler mit Hauptschul- und unterem Realschulniveau (n=12), die andere Klasse Schülerinnen und Schüler mit oberem Realschul- und Gymnasialniveau (n=23). Allen Schülerinnen und Schülern wurden alle Aufgaben des Testheftes vorgelegt.

5.2.2 STATISTISCHE AUSWERTUNGEN

Die folgende Tabelle 5.1 gibt einen Überblick über die anhand der Pilotierungsstichprobe ermittelten deskriptiven Kennwerte. Abweichende Trennschärfen erscheinen fett gedruckt. Die Itemantworten des Tests sind binär kodiert (*falsch = 0, richtig = 1*). Aus diesem Grund gibt der Mittelwert der Items im Sinne der klassischen Itemschwierigkeit darüber Auskunft, wie hoch der Anteil an Testpersonen ist, die das jeweilige Item gelöst haben. Ein Wert von $0,54$ (Bro1) sagt dementsprechend aus, dass 54% der Testpersonen dieses Item gelöst haben. Je höher der Mittelwert ausfällt, desto mehr Personen haben das Item gelöst und desto leichter ist es.

Zur Beurteilung der Itemschwierigkeiten wird allgemein im Rahmen dieser Testentwicklung das Kriterium angelegt, dass die Itemschwierigkeit der Personenfähigkeit der Stichprobe angepasst sein sollte. Das bedeutet, dass die gesamte Bandbreite der Kompetenzausprägungen in der Stichprobe aus Schülerinnen und Schülern der

Tabelle 5.1: Deskriptive Beschreibung des Itempools

Items	M	Trennschärfe
Bro 1	0,54	0,33
Bro 2	0,29	**0,08**
Bro 3	0,66	**0,24**
Hur 1	0,46	0,48
Hur 2	0,54	0,32
Hur 3	0,57	**0,20**
Sch 1	0,34	0,45
Sch 2	0,40	**0,21**
Sch 3	0,74	0,60
Sti 1	0,80	**0,22**
Sti 2	0,63	0,55
Sti 3	0,43	0,52
KLe 1	0,40	**0,14**
KLe 2	0,09	**-0,01**
KLe 3	0,60	**-0,03**
Kli 1	0,49	0,40
Kli 2	0,46	0,49
Kli 3	0,51	0,51
Reg 1	0,40	**0,20**
Reg 2	0,23	**0,11**
Reg 3	0,57	**0,25**
Ros 1	0,54	0,46
Ros 2	0,37	0,40
Ros 3	0,63	0,54
Sko 1	0,66	0,42
Sko 2	0,60	0,53
Sko 3	0,37	**0,09**
Sol 1	0,69	0,42
Sol 2	0,37	0,36
Sol 3	0,40	0,37
M	0,49	0,33

M = Mittelwert

neunten Klasse (Hauptschule, Realschule und Gymnasium) durch den Test erfassbar sein sollte. Abbildung 5.4 stellt den in Tabelle 5.1 dargestellten Itemschwierigkeiten die Personenfähigkeiten gegenüber. Aufgrund der zu geringen Stichprobengröße können hier nicht wie in den später folgenden Darstellungen der Feld- und Haupttestdaten die probabilistischen Personenkennwerte (WLEs) für die Darstellung der Kompetenzbandbreite herangezogen werden. Um dennoch eine Beurteilung der Itemschwierigkeiten zu ermöglichen, werden sie dem Anteil der durch die Testpersonen gelösten Aufgaben gegenübergestellt. Ein Anteil von $0,60$ bedeutet hier, dass die

Testpersonen 60% der Items gelöst haben. Je höher der Anteil, desto kompetenter die Person. Die Graphik zeigt, dass es sowohl im unteren, als auch im oberen Kompetenz-

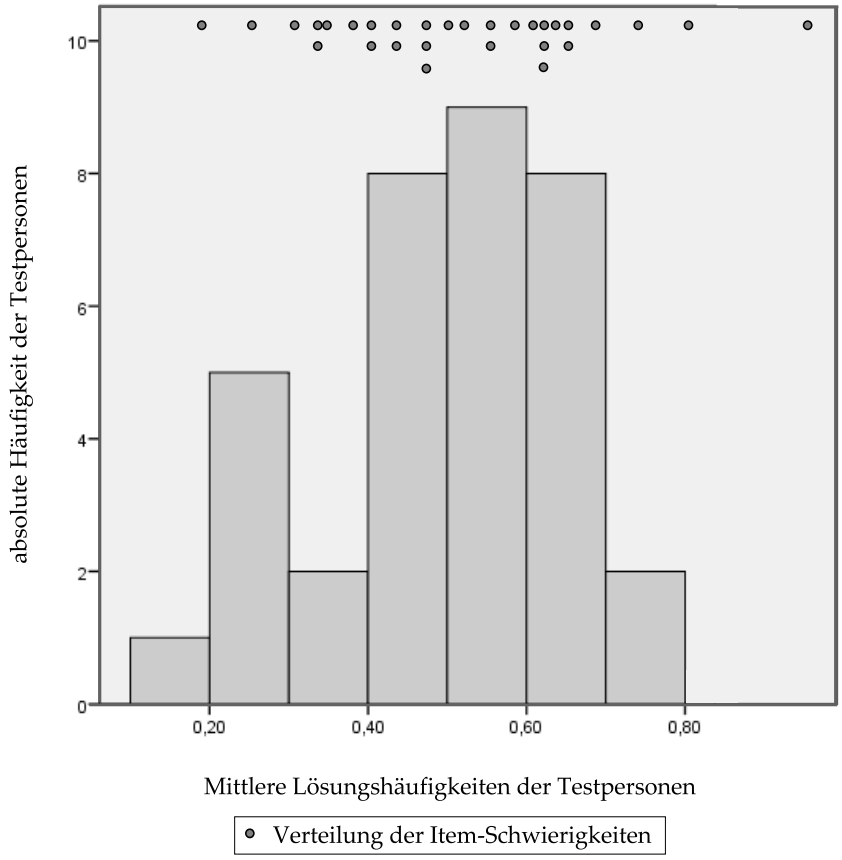


Abbildung 5.4: Verteilung von Itemschwierigkeiten und Personenfähigkeiten

bereich an Items mangelt. Es gibt eine Itemhäufung im mittleren Kompetenz- bzw. Schwierigkeitsbereich. Die extremsten Itemschwierigkeiten zeigen die Items Sti1, das mit einer Lösungswahrscheinlichkeit von $0,80$ das leichteste Item darstellt (hier angepasst an die Fähigkeitsskala der Testpersonen umkodiert in den Wert $0,2$), und das Item KLe2, das mit einer Lösungswahrscheinlichkeit von $0,09$ (umkodiert in den Wert $0,91$) am schwersten ausfällt.

Legt man bezüglich der Trennschärfen das Kriterium von Kelava et al. (2007) an, dass optimale Trennschärfen zwischen $0,4$ und $0,7$ liegen sollten, um als gut bezeichnet werden zu können und räumt man eine zusätzliche Toleranz ein, die bis zu einer Schwelle von $0,3$ reicht, so weisen folgende Items unzureichende Trennschärfen auf: *Bro2* und *Bro3* des Sets *Brot backen*, *Hur3* des Sets *Hurrikans*, *Sch2* des Sets *Schimmel*, *Sti1* des Sets *Stichlinge*, *KLe1-3* des Sets *Kleine Lebewesen, kleine Teilchen*, *Reg1-3* des Sets *Regenbogenforelle* und *Sko3* des Sets *Skorbut*.

Diese aufgrund der kleinen Stichprobe mit Vorbehalt zu betrachtenden statistischen Hinweise auf unzureichende Items und Itemsets führten in Kombination mit den verbalen Daten der Pilotierung zu den im Folgenden dargestellten Erkenntnissen.

5.2.3 ERKENNTNISSE DER TESTDURCHFÜHRUNG

Die Pilotierung diente vorrangig dem Ziel, die spätere Feldtestung zu simulieren, den zeitlichen Rahmen der Testdurchführung zu prüfen und noch einmal die Möglichkeit zu nutzen, nicht verständliche oder missverständliche Textpassagen zu identifizieren. Folgende Erkenntnisse ergaben sich aus der Durchführung der Pilotierung:

- Die Einleitung der Testung, die neben den eigentlichen Testinstruktionen auch den Hintergrund und das Ziel der Testdurchführung erklären sollte, wies noch Schwächen auf. Beispielsweise musste noch genauer ausgeführt werden, worin der Hintergrund der Doktorarbeit besteht und warum für die Testentwicklung die Hilfe der Schülerinnen und Schüler benötigt wird. Zu diesen beiden Punkten gab es die meisten Nachfragen.

- Einige Schülerinnen und Schüler beklagten sich über den Textumfang des Tests. Bezüglich der Testüberarbeitung resultierte daraus, dass alle Aufgaben diesbezüglich noch einmal kritisch betrachtet wurden. Es wurden alle Textteile entfernt, die unnötige Informationen enthielten.

- Es gab einige Ausdrücke, welche die Schülerinnen und Schüler nicht verstanden und die somit überarbeitet werden mussten. Beispiele für solche Ausdrücke sind das Wort *Wandrelief* aus der Einleitung des Sets *Brot backen* oder die Bezeichnung *Bakterium E.coli* aus dem Set *Kleine Lebewesen, kleine Teilchen*. Derartige Ausdrücke wurden vor der Feldtestung verändert oder aber aus der Aufgabe entfernt, wenn sie nicht unbedingt als Informationsbasis notwendig waren.

- Bezüglich der Dauer der Testung konnte festgestellt werden, dass die veranschlagte Testzeit von einer Doppelstunde für die Testung ausreichend sein sollte.

Alle Items wurden unter Kombination der Experten- und Schülerkommentare der Präpilotphase sowie der ersten statistischen Daten und Erkenntnisse der Pilotierung betrachtet und, wenn es angezeigt war, einer Überarbeitung unterzogen. Besonders genau wurden die Items betrachtet, die sich mehrfach (sowohl aufgrund der verbalen als auch aufgrund der statistischen Daten) als auffällig erwiesen hatten. Dies betraf insbesondere die Itemsets *Kleine Lebewesen, kleine Teilchen* und *Regenbogenforelle*.

5.3 ERGEBNISSE DES FELDTESTS

Der Feldtest stellte die erste statistische Prüfung der Testdaten dar, die aufgrund einer ausreichend großen Stichprobe als aussagekräftig gelten konnte. Die Daten wurden genutzt, um festzustellen, welche Items einer weiteren Überarbeitung bedurften. Im Besonderen wurden in diesem Rahmen auch die Antwortalternativen *der Items auf Unzulänglichkeiten geprüft. Die Erkenntnisse der statistischen Auswertungen wurden genutzt, um das zweite Expertenpanel* sowie das zweite *Cognitive-Lab* gezielter einsetzen zu können, um Ansätze für die Überarbeitung der Items zu finden.

5.3.1 STICHPROBENBESCHREIBUNG

Tabelle 5.2 gibt Auskunft über die Zusammensetzung der Stichprobe. Die Stichprobengrößen fallen pro Schulform gleich groß aus, was durch die zufällige Ziehung gleich großer Unterstichproben bedingt ist. Die Anzahl der Schulen, die pro Schulform zur Datenerhebung herangezogen wurden, ist im Falle von Haupt- und Realschule ausgeglichen. Im Falle des Gymnasiums wurde lediglich eine Schule benötigt, um die notwendigen Schülerzahlen zu erreichen. Hinsichtlich der Anzahl der pro Schulform getesteten Klassen gestalten sich die Zahlen sehr ausgeglichen.

Tabelle 5.2: Beschreibung der Feldtest-Stichprobe

		H	R	G	Summe
Stichprobe	**n**	105	105	105	315
Anzahl Schulen		3	3	1	7
Anzahl Klassen		5	4	5	14

n = Stichprobengröße H = Hauptschule R = Realschule G = Gymnasium

5.3.2 KENNWERTE DES FELDTESTS

Die Betrachtung der Itemkennwerte beginnt mit dem Modellgeltungstest. Es wird geprüft, ob ein eindimensionales Testmodell die Daten besser erklären kann als ein dreidimensionales. Dem liegt die theoretische Annahme zu Grunde, dass die zu messende Kompetenz ein eindimensionales Konstrukt darstellt, welches lediglich anhand von drei repräsentativen Fertigkeiten gemessen wird. Vor dem Hintergrund dieser Annahmen wurden die Testaufgaben entwickelt.

Im Anschluss an den Modellgeltungstest werden der Anteil fehlender Werte pro Item sowie die Itemkennwerte dargestellt. Diese Darstellung wird begleitet von einer Betrachtung der Items auf der Ebene der Antwortalternativen. Hier wird geprüft, wie sich die Antworten auf die einzelnen Alternativen verteilen und inwiefern die insgesamt fähigsten Testpersonen die richtige Antwortalternative wählen.

GLOBALE MODELLPRÜFUNG

Bevor eine Betrachtung der Itemkennwerte erfolgen kann, ist es zunächst nötig festzustellen, welches Testmodell die Daten am besten zu erklären vermag.

Tabelle 5.3 stellt die Modellselektionsmaße CAIC und BIC des eindimensionalen Modells denen des dreidimensionalen Modells gegenüber. Eine inferenzstatistische Prüfung der informationstheoretischen Maße ist nicht möglich. Es kann lediglich die Höhe der Werte beurteilt werden. Kleine CAIC- und BIC-Werte sprechen für eine bessere Passung des Modells. Die dargestellten Werte weisen das dreidimensionale

Tabelle 5.3: Globale Modellprüfung anhand des CAIC und BIC

Modell	unabh. Paramter	Deviance	CAIC	BIC
Eindimensional	31	11824,1	12033,4	12002,4
Dreidimensional	36	11789,3	12032,4	11996,4

Deviance = globale Fit-Statistik CAIC = Consistent Akaike´s Information Criterion
BIC = Bayes Information Criterion

Modell als leicht überlegen aus. Berücksichtigt man bei der Beurteilung der Ergebnisse auf der einen Seite, dass CAIC-Unterschiede zwischen 0 und 2 als vernachlässigbar zu bewerten sind (Burnham & Anderson, 2002) und auf der anderen Seite, dass das dreidimensionale Modell bei Betrachtung des BIC über eine zu vernachlässigende Differenz hinaus ein wenig besser geeignet scheint, so fällt die eindeutige Entscheidung für ein Modell schwer.

Um eine Entscheidung über den Umgang mit diesem nicht eindeutigen Ergebnis treffen zu können, wurden die latenten Korrelationen betrachtet, die in Tabelle 5.4 dargestellt sind. Die latenten Dimensionen weisen hohe bis sehr hohe Korrelationen auf. Es erscheint daher angemessen, die drei Dimensionen zu einer zusammenzufassen. Vor dem Hintergrund, dass das Ziel der Testentwicklung darin besteht, ein Instrument für ein eindimensionales Screening *prozessbezogener naturwissenschaftlicher*

Tabelle 5.4: Latente Korrelationen des dreidimensionalen Modells

	Dimension 1	Dimension 2
Dimension 1		
Dimension 2	0,89	
Dimension 3	0,83	0,91

Grundbildung zu schaffen, erscheint es unter Berücksichtigung der hier dargestellten Ergebnisse sinnvoll, die Testitems basierend auf Parameterschätzungen des eindimensionalen Modells zu beurteilen und weiter zu entwickeln. Als Beispiel für die Zusammenfassung von Teilskalen zu einer aussagekräftigen Gesamtskala kann hier die PISA-Untersuchung herangezogen werden (OECD, 2009), deren drei naturwissenschaftlichen Teilkompetenzen ebenfalls sehr hohe Korrelationen aufweisen und ebenfalls zusammengefasst als naturwissenschaftliche Kompetenz ausgewertet wurden.

Im Rahmen des Haupttests werden erneut Modellprüfungen vorgenommen. Nach der Anpassung der Items aufgrund der Feldtestdaten und basierend auf einer neuen und größeren Stichprobe soll so festgestellt werden, ob sich die Ergebnisse des Feldtests hinsichtlich der Dimensionalität im Haupttest bestätigen.

FEHLENDE WERTE

Der grundsätzliche Umgang mit den im Laufe der Testentwicklung auftretenden fehlenden Werten wurde bereits in Abschnitt 3.3.3 erläutert. An dieser Stelle werden die einzelnen Items im Sinne einer Qualitätsprüfung hinsichtlich ihres Anteils an fehlenden Werten untersucht, um eine systematische Häufung fehlender Werte ausschließen zu können. Wie bereits in Abschnitt 3.3.3 erläutert, liegt der maximal tolerierbare Anteil fehlender Werte bei 5%. Das Kriterium ist streng genug, um sensibel für die systematische Häufung fehlender Werte zu sein.

Tabelle 5.5 gibt Auskunft über die pro Item fehlenden Werte. Den höchsten Prozentsatz fehlender Werte weisen die Items *Sch2*, *Ros1+3*, *Sol3* auf. Die durchschnittliche Anzahl fehlender Werte pro Item liegt bei 3,9. Vor diesem Hintergrund muss insbesondere das Item *Sch2* als stark abweichend betrachtet werden. Es sollte in der Haupttestauswertung geprüft werden, ob dieses Item auch dort durch einen im Vergleich zu den übrigen Items überdurchschnittlichen Anteil fehlender Werte auffällt.

Tabelle 5.5: Anzahl und Anteil fehlender Werte pro Item

Items	fehlende Werte n	%
Bro 1	4	1,2
Bro 2	2	0,6
Bro 3	3	0,9
Hur 1	2	0,6
Hur 2	6	1,9
Hur 3	4	1,2
Sch 1	1	0,3
Sch 2	11	3,5
Sch 3	3	0,9
Sti 1	3	0,9
Sti 2	5	1,6
Sti 3	4	1,2
KLe 1	3	0,9
KLe 2	4	1,2
KLe 3	6	1,9
Kli 1	4	1,2
Kli 2	4	1,2
Kli 3	1	0,3
Reg 1	1	0,3
Reg 2	5	1,6
Reg 3	2	0,6
Ros 1	9	2,8
Ros 2	1	0,3
Ros 3	9	2,8
Sko 1	1	0,3
Sko 2	2	0,6
Sko 3	5	1,6
Sol 1	0	0,0
Sol 2	4	1,2
Sol 3	8	2,5

n = Absolute Anzahl fehlender Werte

Auf diese Weise kann festgestellt werden, ob es sich hier um eine zufällige Häufung handelt.

Insgesamt zeigt sich, dass keines der Items mehr als 5% fehlende Werte aufweist, der Anteil also als sehr gering eingeschätzt werden kann.

DESKRIPTIVE DATEN DES FELDTESTS

Die Betrachtung der Itemkennwerte beginnt mit einem Überblick über die deskriptiven Itemdaten, welche anhand von Tabelle 5.6 dargestellt werden. Hier fällt insbe-

Tabelle 5.6: Deskriptive Daten des Itempools

Items	M	SD
Bro 1	0,39	0,488
Bro 2	0,29	0,455
Bro 3	0,63	0,485
Hur 1	0,34	0,473
Hur 2	0,66	0,474
Hur 3	0,45	0,499
Sch 1	0,40	0,492
Sch 2	0,48	0,500
Sch 3	0,62	0,487
Sti 1	0,60	0,491
Sti 2	0,41	0,493
Sti 3	0,37	0,484
KLe 1	0,36	0,480
KLe 2	**0,07**	0,267
KLe 3	0,35	0,479
Kli 1	0,37	0,483
Kli 2	0,40	0,491
Kli 3	0,41	0,493
Reg 1	0,50	0,501
Reg 2	0,25	0,436
Reg 3	0,40	0,491
Ros 1	0,68	0,467
Ros 2	0,39	0,487
Ros 3	0,59	0,493
Sko 1	0,50	0,501
Sko 2	0,67	0,471
Sko 3	0,36	0,481
Sol 1	0,56	0,497
Sol 2	0,39	0,489
Sol 3	0,38	0,486
M	0,44	

M = Mittelwert SD = Standardabweichungen

sondere das Item *KLe2* auf, das nur von einem geringen Anteil der Schülerinnen und Schüler richtig beantwortet wird und damit die ersten statistischen Ergebnisse der

Pilotierung bestätigt. Dieses Item kann gemäß Tabelle D.1 (s. Anhang D.1.1, S. D.2) nur von den Fähigsten beantwortet werden. Es ist aufgrund des extremen Wertes zu erwarten, dass die Trennschärfe des Items dementsprechend gering ausfällt. Diese Annahme ist anhand der probabilistischen Itemkennwerte (Itemschwierigkeiten und -trennschärfen) zu prüfen, die nachfolgend betrachtet und bewertet werden.

ITEMKENNWERTE

Anhand der in Tabelle 5.7 dargestellten probabilistischen Kennwerte (Schwierigkeit, MNSQ und T-Wert) erfolgt eine vertiefende Analyse der Items. Hier wird zum einen geprüft, welche Items durch unzureichende oder abweichende Trennschärfen auffallen und zum anderen, ob die Aufgabenschwierigkeiten sich gleichmäßig über die Personenfähigkeiten verteilen (Wilson, 2005). Items, die abweichende Werte aufweisen, erscheinen in der Tabelle fett gedruckt.

Die *Itemschwierigkeiten* werden einerseits vor dem Hintergrund betrachtet, dass die mittlere Personenfähigkeit bei der Schätzung der Itemkennwerte auf Null fixiert wurde und andererseits unter Berücksichtigung der Zielsetzung, dass die Itemschwierigkeiten die Personenfähigkeiten abdecken sollten. Damit ist gemeint, dass es für jedes Fähigkeitsniveau in der Schwierigkeit angemessene Items geben sollte.

Allgemein ist zunächst festzustellen, dass die mittlere Itemschwierigkeit mit einem Wert von $0,32$ von der mittleren Personenfähigkeit abweicht. Der Test ist für die Stichprobe im Durchschnitt also deutlich zu schwer. Abbildung 5.5 visualisiert diesen Umstand in Form einer Gegenüberstellung der Verteilungen von Personenfähigkeiten und Itemschwierigkeiten. Die Personenfähigkeiten sind links, die Itemschwierigkeiten rechts in *Logits* (vgl. Abschnitt 3.1.2) abgetragen. Es ist zu erkennen, dass 21 der insgesamt 30 Items mit ihrer Schwierigkeit über der mittleren Personenfähigkeit von *0* liegen. Es fehlt insbesondere an leichten Items für Personen des unteren Fähigkeitsdrittels. Auch für eine Differenzierung von Personen mit sehr hoher Fähigkeitsausprägung fehlt es an Items. Hier ist insbesondere der Schwierigkeitsbereich zwischen Item 20 und Item 14 angesprochen, wobei Item 14 in seiner Schwierigkeit bereits über die höchste Personenfähigkeit hinausgeht. Dem Umstand, dass es mehr Items für den unteren Leistungsbereich geben sollte, wurde in die Überarbeitung der Items dadurch Rechnung getragen, dass Komplexität und Umfang der Aufgabentexte, wenn möglich, noch weiter reduziert wurden.

Für die Betrachtung der *Itemtrennschärfen* war es wichtig zu beachten, dass sie zwischen $0,4$ und $0,7$ liegen sollten, um als gut bezeichnet werden zu können (Kelava & Moosbrugger, 2007). Trennschärfen zwischen $0,3$ und $0,4$ werden hier als noch

Tabelle 5.7: Itemkennwerte des Feldtests

Item-Nr.	Items	Schwierigkeit	Trennschärfe	MNSQ	T-Wert
1	Bro 1	0,43	0,45	0,96	-0,9
2	Bro 2	1,01	0,30	1,03	0,5
3	Bro 3	-0,51	0,50	0,94	-1,4
4	Hur 1	0,61	**0,19**	1,11	2,5
5	Hur 2	-0,65	0,46	0,88	-2,7
6	Hur 3	0,20	0,40	0,96	-1,2
7	Sch 1	0,36	0,53	0,89	-3,0
8	Sch 2	0,22	**0,29**	1,08	2,1
9	Sch 3	-0,45	0,48	0,93	-1,8
10	Sti 1	-0,34	0,36	1,00	0,1
11	Sti 2	0,39	**0,18**	1,10	**2,5**
12	Sti 3	0,52	0,35	1,04	0,9
13	KLe 1	0,69	0,32	1,02	0,5
14	KLe 2	2,60	**0,11**	1,06	0,4
15	KLe 3	0,64	**0,14**	1,14	**3,0**
16	Kli 1	0,72	0,33	1,02	0,5
17	Kli 2	0,38	0,31	1,01	0,3
18	Kli 3	0,53	0,30	1,04	1,1
19	Reg 1	0,12	0,34	1,02	0,5
20	Reg 2	1,22	**0,22**	1,06	0,9
21	Reg 3	0,48	**0,25**	1,04	1,1
22	Ros 1	-0,62	0,48	0,94	-1,4
23	Ros 2	0,48	0,47	0,96	-1,0
24	Ros 3	-0,23	0,42	0,99	-0,2
25	Sko 1	-0,11	0,34	0,99	-0,4
26	Sko 2	-0,67	0,49	0,89	-2,6
27	Sko 3	0,72	0,35	0,96	-0,8
28	Sol 1	-0,27	0,34	1,00	0,1
29	Sol 2	0,40	0,46	0,93	-1,9
30	Sol 3	0,67	**0,24**	1,04	0,9
	Mittelwert	0,32	0,35	1,00	

MNSQ = Infit Mean Square T-Wert = Prüfgröße des Infit-Maßes

ausreichend betrachtet. Unter Berücksichtigung dieses Kriteriums fallen im Feldtest acht Items durch Trennschärfen $< 0,3$ auf: *Hur1*, *Sch2*, *Sti2*, *KLe2+3*, *Reg2+3* und *Sol3*. Die Trennschärfen der Items *Hur1* und *KLe2+3* liegen sogar unter $0,2$. Als Konsequenz dieser Betrachtungen wurden die genannten Items mit dem Label *zu überarbeiten* versehen. Hinweise auf Ansatzpunkte für die Überarbeitung lieferten zum einen die Analyse der Antwortalternativen im folgenden Abschnitt sowie die Auswertung

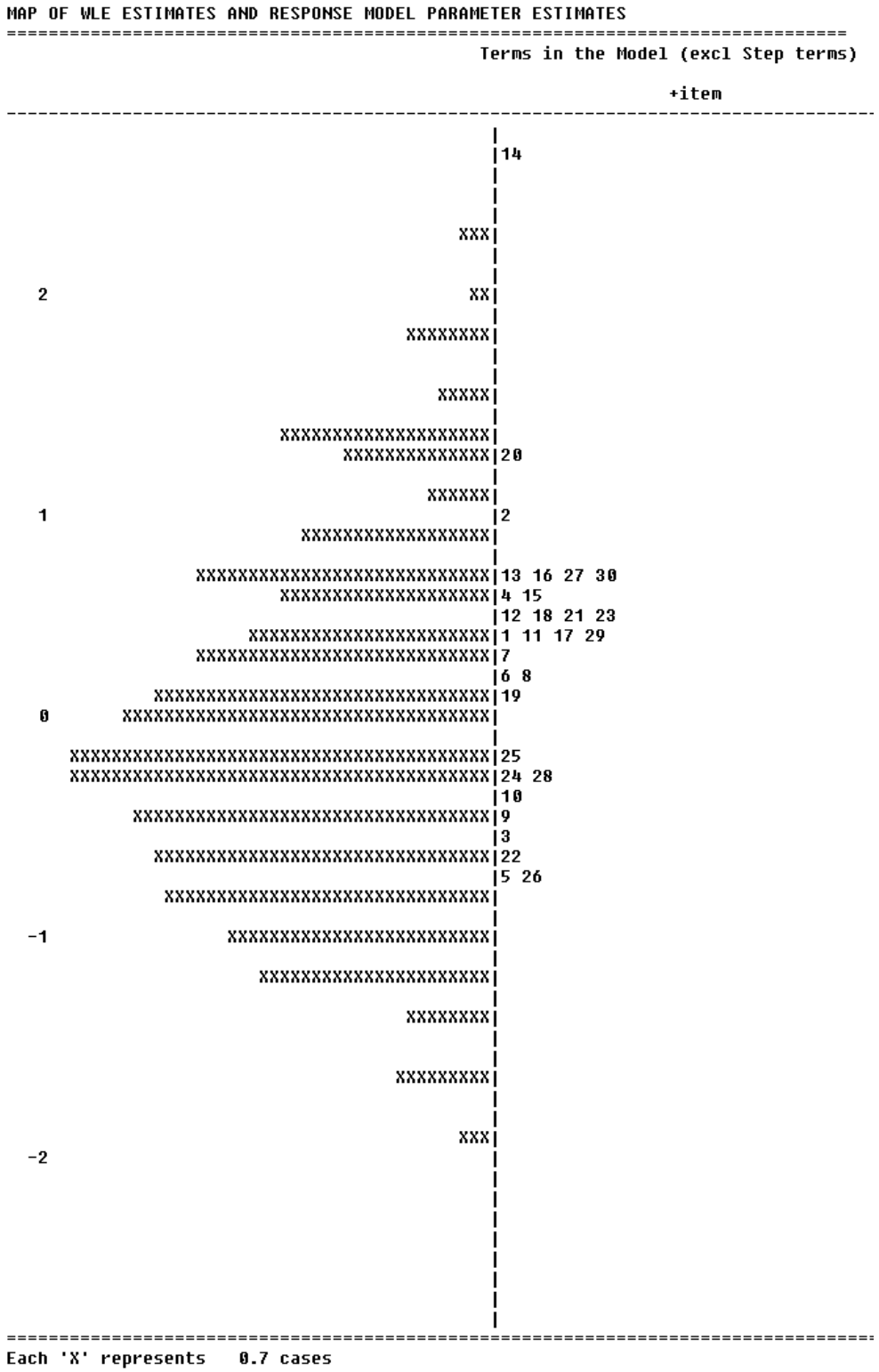


Abbildung 5.5: Verteilung der Itemschwierigkeiten über die Personenfähigkeiten

der Experten- und Schülerkommentare (s. Abschnitt 5.3.3, ab S. 175).

Die durchschnittliche Trennschärfe des Feldtests beträgt $0,35$. Dieser Durchschnittswert kann noch nicht als gut bezeichnet werden.

Der *T-Wert* stellt die Prüfgröße des Infit-Maßes, der Mean Squares (MNSQ), dar und dient als Indikator für abweichende Trennschärfen (vgl. Abschnitt 3.2.1), die eine Verletzung des angenommenen Modells darstellen würden. Es wird an dieser Stelle zunächst geprüft, welche Items T-Werte aufweisen, die > 2[1] sind und damit ein Anzeichen für zu geringe Trennschärfen darstellen können. Im Falle solcher Items schließt sich zunächst eine Betrachtung der *Infit Mean Squares (MNSQ)* an. Als akzeptabel wird ein Intervall zwischen $0,75$ und $1,33$ betrachtet (R. Adams & Khoo, 1996). Zur abschließenden Prüfung, inwiefern abweichende T-Werte tatsächlich für zu geringe Trennschärfen sprechen, werden die Abweichungen der beobachteten von den erwarteten Item-Characteristic-Kurven (ICCs) der jeweiligen Items betrachtet.

Bei der Betrachtung der T-Werte fallen das Item *Hur1* des Sets Hurrikans, das Item *Sti2* des Sets *Stichlinge* und das Item *KLe3* des Sets *Kleine Lebewesen* auf. Die positiven T-Werte dieser Items stellen ein mögliches Anzeichen für unterdurschnittliche Trennschärfen und für eine mögliche Modellverletzung dar.

Die Betrachtung des Infits (MNSQ) legt zunächst keine Modellverletzung nahe. Alle drei Werte weichen zwar deutlich von 1 ab, liegen aber noch in dem von Adams und Khoo (1996) vorgeschlagenen Rahmen. Allerdings zeigt die abschließende Untersuchung der ICCs, die anhand der Abbildungen 5.6, 5.7 und 5.8 erfolgt, dass im Falle dieser Items durchaus von abweichenden Trennschärfen gesprochen werden kann.

Die beobachteten ICCs der Items 11 *(Sti2)* und 15 *(KLe3)* weisen deutliche Abweichungen von den erwarteten ICCs auf, bei Item 4 *(Hur1)* fällt die Abweichung weniger deutlich aus. Die Steigungen der Kurven (im jeweils steilsten Punkt) fallen bei allen drei Items flacher aus und die Trennschärfen sind folglich geringer als erwartet. Zum Vergleich und zur besseren Einschätzung der Abweichung wird diesen Abbildungen die ICC des Items *3 (Bro3)* gegenübergestellt (s. Abb. 5.9).

An Abbildung 5.9 ist zu erkennen, dass die beobachtete ICC näher an der erwarteten verläuft als bei den Items *4 (Hur1)*, *11 (Sti2)* und *15 (KLe3)*. Man kann also im Falle dieser drei Items durchaus von abweichenden Trennschärfen sprechen, die ein Indiz dafür sind, dass die Items etwas anderes messen als angenommen.

Die Schlussfolgerung dieser Ergebnisse besteht darin, dass Items *Sti2* und *KLe3* einer Überarbeitung bedürfen. Auch in diesem Fall wurden *Expertenurteile* und *Cogni-*

1 Hierbei handelt es sich um einen Wert, der stichprobenabhängig ist und damit lediglich als Richtwert betrachtet werden kann (vgl. Abschnitt 3.2.1)

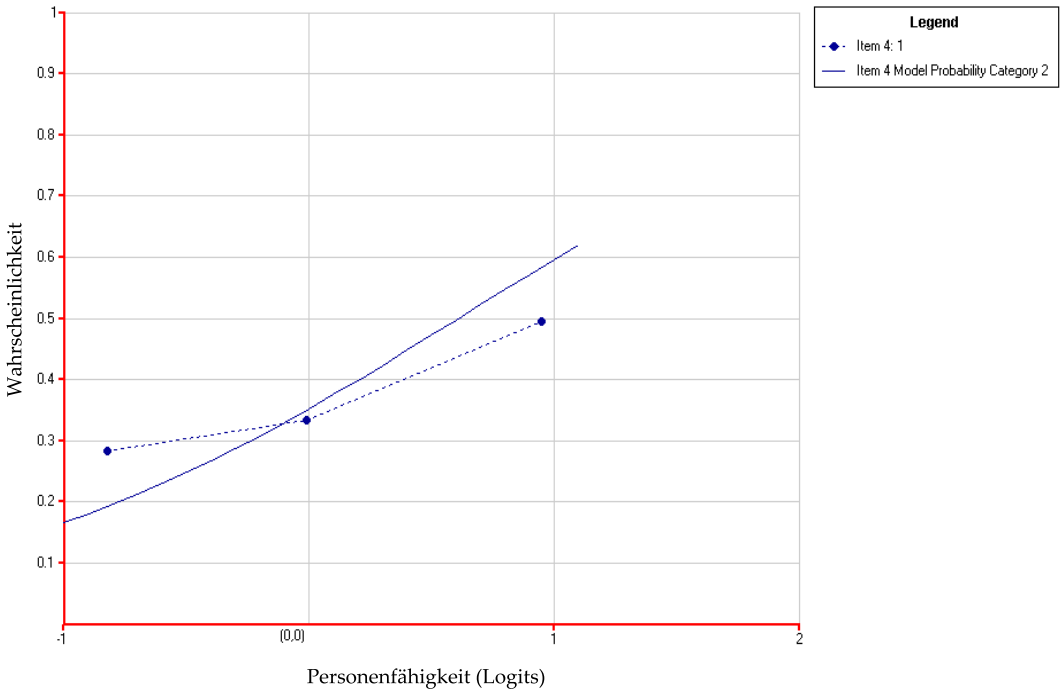


Abbildung 5.6: Erwartete (Model) und beobachtete ICC des Items *Hur1*

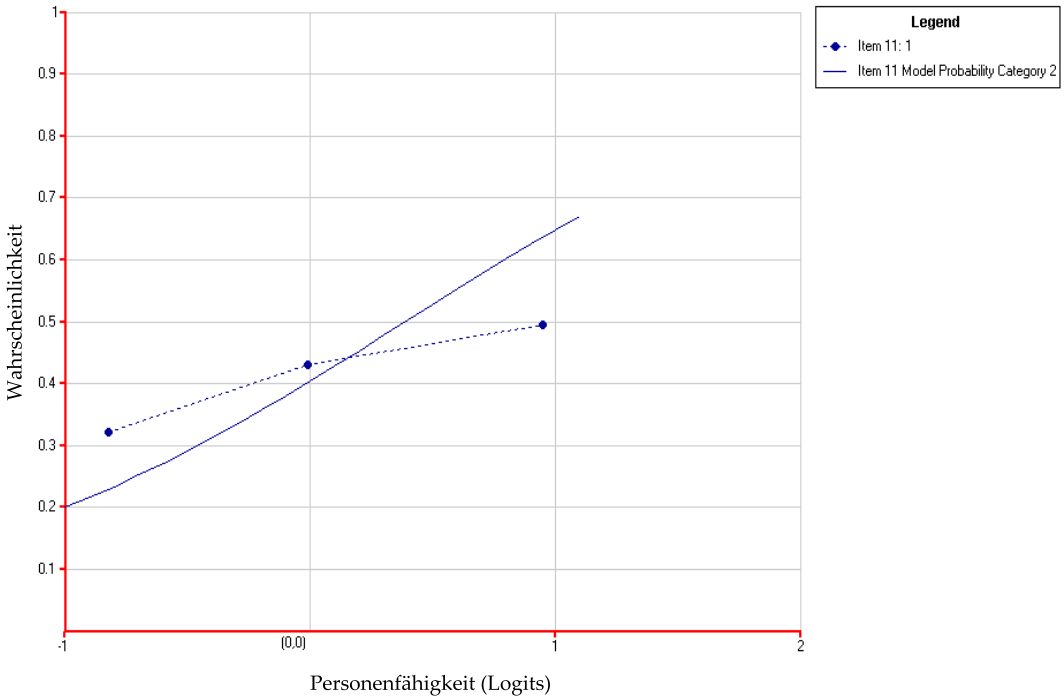


Abbildung 5.7: Erwartete (Model) und beobachtete ICC des Items *Sti2*

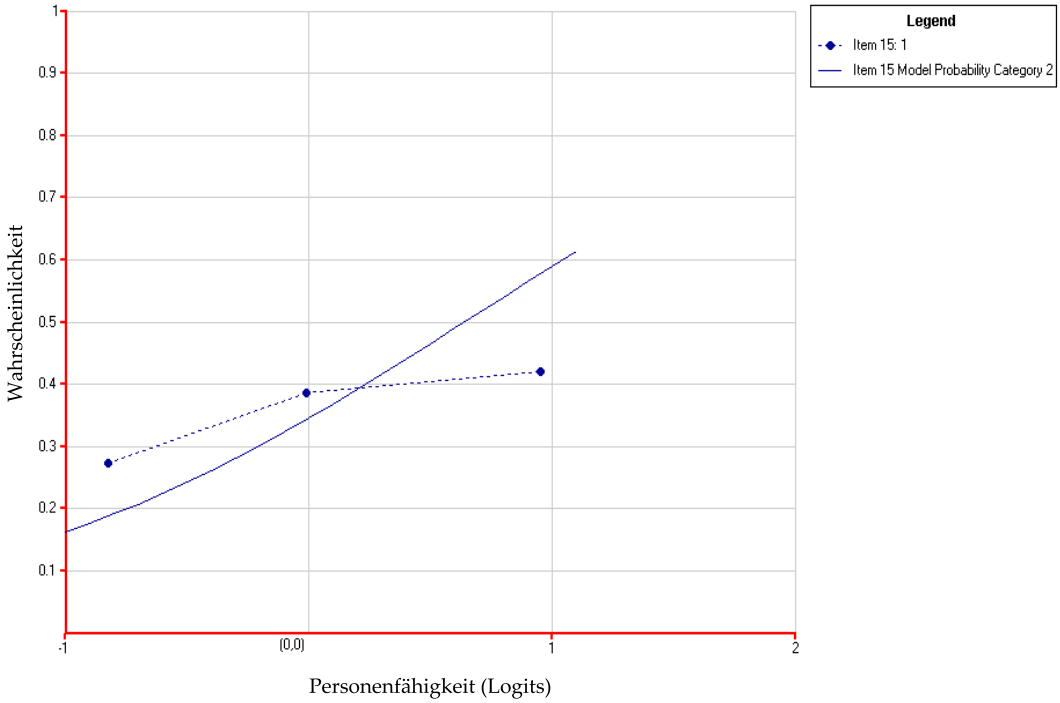


Abbildung 5.8: Erwartete (Model) und beobachtete ICC des Items *KLe3*

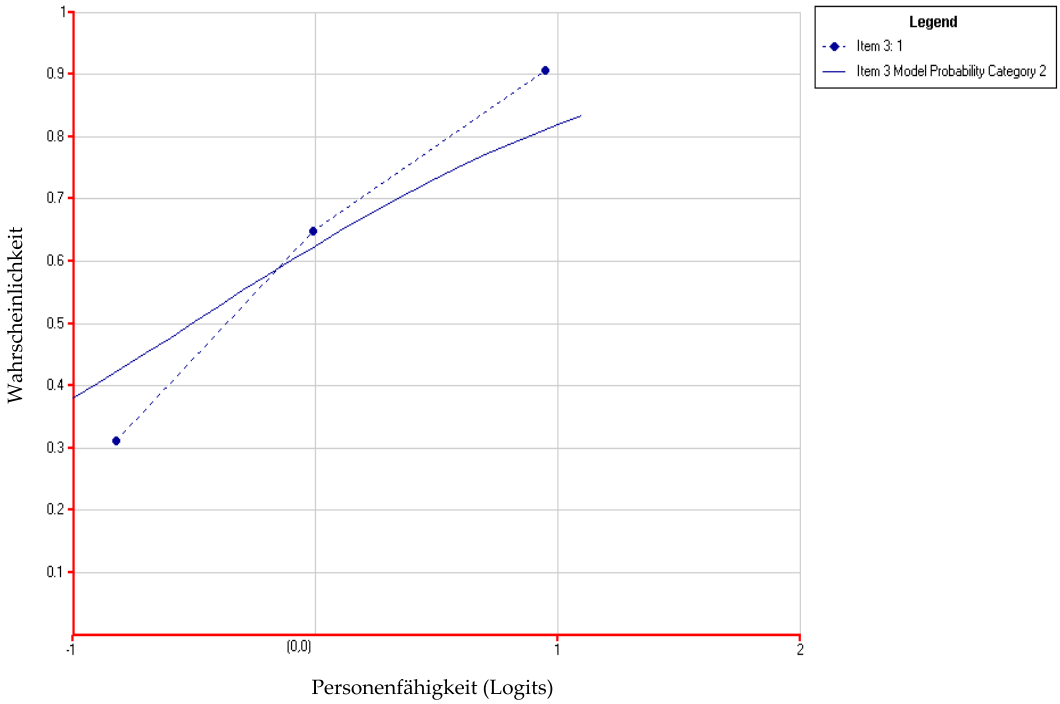


Abbildung 5.9: Erwartete (Model) und beobachtete ICC des Items Bro3

tive-Lab-Interviews für die Generierung möglicher Gründe für die Abweichung dieser Items herangezogen.

ANALYSE DER ANTWORTALTERNATIVEN

Da Itemschwierigkeit und Itemtrennschärfe von der Aufgabe als Ganzes abhängen, können ungenügende Kennwerte sowohl auf Probleme mit dem Itemstamm als auch auf Probleme mit den Antwortalternativen zurückzuführen sein. Um dies zu prüfen, werden insbesondere die Items genauer betrachtet, die sich hinsichtlich ihrer Kennwerte als ungenügende erwiesen haben. Allerdings müssen auch die Items mit guten Kennwerten einer Analyse in Bezug auf folgende Punkte unterzogen werden:

- Die Wahl der richtigen Antwortalternative sollte durch die im Durchschnitt fähigsten Testpersonen erfolgen und
- die Distraktoren sollten in etwa gleich häufig gewählt werden (Lienert & Raatz, 1998).

Eine Aufstellung der Werte aller Items findet sich in Anhang D.1.1, S. 251-253. Zur Veranschaulichung der Datenauswertung werden an dieser Stelle drei Beispiele gegeben, die ein hinsichtlich der beiden oben genannten Punkte gelungenes Item und zwei weniger gut gelungene Items zeigen. Tabelle 5.8 stellt einen Ausschnitt der Ergebnisse dar. Hier sind die Items jeweils mit ihren vier Antwortalternativen dargestellt. Die unter *Score* mit 1 bezifferte Antwortalternative stellt die Lösung dar, die übrigen die Distraktoren sowie die mit dem Label 9 gekennzeichneten fehlenden Werte. Die folgenden beiden Spalten stellen die Anzahl und den Prozentsatz der Personen dar, die sich für die jeweilige Alternative entschieden haben. Unter *Pt Bis* wird die punktbiseriale Korrelation der Antwortalternative mit dem Gesamttestwert ausgewiesen. Die punktbiseriale Korrelation der richtigen Antwortalternative mit dem Gesamtteswert stellt die Trennschärfe der Aufgabe dar. *PV1Avg* drückt die durchschnittliche Fähigkeit der Personen aus, die sich für die jeweilige Alternative entschieden haben.

Das Item *Hur3* stellt ein positives Beispiel dar. Es ist zu erkennen, dass die im Durchschnitt fähigsten Testpersonen sich für die richtige Antwortalternative entschieden haben. Auf die Lösung der Aufgabe entfallen die meisten Antworten und die übrigen Antworten verteilen sich gleichmäßig auf die Distraktoren.

Itembeispiel *KLe2* steht für ein weniger gelungenes Item. Hier entfallen die meisten Antworten auf den Distraktor a. Die fähigsten Personen entscheiden sich zwar

Tabelle 5.8: Analyse der Antwortalternativen

Item	Label	Score	Count	% of tot	Pt Bis	PV1Avg:1
Hur3	a	0.00	50	15.87	-0.21	-0.64
	b	1.00	143	45.40	0.40	-0.08
	c	0.00	64	20.32	-0.16	-0.53
	d	0.00	51	16.19	-0.10	-0.50
	9	0.00	7	2.22	-0.13	-0.69
⋮						
KLe2	a	0.00	169	53.65	0.29	-0.17
	b	0.00	67	21.27	-0.25	-0.66
	c	0.00	46	14.60	-0.20	-0.65
	d	1.00	26	8.25	0.11	-0.05
	9	0.00	7	2.22	-0.01	-0.35
KLe3	a	0.00	103	32.70	0.11	-0.22
	b	0.00	64	20.32	-0.18	-0.61
	c	1.00	113	35.87	0.14	-0.22
	d	0.00	25	7.94	-0.14	-0.61
	9	0.00	10	3.17	-0.03	-0.53

für die richtige Alternative, sie stellen aber den geringsten Prozentsatz an Antworten dar. Dies ist der Hintergrund der weiter oben beschriebenen Itemkennwerte. Wie dort bereits erwähnt, stellt *KLe2* ein extrem schweres Item dar. Nur die Fähigsten können es lösen und aufgrund der dadurch eingeschränkten Differenzierung fällt die Trennschärfe des Items extrem unterdurchschnittlich aus.

Eine andere Art der ungünstigen Antwortverteilung wird anhand des Items *KLe3* offenbar. Hier teilt sich die Gruppe der fähigsten Personen auf die Antwortalternativen a und c auf. Zwar entscheidet sich die Hälfte dieser Personen für die richtige Antwortalternative c, aber der anderen Hälfte erscheint Distraktor a attraktiver.

Auf die beschriebene Weise wurden alle Items untersucht, um Gründe für unzureichende Kennwerte zu finden. Im Anschluss wurden zusammenfassende Itembeurteilungen erstellt. Unter Berücksichtigung dieser Bewertungen wurden mit Blick auf die einzelnen Aufgaben vor der Durchführung von *Expertenpanel* und *Cognitive-Lab* erste Ideen generiert, in welcher Hinsicht die betreffenden Items überarbeitet werden sollten. Die Tabellen D.4-D.6 (S. 254 - 256) im Anhang geben Auskunft über die einzelnen Itembewertungen.

5.3.3 AUSWERTUNG DER SUBJEKTIVEN EINSCHÄTZUNGEN

Das zweite *Expertenpanel* und die zweite Runde der *Cognitive-Lab-Interviews* wurden erneut gezielt dazu genutzt, konkrete Ansätze zur Überarbeitung der Items zu finden. Um von den Expertinnen und Experten zielgerichtete Vorschläge zu erhalten, wurden ihnen Itemkennwerte und eine Zusammenfassung der Antwortanalysen zur Verfügung gestellt.

Die *Cognitive-Lab-Interviews* wurden auf ähnliche Weise wie in der Präpilotphase durchgeführt. Anstatt Einzelinterviews durchzuführen, bearbeitete hier jedoch eine Gruppe von Schülerinnen und Schülern die Aufgaben nach der Laut-Denken-Methode.

Die folgenden Darstellungen beziehen sich der Übersichtlichkeit halber erneut auf ein ausgewähltes Aufgabenbeispiel, um zu verdeutlichen, inwiefern die Kommentare von Expertinnen und Experten sowie von Schülerinnen und Schülern zur Überarbeitung der Items beigetragen haben.

EXPERTENKOMMENTARE

Die Expertenurteile zur Bewertung der Feldtestitems entstammen einer heterogenen Gruppe aus sechs Personen:

- 1 Physik-/Mathematiklehrerin, 1 Chemielehrer
- 1 Student (Lehramt Physik/Mathematik), 1 Studentin (Psychologie)
- 1 Buchhändlerin und 1 Architektin

Das Item, welches hier sowohl für die Experten- als auch für die Schülerkommentare als Beispiel herangezogen wird, ist das in Abbildung 5.10 dargestellte Item *Bro2*. Die Bewertung des Items ist in Tabelle 5.9 dargestellt. Folgende Punkte wurden von den Expertinnen und Experten bemängelt:

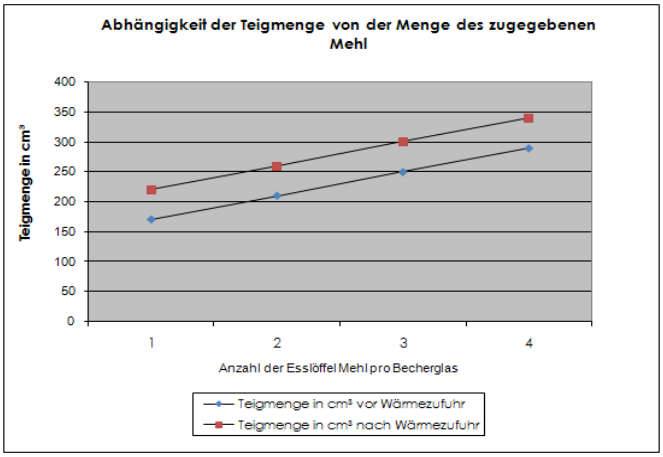
- Die Parallelität der Datenlinien wird von den Schülerinnen und Schülern nicht erkannt.
- Der Abstand zwischen den beiden Datenlinien ist nicht präzise genug ablesbar.
- Einführungstext und Distraktoren sind nicht einheitlich formuliert: Im Einführungstext wird davon gesprochen, dass *„der Teig um so stärker aufgeht“*, während im Rahmen der Distraktoren davon gesprochen wird, dass *„die Teigmenge um so stärker zunimmt“*.

Brot backen II

Eine Gruppe von Bäckern hat die Vermutung, dass der Teig um so stärker aufgeht, je mehr Mehl dem Teig hinzugegeben wird. In vier unterschiedliche Bechergläser wird die gleiche Menge an Wasser, Zucker und Hefe gegeben. Die Bechergläser unterscheiden sich lediglich in der Menge des zugegebenen Mehls:

- 1 Esslöffel Mehl in Becherglas 1
- 2 Esslöffel Mehl in Becherglas 2
- 3 Esslöffel Mehl in Becherglas 3
- 4 Esslöffel Mehl in Becherglas 4

Nach der Mischung der Zutaten werden alle Gläser für eine halbe Stunde der gleichen Wärme ausgesetzt. Folgende Abbildung zeigt das Ergebnis dieses kleinen Versuchs:



Was bedeuten die dargestellten Ergebnisse im Hinblick auf die Vermutung der Bäcker?

a) Die Vermutung lässt sich nicht bestätigen, da die Abbildung dazu keine Aussage ermöglicht.
b) Die Vermutung bestätigt sich, da die Teigmenge um so stärker zunimmt, je mehr Mehl der Teig enthält.
c) Die Vermutung bestätigt sich, da die Teigmenge nach der Wärmezufuhr größer ist.
d) Die Vermutung bestätigt sich nicht, da die Teigmenge nicht um so stärker zunimmt, je mehr Mehl der Teig enthält.

Abbildung 5.10: Feldtest-Version der Aufgabe *Brot backen II (Bro2)*

- Ein Balkendiagramm wäre zur Darstellung eventuell vorteilhafter.

- Die Mehlmenge wird auch für das Aufgehen des Teiges verantwortlich gemacht. Grundsätzliche Frage nach dem Verständnis des *Aufgehens*.

Tabelle 5.9: Bewertung des Items *Brot backen II (Bro2)*

Item (Trennschärfe)	**Bewertung**
Bro2 (0,30) δ = 0,69	- Itemtrennschärfe ist nicht ausreichend - Item ist relativ schwierig - die richtige Antwort d wird von Personen mit der durchschnittl. höchsten Fähigkeit gewählt - Distraktor b zieht die meisten Antworten auf sich - Distraktoren a und c sehen gut aus
Bewertung:	- Distraktor b sollte überarbeitet werden; es sollte untersucht werden, was die Aufgabe so schwierig macht

Die Kommentare der Expertinnen und Experten beziehen sich unter Berücksichtigung der Itembeurteilung zum einen auf die Graphik und zum anderen auf die Textpassage, die das Aufgehen oder Vermehren des Teiges betrifft. Die Schwierigkeit der Aufgabe kann tatsächlich zum einen darin begründet sein, dass die Schülerinnen und Schüler die Graphik nicht verstehen. Da die Itembeurteilung jedoch auch darauf eingeht, dass Distraktor b zu häufig gewählt wird und von den Expertinnen und Experten bemängelt wurde, dass in Distraktor b der Ausdruck, *die Teigmenge nimmt um so stärker zu* als ungünstig bezeichnet werden kann, sollte das Augenmerk während der *Cognitive-Lab-Interviews* auch auf diesen Punkt gelegt werden.

COGNITIVE LAB

Die Gruppe, die im Rahmen der *Cognitive-Lab-Interviews* befragt wurde, setzte sich aus zwei fünfzehnjährigen Realschülerinnen und einem fünfzehnjährigen Realschüler zusammen. Die Jugendlichen haben sich freiwillig für das Interview gemeldet.

Während der Bearbeitung der Aufgabe fiel auf, dass weder der Schüler noch die Schülerinnen die richtige Antwortalternative d wählten: der Schüler wählte Alternative *c*, eine der Schülerinnen wählte Alternative *c*, die andere Alternative *b*. Die Schülergruppe gab bezüglich des Items *Bro2* folgende Kommentare ab:

- Ich habe *c* gewählt, weil man in der Abbildung sieht, dass die Teigmenge nach der Wärmezufuhr größer ist und auch gleichmäßig geblieben ist. Vorher war es niedriger und auch gleichmäßig. Das mit dem Becherglas hat mich irritiert, das hat damit überhaupt nichts zu tun. Habe nicht verstanden, warum die Datenpunkte unterschiedliche Symbole haben.

- Am Anfang wusste ich gar nichts damit anzufangen, habe mich aus den gleichen Gründen für *c* entschieden.

- Weiß nicht ganz genau, warum ich *b* genommen habe. Wenn mehr Mehl drin ist, wird der Teig eben immer größer.

Am ersten Kommentar wird deutlich, dass der Expertenkommentar, die Schülerinnen und Schüler würden die Parallelität der Datenlinien nicht erkennen, hier nicht zutrifft. Der Schüler hat die Parallelität durchaus erkannt, dennoch wählt er die falsche Alternative *c*. Er konzentriert sich nur darauf, dass die Teigmenge nach Wärmezufuhr größer ist und lässt dabei außer Acht, dass die Vermutung geprüft wird, die Teigmenge nehme um so mehr zu, je mehr Mehl hinzugegeben wird. Er hat also nicht verstanden, dass die Teigmenge über die Mehlzugabe hinaus zunehmen sollte, die rote Datenlinie also stärker ansteigen sollte, wenn die Vermutung stimmt. Wichtig ist hinsichtlich der Überarbeitung der Aufgabe, dass der Schüler durch die unterschiedliche Darstellung der Datenpunkt-Symbole irritiert ist. In diesem Zusammenhang ist auch der Expertenkommentar zu beachten, dass eventuell zur besseren Verständlichkeit Balkendiagramme verwendet werden sollten.

Einen weiteren wichtigen Hinweis für die Überarbeitung des Items liefert hier auch der dritte Kommentar. Die Schülerin gibt an, sie habe *b* gewählt, da die Teigmenge durch die Mehlzugabe immer größer wird. Diese Aussage passt zur Kritik eines der Experten. Es ist also wahrscheinlich, dass der Ausdruck, *die Teigmenge nehme zu*, hier missverständlich ist, denn sie nimmt ja tatsächlich schon allein durch die Mehlzugabe zu. Dass die Teigmenge *um so stärker* zunehmen soll, je mehr Mehl hinzugegeben wird, ignoriert die Schülerin an dieser Stelle.

Folgende Schlussfolgerungen wurden hinsichtlich der Überarbeitung aus der Kombination der Experten- und Schülerkommentare gezogen:

- Die Graphik wird so verändert, dass die Teigmenge in Form eines Balkendiagramms dargestellt wird.

- Die Distraktoren *b* und *d* werden überarbeitet. Der Ausdruck, dass *„die Teigmenge um so stärker zunimmt“* wird ersetzt durch den Ausdruck, dass *„der Teig um so stärker aufgeht“*.

Nach dieser Methode wurden alle Items untersucht und überarbeitet, die gemäß den Tabellen D.4 - D.6 (S. 254 - 256) farbig gekennzeichnet sind.

5.3.4 KONSEQUENZEN DER AUSWERTUNG

Nach Auswertung aller hier dargestellten Datenquellen wurden die in Tabelle 5.10 dargestellten Items aufgrund der Feldtestdaten überarbeitet. Es fällt auf, dass auch Items überarbeitet wurden, die gemäß den Itembewertungen keiner Überarbeitung bedurften. Dies betraf die Items *Hur2*, *Sch1* und *Reg2*. Sie wurden überarbeitet, weil sich im Rahmen der wiederholten Durchsicht der Items noch unscharfe Formulierungen und insbesondere qualitative Mängel der Graphiken feststellen ließen. Die Überarbeitungen waren lediglich geringfügiger Natur. Umgekehrt ließ sich wiederum das Item Sko1, das gemäß der Bewertung (s. Tabelle D.6, S. 256) verändert werden sollte, nicht überarbeiten, da sich keine Ansätze zur Verbesserung finden ließen. Insgesamt wurden 21 der 30 Testitems mehr oder minder umfangreich überarbeitet. Inwiefern sich die Items verbessert haben, kann anhand der Itemkennwerte zwischen Feld- und Haupttest nicht direkt verglichen werden, weil es sich um zwar ähnlich zusammengesetzte, aber unterschiedliche Stichproben handelt und weil der Großteil der Items verändert worden ist. Die Items werden demnach in den folgenden Darstellungen der Haupttestergebnisse gesondert und nicht im Vergleich zum Feldtest betrachtet.

5.4 ERGEBNISSE DES HAUPTTESTS

Das Ziel dieser Arbeit besteht darin, einen Test zu entwickeln, der ein eindimensionales Screening der prozessbezogenen naturwissenschaftlichen Grundbildung ermöglicht. Gemäß dieser Zielsetzung wurden Testitems entwickelt und im Rahmen des Feldtests statistisch geprüft. Der Modellgeltungstest im Feldtest führte hinsichtlich der Dimensionalität zu keiner klaren Entscheidung für das ein- oder dreidimensionale Modell. Das dreidimensionale Modell zeigte sich leicht im Vorteil, doch die hohen Korrelationen der latenten Dimensionen legten letztendlich eine eindimensionale Betrachtung der Daten nahe.

Bevor die Linie dieser eindimensionalen Betrachtung im Haupttest fortgesetzt werden kann, ist zunächst eine weitere Modellgeltungsprüfung notwendig, die sich direkt an die Beschreibung der Haupttestst ichprobe anschließen wird. Die vorliegenden Daten werden auf der Grundlage des Modells betrachtet, das die Daten am besten erklärt.

Ziel der Haupttestauswertungen ist anders als noch im Feldtest eine abschließende Entscheidung darüber, welche Items und, auf nächsthöherer Ebene, welche Itemsets sich endgültig als unzureichend erweisen und demzufolge aus dem Test entfernt werden sollten. Nach der Veränderung der Testitems auf Grundlage der Feldtestda-

Tabelle 5.10: Konsequenzen der Feldtestauswertung

Item	Überarbeitung
Bro1	- Itemstamm überarbeitet: Inhalte präzisiert und kürzer dargestellt - Distraktoren c und d überarbeitet
Bro2	- Graphik verändert: : Liniendiagramm in Balkendiagramm umgewandelt - Distraktor b und Lösung d überarbeitet
Hur1	- Graphik verändert: beispielhafte Beschreibung von Datenpunkten ergänzt - Distraktoren a, und d sowie Lösung c überarbeitet
Hur2	- Graphiken verändert: Qualität verbessert, Daten sind besser ablesbar
Sch1	- Itemstamm überarbeitet: Beschreibung der Abbildungen präzisiert
Sch2	- Distraktoren a, b und c überarbeitet
Sch3	- Distraktor a überarbeitet
Sti 1	- Itemstamm überarbeitet: Beschreibung der Annahme präzisiert
Sti2	- Distraktor d überarbeitet
Sti3	- Itemstamm überarbeitet: Stichlingattrappe 3 verändert - Distraktor c überarbeitet - Distraktoren a und d sowie b und c getauscht
KLe1	- Itemstamm überarbeitet: gekürzt
KLe2	- Distraktoren a und b überarbeitet
KLe3	- Itemstamm überarbeitet: Text deutlich gekürzt, überflüssige Infos entfernt - Distraktor a und Lösung c überarbeitet - Graphiken verändert: Qualität verbessert, Daten sind besser ablesbar
Kli 1	- Itemstamm überarbeitet: Beschreibung der Untersuchung präzisiert
Reg1	- Itemstamm überarbeitet: Beschreibung der Untersuchung präzisiert
Reg2	- Itemstamm überarbeitet: Beschreibung der Untersuchung präzisiert - Abbildungen im Itemstamm überarbeitet
Reg3	- Graphik verändert: Bedeutung der Balken verändert
Ros3	- Distraktor d überarbeitet
Sko3	- Itemstamm überarbeitet: Beschreibung der Untersuchung präzisiert - Distraktor a und Lösung c überarbeitet
Sol1	- Distraktor d überarbeitet
Sol3	- Itemstamm überarbeitet: Beschreibung der Vermutung präzisiert - Distraktor a und Lösung c überarbeitet

rot = Überarbeitung dringend notwendig
orange = Überarbeitung notwendig
grün = Überarbeitung nicht unbedingt notwendig
schwarz = Überarbeitung nicht notwendig

ten kommt dem Haupttest, wie schon dem Feldtest, also zunächst die Aufgabe einer allgemeinen Prüfung der einzelnen Items zu.

Die Darstellung der Haupttestergebnisse gliedert sich derart, dass, ausgehend von einer Beschreibung der Stichprobe, welche die Daten für den Haupttest lieferte, die Prüfung der Modellgeltung und die deskriptive Betrachtung des Itempools erfolgen. Dieser wird zunächst auf Basis der einzelnen Items und pro Schulniveau auf den Anteil fehlender Werte geprüft, um festzustellen, ob einzelne Items eine überzufällige Häufung fehlender Werte aufweisen. Im Anschluss daran werden die klassischen und probabilistischen Itemkennwerte dargestellt, auf deren Grundlage die Itemsets bewertet und gegebenenfalls aus dem Test entfernt werden.

Der um qualitativ unzureichende Items reduzierte Itempool wird nachfolgend einer ersten Validierung unterzogen. Diese Validierung sowie die Prüfung der Reliabilität bilden den Abschluss dieses Abschnitts. Den Betrachtungen der abschließenden Testversion sowie der Prüfung der Gütekriterien anhand dieser Version kommt die Aufgabe eines Ausblicks zu. Die Daten können erst als gesichert gelten, nachdem sie an einer weiteren Stichprobe kreuzvalidiert wurden.

5.4.1 STICHPROBENBESCHREIBUNG

Den Ausgangspunkt der Haupttest-Ergebnisdarstellung bildet die Beschreibung der Stichprobe, die den folgenden Auswertungen als Basis diente. Tabelle 5.11 gibt Auskunft über die Zusammensetzung der Stichprobe. Die Tabelle zeigt, dass die Stich-

Tabelle 5.11: Beschreibung der Haupttest-Stichprobe

		H	R	G	Summe
Stichprobe	**n**	250	250	250	750
Geschlecht	**m**	126	126	121	373
	w	123	124	128	375
Fehlende Angabe		1	0	1	2
Anzahl Schulen		7	4	3	14
Anzahl Klassen		18	15	14	47

n = Stichprobengröße H = Hauptschule R = Realschule G = Gymnasium

probengrößen pro Schulform gleich groß sind, was durch die zufällige Ziehung gleich großer Unterstichproben bedingt ist. Die Geschlechteraufteilung ist annähernd gleich,

so dass sowohl die Unterstichproben als auch die Gesamtstichprobe diesbezüglich als ausgeglichen bezeichnet werden können.

Zur Datenerhebung wurden knapp doppelt so viele Hauptschulen benötigt wie Realschulen und Gymnasien, um die notwendigen Schülerzahlen zu erreichen. Hinsichtlich der Anzahl der pro Schulform getesteten Klassen gestalten sich die Zahlen ausgeglichener.

GLOBALE MODELLPRÜFUNG

Bevor die Item- und Personenkennwerte des Haupttests einer Prüfung unterzogen werden konnten, war es zunächst notwendig, eine weitere Prüfung der Modellselektionsmaße CAIC und BIC vorzunehmen, um zu klären, vor dem Hintergrund welchen Modells die Haupttestdaten betrachtet und bewertet werden sollen.

Tabelle 5.12 stellt die Modellselektionsmaße CAIC und BIC des eindimensionalen Modells denen des dreidimensionalen Modells gegenüber. Wie die Tabelle zeigt, kann

Tabelle 5.12: Globale Modellprüfung anhand des CAIC und BIC

Modell	unabh. Paramter	Deviance	CAIC	BIC
Eindimensional	31	27114,1	27350,4	27319,4
Dreidimensional	36	27022,2	27296,6	27260,6

Deviance = globale Fit-Statistik CAIC = Consistent Akaike´s Information Criterion
BIC = Bayes Information Criterion

hier das dreidimensionale Modell als überlegen bezeichnet werden.

Da bereits der Feldtest ergeben hat, dass die latenten Dimensionen des leicht favorisierten dreidimensionalen Modells hoch miteinander korrelieren, wurden diese auch im Haupttest auf die Höhe ihrer Korrelation geprüft. Gemäß Tabelle 5.13 fallen die latenten Korrelationen der drei Dimensionen höher aus als im Feldtest ($> 0,90$). Der Argumentation des Feldtests folgend, ist es aufgrund dieser hohen Korrelationen angemessen, das Konstrukt der *prozessbezogenen naturwissenschaftlichen Grundbildung* als eindimensional zu betrachten. Würde man die drei Dimensionen einzeln analysieren und die drei Fähigkeiten, die eigentlich als Indikator für die Ausprägung der *prozessbezogenen naturwissenschaftlichen Grundbildung* dienen, damit differenziert erfassen wollen, so sähe man sich darüber hinaus dem Problem unzureichender WLE-Reliabilitäten (vgl. Abschnitt 3.2.2) ausgesetzt:

- Identifizieren wissenschaftlicher Hypothesen: $r = 0,52$

Tabelle 5.13: Latente Korrelationen des dreidimensionalen Modells

	Dimension 1	Dimension 2
Dimension 1		
Dimension 2	0,93	
Dimension 3	0,96	0,93

- Planen einer wissenschaftlichen Untersuchung: $r = 0,65$
- Nutzen wissenschaftlicher Ergebnisse: $r = 0,48$

Daraus folgt, dass die im Weiteren dargestellten probabilistischen Parameter vor dem Hintergrund des eindimensionalen Testmodells geschätzt und auf dieser Basis bewertet werden. Dieses Vorgehen wird in der Diskussion der Haupttest-Ergebnisse (s. Abschnitt 6.1.3, S. 226) noch einmal aufgegriffen.

5.4.2 DER ITEMPOOL

Vor dem Hintergrund der Modellbetrachtungen erfolgt nun die Beschreibung der Itemstichprobe. Der Itempool (n=30), der den Schülerinnen und Schülern jeweils in kompletter Form vorgelegt wurde, wird im Folgenden zunächst auf den Anteil fehlender Werte untersucht. Die Beschreibung und Bewertung der Itemkennwerte schließt sich diesen Betrachtungen an.

FEHLENDE WERTE

Im Rahmen der Qualitätsprüfung der Haupttestitems ging es darum, differenzierter als noch im Feldtest festzustellen, ob es zu einer systematischen Häufung fehlender Werte kommt. Zu diesem Zweck wurden zunächst die fehlenden Werte pro Item und im Anschluss die fehlenden Werte pro Schulniveau betrachtet.

Wie schon im Feldtest ausgeführt, wird hier ein Kriterium für den Anteil fehlender Werte von $< 5\%$ angelegt. Items, die diesen Prozentsatz überschreiten oder die wiederholt durch einen erhöhten Anteil fehlender Werte auffallen, sollten kritisch daraufhin überprüft werden, welche Merkmale zu diesem hohen Anteil fehlender Werte geführt haben könnten.

Die folgende Tabelle 5.14 gibt einen Überblick über die Anzahl und den Anteil der fehlenden Werte, die hier pro Item sowie aggregiert auf der Ebene der einzelnen

Fähigkeiten aufgeführt sind. Acht der insgesamt dreißig Items weisen mehr als 2% fehlende Werte auf und lediglich zwei Items (*Sch2 und Sti2*) zeigen mehr als 3% Prozent fehlende Werte. Das Item *Sch2 (Schimmel II)* fiel bereits im Feldtest durch einen gegenüber den übrigen Items erhöhten Anteil fehlender Werte auf. Untersucht man das Item (s. Anhang A.2) auf mögliche Ursachen, so lassen sich dafür keine formalen Anhaltspunkte finden, da es keine Graphiken oder Tabellen enthält und kurz und einfach formuliert ist. Der einzige Grund, der hier denkbar wäre, ist der, dass der Aufgabeninhalt die Schülerinnen und Schüler nicht anspricht und ihnen diese Aufgabe zu langweilig erscheint.

Der Anteil fehlender Werte bleibt im Falle aller Items, bezogen auf die Gesamtstichprobe von 750 Testpersonen, unter fünf Prozent. Es besteht also auf dieser Basis kein Grund dafür, Items aus dem Test auszuschließen.

Zusammengefasst zu Fähigkeitskomplexen wird ein Blick auf die Verteilung der fehlenden Werte auf diese Bereiche ermöglicht. Hier wird deutlich, dass auf den Komplex *Identifizieren wissenschaftlicher Hypothesen* mit 155 mehr fehlende Werte entfallen als auf die beiden anderen. Die Komplexe *Planen einer wissenschaftlichen Untersuchung* und *Nutzen wissenschaftlicher Ergebnisse* zeigen mit 114 und 109 annähernd gleich viele fehlende Werte. Der Prozentsatz fehlender Werte fällt insgesamt sehr gering aus. Die Abweichung des Fähigkeitsbereichs *Identifizieren wissenschaftlicher Hypothesen* ist daher unbedenklich. Dennoch wird dieser Unterschied vor dem Hintergrund aller Ergebnisse in Abschnitt 6.1.2 noch einmal diskutiert, um mögliche Fehlerquellen in der Itementwicklung aufzuzeigen, die für zukünftige Item- und Testentwicklungen relevant sein könnten.

Tabelle 5.15 ermöglicht einen differenzierten Blick auf die nach Schulniveau differenzierten fehlenden Werte. Insgesamt betrachtet fällt die Anzahl fehlender Werte sowie der Anteil fehlender Werte pro Schulniveau gemäß dem angelegten 5%-Kriterium unbedenklich aus. Vergleicht man jedoch die Schulniveaus, so zeigen sich Unterschiede. Insbesondere die Hauptschule fällt hier im Vergleich zu den anderen Schulformen durch einen erhöhten Anteil fehlender Werte auf. Die Hauptschule hat mit 2,44% einen um über ein Prozent höheren Anteil fehlender Werte als die Realschule (1,41%), die leicht über dem Anteil fehlender Werte des Gymnasiums (1,19%) liegt. Der Kruskall-Wallis-Test zur nicht-parametrischen Berechnung von Gruppenunterschieden weist mit einem χ^2-Wert von $21,87$ $(p < .001)$ aus, dass es signifikante Unterschiede zwischen den Schulniveaus hinsichtlich des Anteils fehlender Werte gibt. In Abschnitt 6.1.2 wird diskutiert, welche Ursachen diese Unterschiede haben könnten.

Tabelle 5.14: Fehlende Werte der Haupttestst ichprobe pro Item und Fähigkeitskomplex

Items	fehlende Werte n	%
Bro 1	8	1,1
Bro 2	9	1,2
Bro 3	12	1,6
Hur 1	7	0,9
Hur 2	21	2,8
Hur 3	17	2,3
Sch 1	4	0,5
Sch 2	28	3,7
Sch 3	12	1,6
Sti 1	4	0,5
Sti 2	25	3,3
Sti 3	9	1,2
KLe 1	4	0,5
KLe 2	12	0,5
KLe 3	20	1,6
Kli 1	7	0,9
Kli 2	13	1,7
Kli 3	14	1,9
Reg 1	5	0,7
Reg 2	16	2,1
Reg 3	20	2,7
Ros 1	11	1,5
Ros 2	10	1,3
Ros 3	17	2,3
Sko 1	6	0,8
Sko 2	6	0,8
Sko 3	17	2,3
Sol 1	8	1,1
Sol 2	16	2,1
Sol 3	20	2,7

Fähigkeitskomplex	fehlende Werte aggregiert pro Fähigkeit n	%
Identifizieren wissenschaftlicher Hypothesen	155	2,1
Planen einer wissenschaftlichen Untersuchung	114	1,5
Nutzen wissenschaftlicher Ergebnisse	109	1,4

n = absolute Anzahl fehlender Werte

Vergleicht man die Geschlechterstichproben hinsichtlich ihrer fehlenden Werte, so ergeben sich keine großen Unterschiede. Die Schülerstichprobe zeigt 195 (entspricht $1,74\%$) fehlende Wert und die Schülerinnenstichprobe 183 (entspricht $1,64\%$). Die Anteile fehlender Werte fallen sehr gering aus und die Stichproben unterscheiden

Tabelle 5.15: Fehlende Werte pro Person und Schulniveau

Niveau	n	fehlende Werte	fehlende Werte/Person	Anteil fehlender Wert an insgesamt möglichen Item-Antworten in %
Hauptschule	250	183	,73	2,44
Realschule	250	106	,42	1,41
Gymnasium	250	89	,36	1,19

n = Stichprobengröße

sich hinsichtlich der Anzahl fehlender Werte nicht signifikant (χ^2-Wert von $3,558$ $(p > .05)$).

Die dargestellten Daten können so zusammengefasst werden, dass es einzelne auffällige Häufungen fehlender Werte gibt, ihr Anteil aber jeweils als sehr gering zu bewerten ist. Die geringen und relativ ausgeglichenen Anteile fehlender Werte pro Item sprechen dafür, dass sie nicht in besonderen Eigenschaften der Items begründet sind. Den im Verhältnis zu den anderen Schulniveaus erhöhten Anteil fehlender Werte der Hauptschule gilt es noch zu diskutieren.

ITEMKENNWERTE

Zu Beginn dieses Abschnitts soll zunächst anhand von Tabelle 5.16 ein klassisch deskriptiver Überblick über die Mittelwerte (M) und Standardabweichungen (SD) der einzelnen Items sowie der drei Fähigkeitsskalen gegeben werden. Wie bereits im Rahmen der Pilotierung ausgeführt, gibt der Mittelwert aufgrund der binären Kodierung der Aufgabenantworten hier die klassische Itemschwierigkeit an. Je höher der Mittelwert ausfällt, desto mehr Personen haben das Item gelöst und desto leichter es. Eine mittlere Itemschwierigkeit von $0,5$ führt zu einer für die Differenzierung der Testpersonen optimalen Itemvarianz[2]. Dies bedeutet nicht, dass alle Items über die Gesamtstichprobe hinweg eine Schwierigkeit von $0,5$ besitzen sollten. Wie bereits im Feldtest ausgeführt, sollten sich die Itemschwierigkeiten möglichst ausgeglichen über die Personenfähigkeiten der Stichprobe verteilen. Es sollte also für jedes Kompetenzniveau Items mit einer Schwierigkeit von $0,5$ geben. Die Verteilung der Itemschwierigkeiten über die Personenfähigkeiten wird weiter unten im Rahmen der probabilistischen

2 Wegen der Gleichung $Var(x_i) = p_i \cdot (1 - p_i)$ erreicht die Itemvarianz (aus der die hier präsentierte Standardabweichung durch $\sqrt{Var}$ errechnet wird) ihr Maximum bei einer mittleren Itemschwierigkeit von $0,5$. Die Itemvarianz $Var(x_i)$ entspricht dem Produkt der Wahrscheinlichkeit, das Item i zu lösen (p_i), und der Gegenwahrscheinlichkeit, das Item i nicht zu lösen $(1 - p_i)$.

Tabelle 5.16: Deskriptive Beschreibung des Itempools

Items	M	SD
Bro 1	0,50	0,500
Bro 2	0,33	0,469
Bro 3	0,67	0,472
Hur 1	0,33	0,472
Hur 2	0,60	0,490
Hur 3	0,54	0,499
Sch 1	0,38	0,487
Sch 2	0,59	0,492
Sch 3	0,81	0,390
Sti 1	0,59	0,493
Sti 2	0,56	0,497
Sti 3	0,37	0,484
KLe 1	0,32	0,466
KLe 2	0,07	0,258
KLe 3	0,55	0,498
Kli 1	0,46	0,499
Kli 2	0,48	0,500
Kli 3	0,40	0,491
Reg 1	0,50	0,500
Reg 2	0,46	0,499
Reg 3	0,38	0,486
Ros 1	0,68	0,465
Ros 2	0,44	0,497
Ros 3	0,50	0,500
Sko 1	0,58	0,494
Sko 2	0,68	0,467
Sko 3	0,64	0,480
Sol 1	0,62	0,486
Sol 2	0,44	0,497
Sol 3	0,50	0,500

M = Mittelwert SD = Standardabweichungen

Auswertungen noch einmal aufgegriffen. In diesem Rahmen sollte neben der Prüfung einer gleichmäßigen Verteilung insbesondere das Item *KLe2* kritisch betrachtet werden, das wie schon im Feldtest durch eine sehr hohe Itemschwierigkeit auffällt.

Die in Tabelle 5.17 zu Fähigkeitskomplexen zusammengefassten Items (10 Items pro Fähigkeitskomplex) weisen kaum Unterschiede in den mittleren Itemschwierigkeiten auf, sie erscheinen also für die Testpersonen gleich schwer.

Tabelle 5.17: Deskriptive Beschreibung der zusammengefassten Fähigkeitskomplexe

Fähigkeitskomplex	**M**	**SD**
Identifizieren wissenschaftlicher Hypothesen	0,49	0,23
Planen einer wissenschaftlichen Untersuchung	0,48	0,25
Nutzen wissenschaftlicher Ergebnisse	0,50	0,22

M = Mittelwert SD = Standardabweichungen

Als Grundlage einer abschließenden Entscheidung darüber, welche Items aus dem Test entfernt werden sollten, weil sie bezüglich ihrer Trennschärfe oder Schwierigkeit unpassend erscheinen oder aber weil sie das hier angelegte Rasch-Kriterium gleicher Trennschärfen nicht erfüllen, werden nun alle zur qualitativen Beurteilung notwendigen Kennwerte präsentiert. Tabelle 5.18 gibt Auskunft über probabilistische Itemschwierigkeiten, klassische Trennschärfen und den Infit der Haupttest-Items. Items, die abweichende Werte aufweisen, erscheinen fett gedruckt.

Die Betrachtung der *Itemschwierigkeiten* erfolgt vor dem Hintergrund einer gleichmäßigen Verteilung der Itemschwierigkeiten über die Personenfähigkeiten. Die mittlere Personenfähigkeit wurde zur Schätzung der Itemkennwerte auf Null fixiert. Bei einem ersten Blick auf die Itemschwierigkeiten fallen zunächst die Items 14 (*KLe2*) und 9 (*Sch3*) durch extreme Schwierigkeiten auf, die allerdings erst anhand der Wright-Map eingeordnet werden können. Die in Abbildung 5.11 dargestellte Wright-Map verdeutlicht auf der rechten Seite die Lokalisierung der Items im Verhältnis zur Verteilung der Personenfähigkeiten auf der linken Seite. Es ist ersichtlich, dass sich die Itemschwierigkeiten gut über den mittleren Fähigkeitsbereich verteilen und es in Richtung der Randbereiche - sowohl im hohen als auch im niedrigen Fähigkeitsbereich - an Items fehlt. Während Item 14 so schwer ist, dass es nur von den fähigsten Personen gelöst werden kann, und es an Items zur Überbrückung zwischen diesem Punkt und dem mittleren Fähigkeitsbereich fehlt, stellt Item 9 nicht den Extrempunkt des unteren Fähigkeitsbereiches dar. Hier sollte es noch leichtere Items geben, um gut zwischen Personen des untersten Fähigkeitsbereiches differenzieren zu können.

Bezogen auf die hinsichtlich ihrer Leistung sehr heterogenen Stichprobe kann der Test insgesamt mit einer durchschnittlichen Itemschwierigkeit von $0,05$ im Vergleich

Tabelle 5.18: Itemkennwerte des Haupttests

Item-Nr.	Items	Schwierigkeit	Trennschärfe	MNSQ	T-Wert
1	Bro 1	-0,01	0,47	0,96	-1,4
2	Bro 2	0,86	0,52	0,89	-3,0
3	Bro 3	-0,77	0,51	0,90	-3,1
4	Hur 1	0,82	0,36	1,03	0,8
5	Hur 2	-0,43	0,42	1,00	-0,2
6	Hur 3	-0,16	0,40	1,02	0,7
7	Sch 1	0,55	0,51	0,92	-2,6
8	Sch 2	-0,34	0,41	1,00	0,2
9	Sch 3	-1,63	0,36	0,99	-0,2
10	Sti 1	-0,41	0,31	1,08	**2,7**
11	Sti 2	-0,23	0,40	1,03	1,1
12	Sti 3	0,62	0,57	0,88	-3,8
13	KLe 1	0,89	0,36	1,03	0,9
14	KLe 2	2,95	**0,17**	1,06	0,5
15	KLe 3	-0,20	**0,25**	1,11	**3,8**
16	Kli 1	0,19	0,44	0,99	-0,3
17	Kli 2	0,11	0,37	1,04	1,3
18	Kli 3	0,49	0,38	1,05	1,6
19	Reg 1	-0,01	0,31	1,06	**2,3**
20	Reg 2	0,20	0,53	0,90	-3,5
21	Reg 3	0,61	0,33	1,08	**2,4**
22	Ros 1	-0,88	0,51	0,91	-2,6
23	Ros 2	0,28	0,56	0,89	-4,0
24	Ros 3	0,04	0,48	0,98	-0,8
25	Sko 1	-0,38	0,43	1,00	0,0
26	Sko 2	-0,87	0,52	0,89	-3,3
27	Sko 3	-0,64	0,40	0,98	-0,5
28	Sol 1	-0,56	**0,28**	1,10	**3,4**
29	Sol 2	0,31	0,47	0,95	-1,7
30	Sol 3	0,05	0,47	0,95	-1,7
	Mittelwert	0,05	0,42	0,99	

MNSQ = Infit Mean Square T-Wert = Prüfgröße des Infit-Maßes

zur mittleren Personenfähigkeit von 0 als angemessen bezeichnet werden.

Die Bewertung der *Itemtrennschärfen* erfolgt wie schon im Falle des Feldtests vor dem Hintergrund, dass sie zwischen $0,4$ und $0,7$ liegen sollten, um als gut bezeichnet werden zu können (Kelava & Moosbrugger, 2007). Auch an dieser Stelle wird eine gewisse Toleranz eingeräumt, so dass Trennschärfen zwischen $0,3$ und $0,4$ als noch ausreichend betrachtet werden. Auf dieser Grundlage fallen drei Items auf, die unzureichende Werte $< 0,3$ aufweisen: *KLe2* und *KLe3* sowie *Sol1*.

Die durchschnittliche Trennschärfe des Haupttests beträgt $0,42$ und liegt damit in

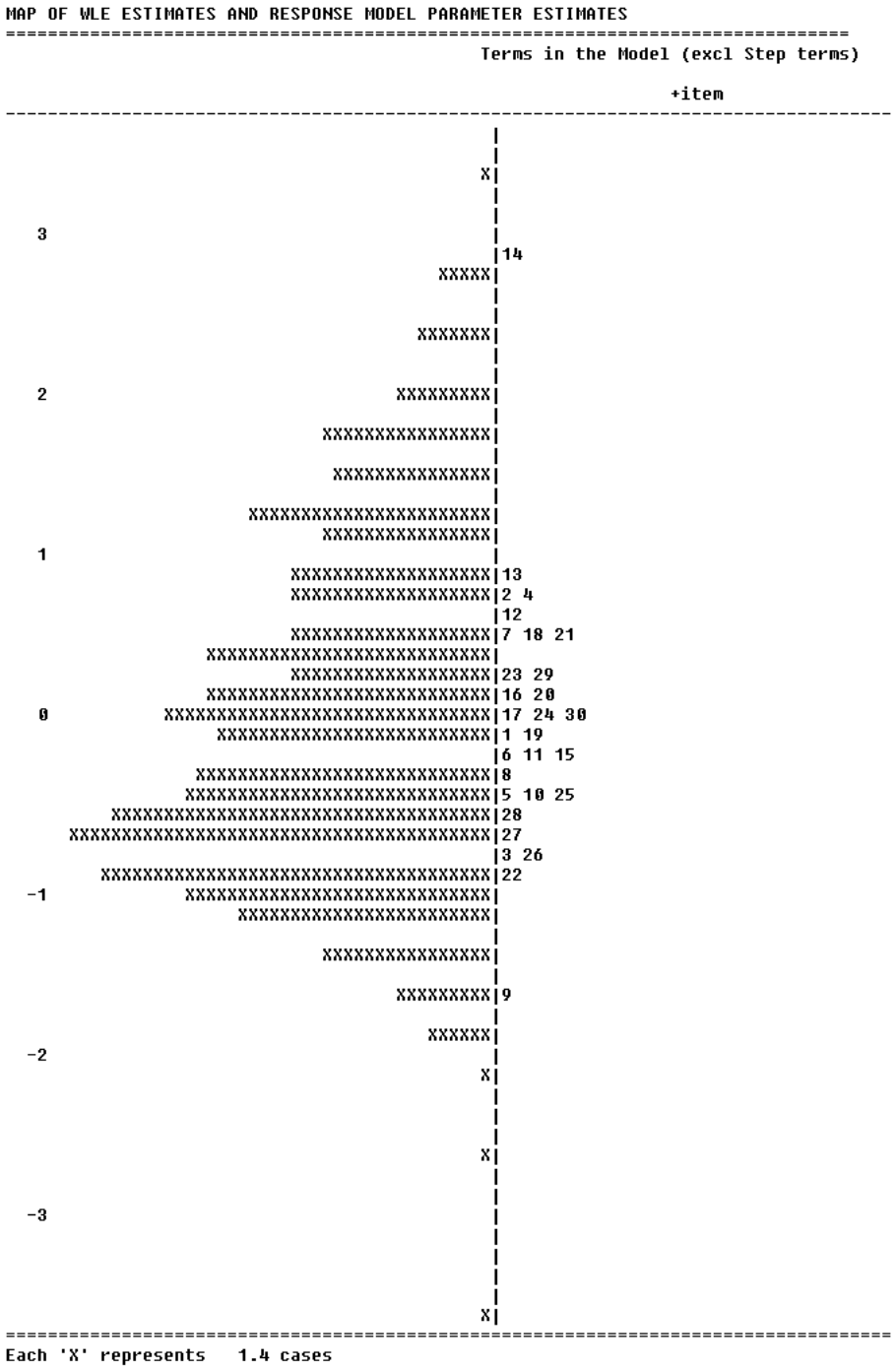


Abbildung 5.11: Verteilung der Itemschwierigkeiten über die Personenfähigkeiten

dem als gut zu bezeichnenden Bereich.

Die T-Werte stellen die Prüfgröße des Infit-Maßes (MNSQ) dar und dienen als mögliche Indikatoren für abweichende Trennschärfen. Als Anhaltspunkt für die Bewertung der T-Werte werden unter Berücksichtigung der Tatsache, dass sie von der Stich-

probengröße abhängen, Werte über $2,0$ und unter $-2,0$ als signifikant betrachtet (vgl. 3.2.1). Abweichungen im negativen Bereich werden eher toleriert als Abweichungen im positiven Bereich, da sie ein Anzeichen für überdurchschnittlich gute Trennschärfen darstellen können. Vier Items fallen durch abweichende positive Werte auf. Dies sind die Items *Sti1* ($T = 2,7$), *KLe3* ($T = 3,8$), *Reg1* ($T = 2,3$), *Reg3* ($T = 2,4$) und *Sol1* ($T = 3,4$).

Zur weiteren Prüfung auffälliger T-Werte schließt sich eine Betrachtung der *MNSQ-Werte* an, die im optimalen Fall bei 1 liegen sollten und für die ein Toleranzbereich zwischen $0,75$ und $1,33$ angelegt wird (R. Adams & Khoo, 1996). Keines der Items kann unter Berücksichtigung dieses Kriteriums als abweichend bezeichnet werden. Alle MNSQ-Werte liegen innerhalb dieser Grenzen, jedoch zeigen die im Bereich der auffälligen T-Werte genannten Items die am stärksten abweichenden Mean Squares. Im Durchschnitt liegt der Mean Square aller Items bei $0,99$, also nahe an dem wünschenswerten Wert von 1.

Um sicher klären zu können, inwiefern die jeweiligen Trennschärfen tatsächlich die Ursache für die abweichenden T-Werte darstellen, werden zum Abschluss dieser Ergebnisdarstellungen die Item-Characteristic-Kurven (ICCs) der jeweiligen Items geprüft. Die Trennschärfe lässt sich als Steigung der Kurve an ihrer steilsten Stelle ablesen. Da alle Items die gleiche Trennschärfe besitzen sollten, sollten sie demnach parallel verlaufen. Ein Anzeichen für nicht modellkonforme Trennschärfen stellt das Abweichen der beobachteten von der per Modell erwarteten ICC dar. Abbildungen 5.12 bis 5.15 zeigen, dass die ICCs der aufgrund abweichender T-Werte auffälligen Items tatsächlich abweichende Trennschärfen aufweisen. Um den Unterschied zu modellkonformen Items darstellen zu können, werden die abweichenden Items den modellkonformen Items der jeweiligen Itemsets gegenübergestellt. Eine Gesamtübersicht über die ICCs aller Items findet sich in Anhang D.2.1 ab Seite 257. Es wurden für die Darstellung Aufgabengruppen gebildet, um die ICCs übersichtlich zu halten.

In Abbildung 5.12 wird das abweichende Item 10 (*Sti1*) dargestellt. Die beobachtete ICC weicht von der erwarteten ab und kreuzt darüber hinaus die ICC des Items 11 (*Sti2*). Auch das Item 15 (*KLe3*) in Abbildung 5.13 weist im Verhältnis zu den beiden übrigen Aufgaben des Sets eine deutliche Abweichung von der erwarteten ICC auf. In Abbildung 5.14 sind die Items 19 (*Reg1*) und 21 (*Reg3*) auffällig. Im Vergleich dazu folgt die beobachtete ICC des Items 20 mit einer sehr geringen Abweichung der erwarteten ICC. Abschließend erweist sich die beobachtete ICC des Items 28 (*Sol1*) in der letzten Abbildung 5.15 im Vergleich zur erwarteten ICC und im Vergleich zu den sehr modellkonform verlaufenden beobachteten ICCs der Items 29 und 30 (*Sol2 und*

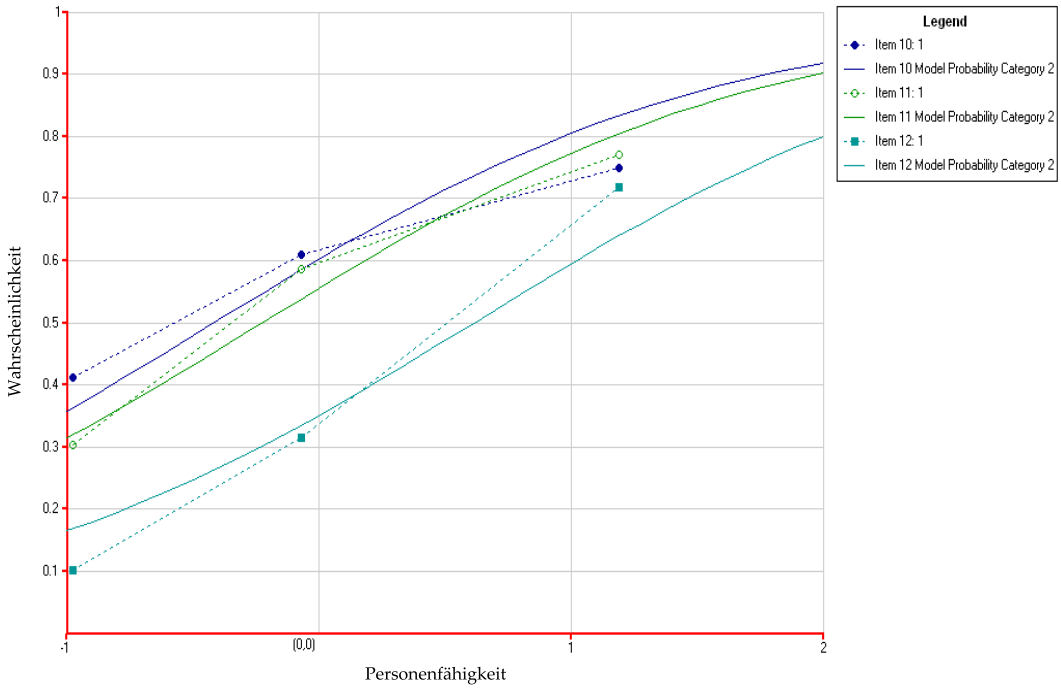


Abbildung 5.12: ICCs der Items 10 bis 12 (*Sti1-3*)

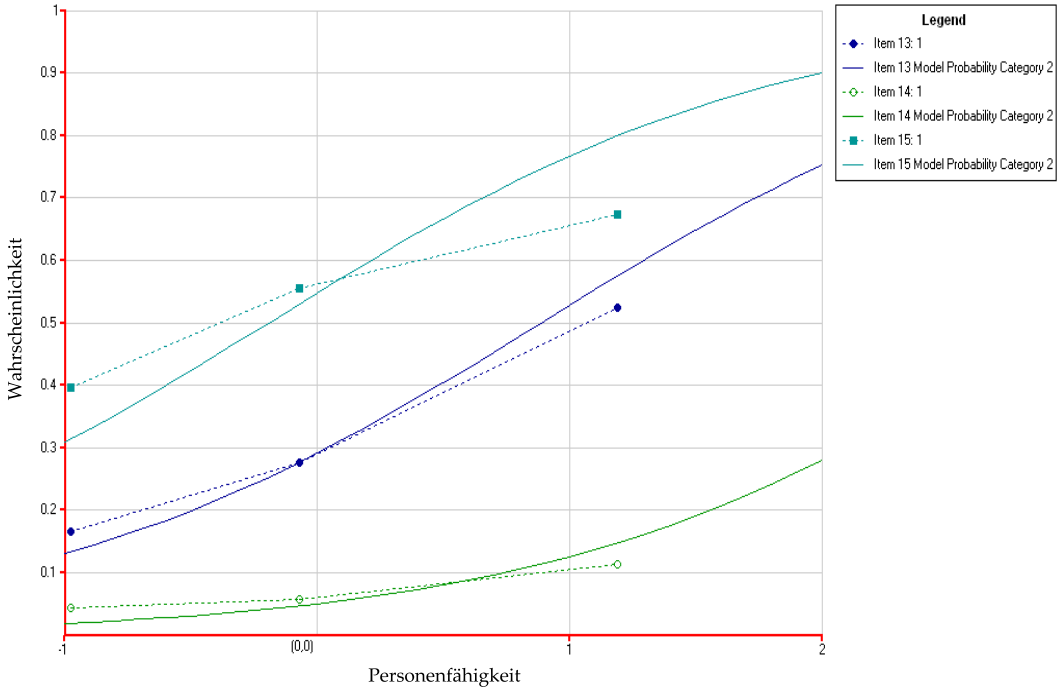


Abbildung 5.13: ICCs der Items 13 bis 15 (*KLe1-3*)

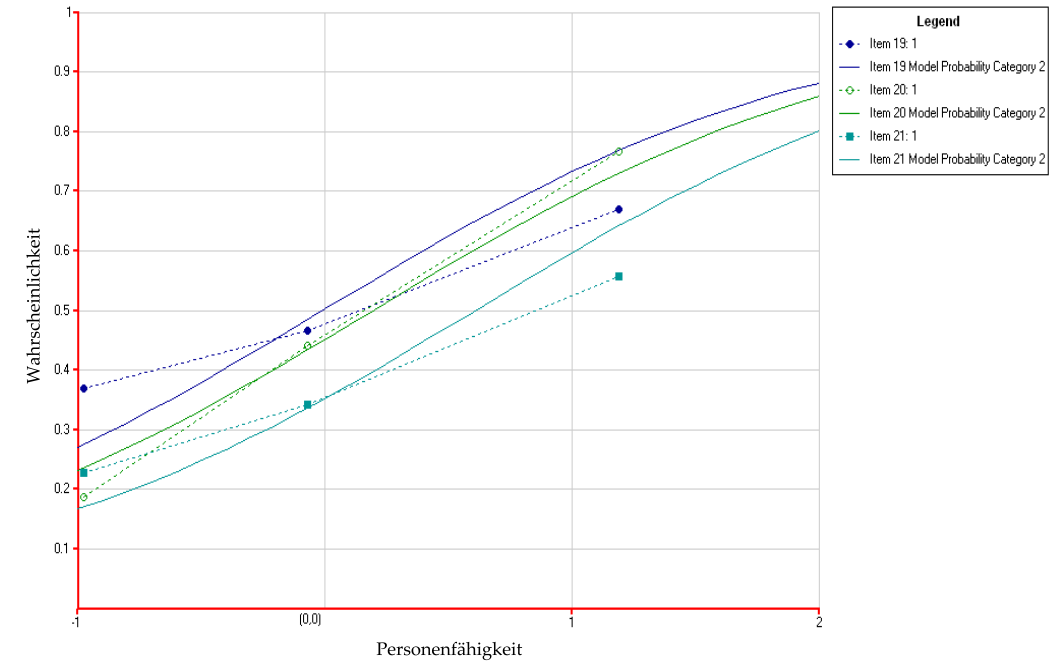


Abbildung 5.14: ICCs der Items 19 bis 21

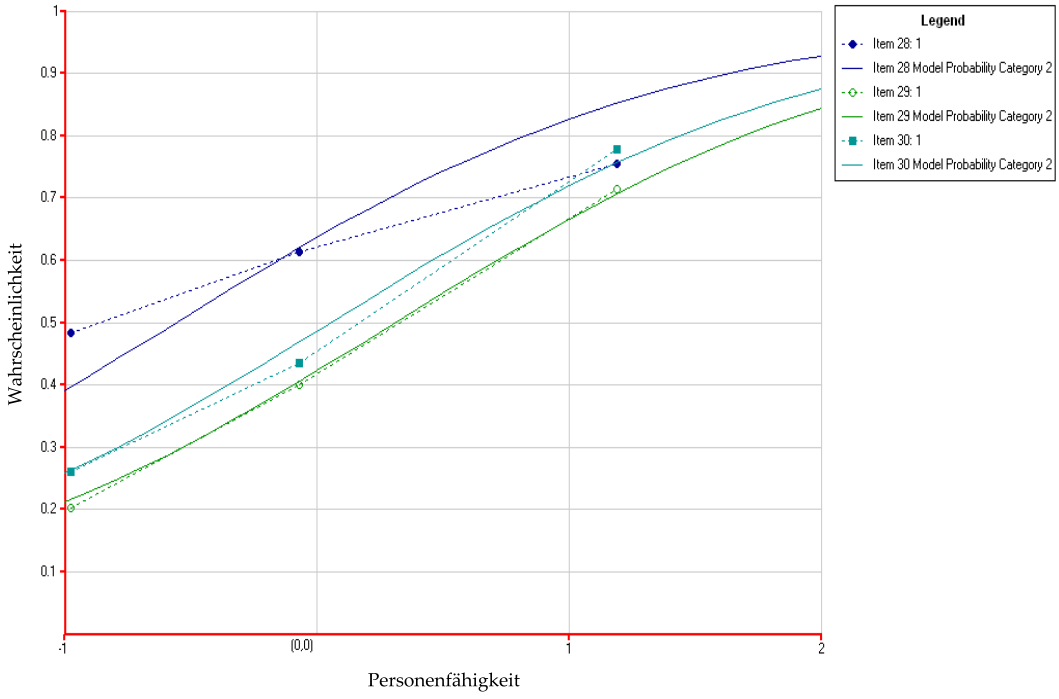


Abbildung 5.15: ICCs der Items 28 bis 30

Sol3) als auffällig.

Diese Ergebnisse sprechen dafür, dass die genannten Items die Modellvoraussetzung gleicher Trennschärfen verletzen und aus dem Test entfern werden sollten. Da eine Entfernung einzelner Items aufgrund ihrer Zusammenstellung zu Sets nicht möglich ist, folgen vor dem Hintergrund der getroffenen Einzelbewertungen die Beurteilungen der Sets, um eine abschließende Entscheidung über ihre Entfernung treffen zu können.

QUALITÄT DER ITEMSETS

Die neben der Prüfung der einzelnen Items wichtige Qualitätsprüfung der thematisch zusammenhängenden Itemsets umfasst eine gemeinsame Betrachtung ihrer probabilistischen Itemschwierigkeiten und klassischen Trennschärfen sowie ihrer T-Werte. Die Betrachtung geschieht vor dem Hintergrund, dass die Feststellung unzureichender Itemkennwerte nicht in den Ausschluss einzelner Items münden kann, sondern ganze Itemsets gestrichen werden müssen. Ansonsten wären die grundsätzlich aus drei Items bestehenden Sets nicht mehr gleichmäßig besetzt. Gleiches gilt für die drei Fähigkeiten, von denen aus auf die *prozessbezogene naturwissenschaftliche Grundbildung* geschlossen wird. Im Folgenden wird jedes Itemset einzeln betrachtet und hinsichtlich der genannten Kriterien beschrieben.

- *Brot (Bro1, Bro2, Bro3)*: Das Itemset zeigt gut verteilte Itemschwierigkeiten, die im Durchschnitt kaum von der durchschnittlichen Schwierigkeit des Gesamttests abweichen. Die Trennschärfen liegen mit einem Durchschnitt von $0,50$ über dem Testdurchschnitt. Die negativen T-Werte sprechen für im Vergleich zum Gesamttest überdurchschnittliche Trennschärfen und für eine tolerierbare Abweichung der Items.

- *Hurrikans (Hur1, Hur2, Hur3)*: Die drei Items weisen gut verteilte Schwierigkeiten auf, die im Durchschnitt ebenfalls kaum von der durchschnittlichen Schwierigkeit des Gesamttests abweichen. Die Trennschärfen liegen durchschnittlich bei $0,39$, wobei lediglich Item *Hur1* einen Wert $< 0,40$ zeigt. Die Trennschärfen liegen unter denen des vorangegangenen Sets, die T-Werte zeigen keine Auffälligkeiten.

- *Schimmel (Sch1, Sch2, Sch3)*: Dieses Set fällt mit einer durchschnittlichen Schwierigkeit von $-0,47$ sehr leicht aus. Insbesondere Item *Sch3* ist sehr leicht. Die durchschnittliche Trennschärfe kann mit einem Wert von $0,43$ als gut bezeichnet werden, wobei *Sch3* eine gerade noch ausreichende Trennschärfe aufweist. Die T-Werte sprechen für modellkonforme Items.

- *Stichling (Sti1, Sti2, Sti3)*: Das Itemset liegt mit einer durchschnittlichen Itemschwierigkeit von $-0,01$ nahe an der durchschnittlichen Schwierigkeit des Tests. Die Trennschärfe liegt im Durchschnitt bei $0,43$, wobei Item *Sti1* mit einem Wert von $0,31$ stark nach unten abweicht. Die anderen beiden Items zeigen gute Trennschärfen, wobei der T-Wert des Items *Sti1* in Höhe von $2,7$ auffällig ist und auf eine möglicherweise abweichende Trennschärfe hindeutet. *Sti3* erweist sich dagegen mit einem T-Wert von $-3,8$ als in Richtung einer deutlich höheren Trennschärfe abweichend. Das Itemset ist demnach hinsichtlich des Infits durch starke Schwankungen gekennzeichnet.

- *Kleine Lebewesen, kleine Teilchen (KLe1, KLe2, KLe3)*: Das Itemset erweist sich mit einer durchschnittlichen Itemschwierigkeit von $1,21$ als das bei weitem schwierigste, wofür insbesondere das Item *KLe2* mit einem Wert von $2,95$ verantwortlich ist. Weiterhin zeigt das Set mit einem Durchschnitt von $0,26$ im Vergleich zu allen anderen Sets die mit Abstand niedrigste Trennschärfe. Bei der Prüfung der Trennschärfenhomogenität fällt insbesondere *KLe3* durch einen hohen T-Wert von $3,8$ auf. Wie die Betrachtungen der ICCs gezeigt haben, liegt hier eine deutliche Abweichung in Richtung einer unterdurchschnittlichen Trennschärfe vor.

- *Klima (Kli1, Kli2, Kli3)*: Dieses Itemset gehört mit einer durchschnittlichen Schwierigkeit von $0,26$ bereits zu den schwierigeren Sets. Die Trennschärfen fallen mit einem Durchschnitt von $0,40$ gerade ausreichend aus und die T-Werte zeigen keine augenfälligen Abweichungen.

- *Regenbogenforelle (Reg1, Reg2, Reg3)*: Ebenso wie das Klima-Set gehört auch dieses mit einem Durchschnittswert von $0,27$ zu den schwierigeren Sets. Auch dieses Itemset weist mit einem Durchschnitt von $0,39$ ausreichende Trennschärfen auf, wobei die Werte der Items *Reg1* und *Reg3* sehr gering ausfallen. Die T-Werte erweisen sich als sehr konträr. Während *Reg1* mit einem Wert von $2,3$ und *Reg3* mit einem Wert von $2,4$ auf eine zu geringe Trennschärfe hindeuten, zeigt *Reg2* mit $-3,5$ eine deutliche Abweichung in Richtung einer im Vergleich zum Testdurchschnitt hohen Trennschärfe.

- *Rost (Ros1, Ros2, Ros3)*: Das Itemset liegt mit einer durchschnittlichen Schwierigkeit von $-0,19$ unter dem Testdurchschnitt. Es besitzt mit einem Wert von $0,52$ die durchschnittlich beste Trennschärfe. Alle T-Werte - insbesondere der Wert des Items *Ros2* - deuten auf überdurchschnittlich gute Trennschärfen hin.

- *Skorbut (Sko1, Sko2, Sko3)*: Das Itemset erweist sich mit einer durchschnittlichen

Schwierigkeit von $-0,63$ als das leichteste. Die durchschnittliche Trennschärfe liegt mit $0,45$ leicht über dem Testdurchschnitt. Die T-Werte erscheinen unauffällig.

- *Solarzelle (Sol1, Sol2, Sol3)*: Das Itemset ist mit einem Durchschnittswert von $-0,07$ unterdurchschnittlich leicht. Die mittlere Trennschärfe liegt bei $0,41$, wobei Item *Sol1* mit $0,28$ eine nicht ausreichende Trennschärfe zeigt. Der T-Wert des Items *Sol1* in Höhe von $3,4$ deutet auf ein in Richtung zu geringer Trennschärfe abweichendes Item hin.

5.4.3 BEWERTUNG DER ITEMSETS

Aus den im vorangegangenen Abschnitt beschriebenen Betrachtungen der Itemkennwerte werden an dieser Stelle Konsequenzen gezogen, die sich auf die Eliminierung von Itemsets beziehen, welche sich in der kombinierten Betrachtung der Itemkennwerte und der T-Werte als unzureichend bzw. nicht modellkonform erwiesen haben. Über die Grundlage dieser Entscheidung gibt Tabelle 5.19 Auskunft, welche die Bewertungen noch einmal in Kürze zusammenfasst. Die Tabelle zeigt die mittle-

Tabelle 5.19: Bewertung der Itemkennwerte

Items	mittlere Schwierigkeit	mittlere Trennschärfe		T-Wert
Brot	0,03	0,50	+ +	+
Hurrikans	0,08	0,39	- +	+ +
Schimmel	-0,47	0,43	+	+ +
Stichlinge	**-0,01**	**0,43**	**+**	**-**
Kleine Lebew.	**1,21**	**0,26**	**- -**	**- -**
Klima	0,26	0,40	- +	+ +
Regenbogenf.	**0,27**	**0,39**	**- +**	**-**
Rost	-0,19	0,52	+ +	+
Skorbut	-0,63	0,45	+	+
Solarzelle	-0,07	0,41	+	- +

ren Schwierigkeiten und Trennschärfen der Itemsets. Die Trennschärfen und T-Werte wurden deskriptiv mit „+“ und „-“ und Kombinationen der beiden bewertet.

Bei der Bewertung der durchschnittlichen Trennschärfe wurde das Kriterium angelegt, dass Trennschärfen zwischen $0,4$ und $0,7$ liegen sollten (Kelava & Moosbrugger, 2007), um als gut bezeichnet werden zu können. Ein „+ +“ bekamen die Itemsets mit einer durchschnittlichen Trennschärfe von $\geq 0,50$ und ein „+“ bekamen die Itemsets mit einer durchschnittlichen Trennschärfe zwischen $0,40$ und $0,49$. Itemsets mit durchschnittlichen Trennschärfen zwischen $0,30$ und $0,39$ wurden mit einem „-“ bewertet, da sie unter der Mindestmarke von $0,4$ liegen. Da es drei Durchschnittswerte gibt, die an der Grenze zwischen „+“ und „-“ liegen, wurden diese mit einem „- +“ bewertet. Itemsets mit einer mittleren Trennschärfe $\leq 0,30$ wurden mit einem „- -“ bewertet.

Bezüglich der T-Werte wurde folgende Bewertung vorgenommen. Bestand das Itemset aus drei unauffälligen Items, so wurde es mit „+ +“ bewertet. Itemsets, deren T-Werte auf eine Abweichung in Richtung überdurchschnittlich hoher Trennschärfen hindeuten, erhielten ein „+“. Mit einem „- +“ bewertete Itemsets zeigen unauffällige T-Werte und lediglich einen Wert, der auf eine zu niedrigen Trennschärfe hindeutet. Itemsets wurden mit einem „-“ bewertet, wenn die T-Werte konträr ausfielen, also einen positiv sowie einen negativ abweichenden T-Wert enthielten. Mit „- -“ wurden schließlich Itemsets mit positiven T-Werten eingestuft, von denen mindestens einer positiv (Wert $\geq 2,0$) abweicht und damit auf eine zu geringe Trennschärfe hindeutet.

Um eine Entscheidung über die Entfernung von Itemsets zu treffen, wird hier das Kriterium angelegt, dass ein Set dann zu entfernen ist, wenn einer der Bereiche mit einem „-“ bewertet wurde, ohne dass diesem ein ausgleichendes „+“ entgegensteht. Bei kombinierter Betrachtung der dargestellten Kennwerte und Bewertungen haben sich gemäß diesem Kriterium drei Itemsets als nicht ausreichend erwiesen: die Sets *Stichlinge*, *Kleine Lebewesen/kleine Teilchen* und *Regenbogenforelle*. Bei zwei dieser drei Sets fiel die Entscheidung aufgrund der Daten leicht. Bei dem Itemset *Stichlinge* fiel die mittlere Trennschärfe zwar ausreichend aus, jedoch erwiesen sich die T-Werte als so heterogen, dass das Set schließlich doch gestrichen wurde.

5.4.4 ENDGÜLTIGE TESTZUSAMMENSTELLUNG

Aus den dargestellte Bewertungen ergibt sich eine abschließende Testzusammenstellung aus den Sets:

- Brot backen
- Hurrikans
- Schimmel

- Klima
- Rost
- Skorbut
- Solarzelle.

Tabelle 5.20: Deskriptive Daten der abschließenden Testversion

Item-Nr.	Item	Schwierigkeit	M_Set	Trennschärfe	M_Set
1	Bro 1	-0,01		0,47	
2	Bro 2	0,88		0,53	
3	Bro 3	-0,79	0,03	0,51	0,50
4	Hur 1	0,84		0,37	
5	Hur 2	-0,44		0,44	
6	Hur 3	-0,16	0,08	0,41	0,41
7	Sch 1	0,57		0,53	
8	Sch 2	-0,35		0,42	
9	Sch 3	-1,66	-0,48	0,37	0,44
10	Kli 1	0,20		0,44	
11	Kli 2	0,11		0,38	
12	Kli 3	0,51	0,27	0,38	0,40
13	Ros 1	-0,89		0,53	
14	Ros 2	0,29		0,57	
15	Ros 3	0,05	-0,18	0,49	0,53
16	Sko 1	-0,39		0,44	
17	Sko 2	-0,89		0,54	
18	Sko 3	-0,65	-0,64	0,43	0,47
19	Sol 1	-0,57		0,30	
20	Sol 2	0,32		0,49	
21	Sol 3	0,06	-0,06	0,49	0,43
	M	-0,14		0,45	

M = Mittelwert

Tabelle 5.20 gibt einen deskriptiven Überblick über die Items der abschließenden Testversion. Die Schwierigkeiten der Items haben sich durch die Reduzierung des Itempools nicht verändert, da ihre Berechnung nicht von den übrigen Items abhängt. Kleine Abweichungen im Vergleich zu den Aufgabenschwierigkeiten in Tabelle 5.18 können als Zufallsschwankungen bezeichnet werden. Ein anderes Bild ergibt sich bei der Betrachtung der Trennschärfen. Bei der Berechnung der klassischen Trennschärfen werden die einzelnen Items mit dem Gesamttest korreliert und daher verändern

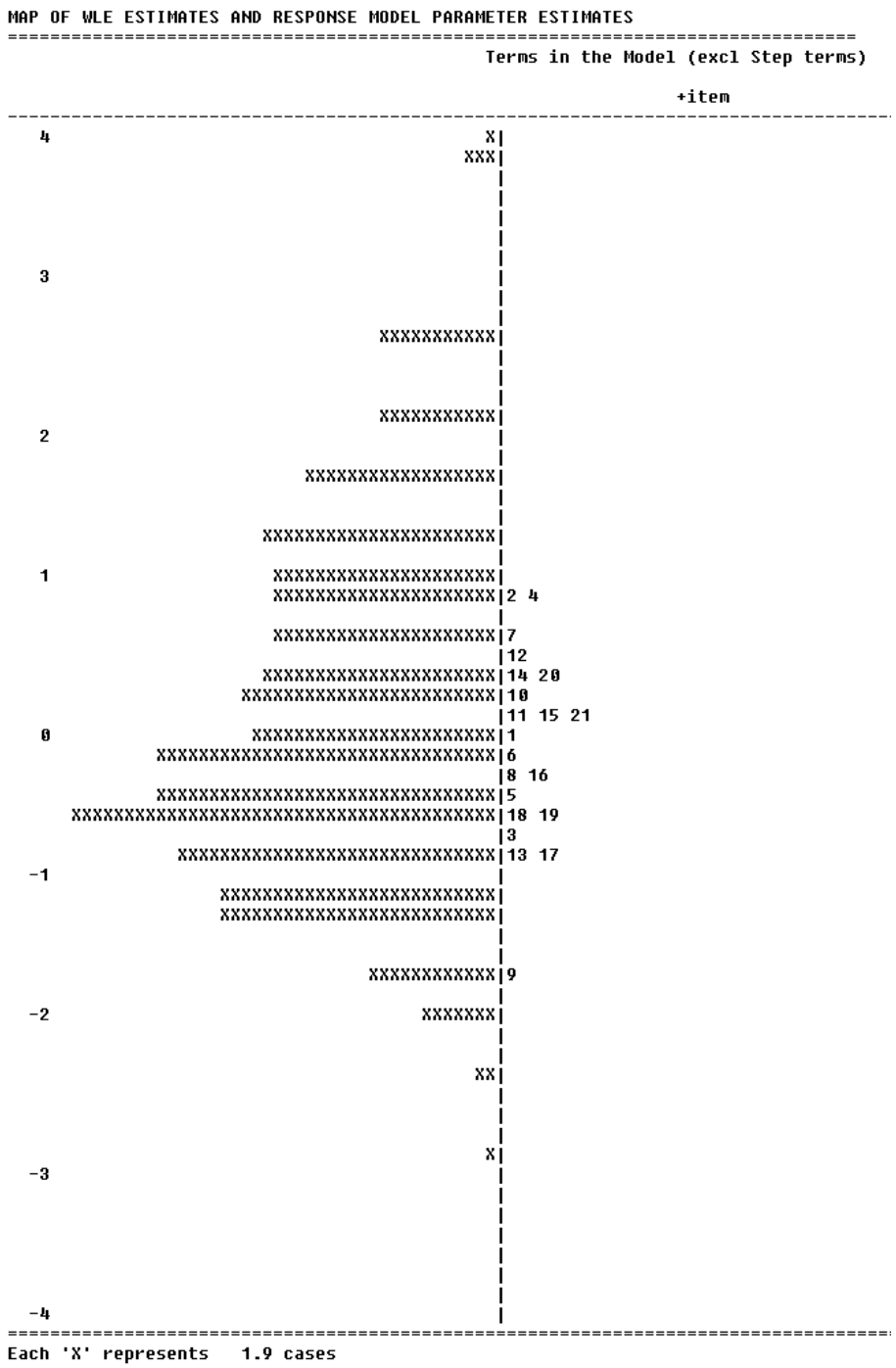


Abbildung 5.16: Verteilung der Itemschwierigkeiten über die Personenfähigkeiten

sich die Trennschärfen, wenn sich der Itempool verändert. Die Daten zeigen, dass sich die durchschnittliche Trennschärfe nach Entfernung der nicht modellkonformen Itemsets leicht von ursprünglich $0,42$ auf $0,45$ verbessert.

Die durchschnittliche Aufgabenschwierigkeit von $-0,14$ des verbleibenden Itempools deutet vor dem Hintergrund einer auf 0 fixierten mittleren Personenfähigkeit darauf hin, dass der Test für die Stichprobe nun ein wenig zu leicht, im Großen und Ganzen jedoch angemessen ausfällt. Die in Abbildung 5.16 dargestellte Wright-Map der abschließenden Testversion stellt die Verteilung der Personenfähigkeiten noch einmal der Verteilung der Aufgabenschwierigkeiten gegenüber. Es ist zu erkennen, dass nach der Entfernung der qualitativ unzureichenden Itemsets zwar für den mittleren Fähigkeitsbereich ausreichend Items zur Verfügung stehen und auch in Richtung des unteren Drittels noch Items vorhanden sind, dass es aber für die Differenzierung der Personen im oberen Fähigkeitsdrittel an Items fehlt.

KOMPETENZWERTE

Um die im Anschluss an diesen Abschnitt folgende Validierung vorzubereiten und einen Überblick über die Kompetenzdaten zu bieten sowie die ermittelten Daten besser einschätzen zu können, werden auf der Datengrundlage der genannten verbliebenen Itemsets die Gesamtstichprobe und die nach Niveau und Geschlecht gruppierten Teilstichproben hinsichtlich ihrer Kompetenzausprägung deskriptiv dargestellt. Tabelle 5.21 zeigt Mittelwerte (M) und Standardfehler der Mittelwerte (SE) des *Weighted Likelihood Schätzers (WLE)*, der als Personenparameterschätzer herangezogen wurde (vgl. Abschnitt 3.1.2). Zur Beurteilung der angegebenen WLE-Schätzungen ist es

Tabelle 5.21: Kompetenzwerte der nach Geschlecht bzw. Schulniveau aufgeteilten Stichproben

Kompetenz		**n**	**M**	**SE**
Gesamt		750	0,157	0,041
Geschlecht	m	373	0,143	0,056
	w	375	0,172	0,060
Schulniveau	Hauptschule	250	-0,614	0,042
	Realschule	250	0,041	0,058
	Gymnasium	250	1,045	0,066

n = Stichprobengröße M = Mittelwert SE = Standardfehler

wichtig zu berücksichtigen, dass die mittlere Itemschwierigkeit auf Null fixiert wur-

de[3]. Dies bedeutet für die ebenfalls auf der Logit-Skala angegebenen Personenparameterschätzungen, dass ihre WLE-Mittelwerte mit dem Mittelwert der Itemschwierigkeit von Null in Beziehung gesetzt werden. Tabelle 5.21 gibt an, dass die *prozessbezogene naturwissenschaftliche Grundbildung* der Gesamtstichprobe (im Folgenden nur als *Kompetenz* bezeichnet) im Durchschnitt bei $0,157$ Logits liegt. Dies bedeutet, dass die auf die Gesamtstichprobe bezogene durchschnittliche Personenfähigkeit nur knapp über der durchschnittlichen Itemschwierigkeit liegt und der Test somit eine gute Passung zwischen Itemschwierigkeit und Personenfähigkeit aufweist.

Betrachtet man die nach Geschlecht unterteilte Gesamtstichprobe, so zeigt sich, dass der WLE-Mittelwert der Schülerinnen $(0,172)$ über dem der Schüler liegt $(0,143)$. Die Schülerinnen erweisen sich damit als etwas kompetenter als die Schüler.

Die nach Schulniveau geteilte Stichprobe weist deutliche Unterschiede zwischen den Teilstichproben aus. Die mittlere Kompetenz der Hauptschülerinnen und Hauptschüler liegt bei $-0,614$, der Test ist folglich für diese Stichprobe zu schwer. Anders sieht es bei den Realschülerinnen und Realschülern aus. Ihre Kompetenz fällt mit $0,041$ Logits im Vergleich zur durchschnittlichen Itemschwierigkeit optimal aus. Die Gymnasiastinnen und Gymnasiasten dagegen liegen mit ihrer Kompetenz $(1,045)$ deutlich über der durchschnittlichen Itemschwierigkeit. Der Test ist für sie zu leicht.

Im Abschnitt 5.5 werden die Gruppenunterschiede noch einmal aufgegriffen, um sie vor dem Hintergrund der Validitätsprüfung auf Signifikanz und auf *Differential Item Functioning (DIF)* zu prüfen.

5.5 PRÜFUNG DER GÜTEKRITERIEN

Den Abschluss des Ergebnisteils bildet die Prüfung der Gütekriterien. Wie bereits im Rahmen des Methodenteils 3.2 ausgeführt, werden die interne und externe Validität des Tests betrachtet. Die Reliabilität wird anhand der WLE-Reliabilität sowie anhand der internen Konsistenz (Cronbachs Alpha) geprüft.

Wichtig für die Einordnung der folgenden Auswertungen ist die Tatsache, dass es sich hier um erste Ansätze zur Prüfung der Gütekriterien handelt. Da diese Prüfung anhand der gleichen Daten durchgeführt wird, die bereits der Beurteilung der Items dienten und sich nun auf den anhand dieser Daten reduzierten Itempool gründet, sind die Ergebnisse mit Vorsicht zu betrachten und entbinden nicht von der Prü-

3 Der im eindimensionalen Rasch-Modell zu schätzende Itemparameter ist die Aufgabenschwierigkeit, die auf Basis der Itemantworten geschätzt wird und auf der Logit-Skala angegeben wird. Zu diesem Zweck muss die Logit-Skala normiert werden, was durch eine Fixierung der mittleren Aufgabenschwierigkeit auf Null geschieht.

fung der Kriterien anhand einer neuen Stichprobe. Aufgabenanalyse und Validierung sollten nicht an der gleichen Stichprobe durchgeführt werden, da selbst im Fall einer hoch repräsentativen Analysestichprobe systematische, dem Beobachter nicht zugängliche, Fehler nicht verhindert werden können. Diese können nur durch eine *Kreuzvalidierung* identifiziert werden, also durch eine Datenerhebung anhand einer neuen Stichprobe und einen Vergleich mit dieser (vgl.Lienert & Raatz, 1998).

Die im folgenden dargestellten Ergebnisse können somit lediglich erste Hinweise auf Validität und Reliabilität des Tests geben und müssen als eben solche eingestuft werden.

5.5.1 PRÜFUNG DER VALIDITÄT

Die Prüfung der Validität teilt sich in die interne Validierung, also die Prüfung des angenommenen Testmodells, und in die externe Validierung, in deren Verlauf die korrelativen Beziehungen zwischen Kompetenz und externen Leistungs- und Motivationskriterien geprüft werden.

INTERNE VALIDIERUNG

Die Prüfung der internen Validität teilt sich in folgende Bestandteile:

1. Es wird die für das einfache Rasch-Modell wichtige Voraussetzung gleicher Trennschärfen anhand der Infit-Werte, also die Itemhomogenität geprüft (vgl. Abschnitt 3.1.2).

2. Es werden Gruppenvergleiche zwischen Schülerinnen und Schülern sowie zwischen den einzelnen Schulniveaus angestellt. Bezüglich des Geschlechtervergleichs wird in Anlehnung an die nationalen Ergebnisse zur naturwissenschaftlichen Kompetenz der PISA-Studie 2006 für Schleswig-Holstein (Rönnebeck, Schöps, Prenzel & Hammann, 2008) angenommen, dass sich Mädchen und Jungen hinsichtlich ihrer *prozessbezogenen naturwissenschaftlichen Grundbildung* nicht signifikant unterscheiden.

   Bezüglich der Schulniveaus wird angenommen, dass sie sich signifikant hinsichtlich der Ausprägung der *prozessbezogenen naturwissenschaftlichen Grundbildung* ihrer Schülerinnen und Schüler unterscheiden. Aufgrund der mit ansteigendem Schulniveau differenzierteren naturwissenschaftlichen Bildung und in Anlehnung an die nationalen PISA-Ergebnisse 2006 für Schleswig-Holstein (Rönnebeck et al.,

2008) wird angenommen, dass Gymnasiastinnen und Gymnasiasten dementsprechend im Test besser abschneiden als Realschülerinnen und Realschüler und diese wiederum besser als Hauptschülerinnen und Hauptschüler.

3. Es wird eine Prüfung der Personenhomogenität der Items durch die getrennte Schätzung der Itemparameter für nach Geschlecht bzw. Schulniveau aufgeteilte Untergruppen der Gesamtstichprobe vorgenommen. Hier wird die Annahme geprüft, dass der Test bzw. die einzelnen Items bei allen Personen dieselbe Personenfähigkeit messen. Dies sollte sich darin ausdrücken, dass die Itemschwierigkeiten in bestimmten Untergruppen, in denen dieselbe Kompetenz gemessen wird, gleich ausfallen. Die Personenhomogenität wird überprüft, indem die Itemschwierigkeiten für Untergruppen getrennt geschätzt werden. Bestätigt sich die getroffene Annahme, so sollten die Schätzungen für die Untergruppen, von Zufallsschwankungen abgesehen, zu ähnlichen Ergebnissen führen, die Items sollten also kein *Differential Item Functioning (DIF)* zeigen (vgl. Abschnitt 3.2.1).

4. Zum Abschluss erfolgt noch einmal eine globale Modellprüfung anhand der informationstheoretischen Modellselektionsmaße CAIC und BIC. Gemäß der theoretischen Grundlagen der Arbeit wird angenommen, dass der entwickelte Test die *prozessbezogene naturwissenschaftliche Grundbildung* als eine Dimension misst. Da zur Itementwicklung drei Fähigkeiten ausgewählt wurden, um als Indikatoren dieser Kompetenz zu fungieren, wird zur globalen Modellprüfung hier das eindimensionale Modell, das eine einzige Dimension annimmt, dem dreidimensionalen Modell gegenübergestellt. Diese Prüfung wird auch deshalb noch einmal anhand des verbleibenden Itempools durchgeführt, da die Prüfungen im Rahmen des Feld- und Haupttests nicht eindeutig ausfielen.

AD 1: PRÜFUNG DES INFITS

Tabelle 5.22 gibt Auskunft über die T-Werte der abschließenden Testversion. Es wird deutlich, dass die auf diese Weise zusammengestellte Testversion lediglich noch ein Item aufweist, das durch einen extrem abweichenden T-Wert auffällt. Das bereits vor der Reduzierung des Itempools auffällige Item *Sol1* legt, gemäß dem in Abschnitt 3.2.1 festgelegten Kriterium, mit einem T-Wert von $3,5$ eine Verletzung des Modells in Richtung einer unterdurchschnittlichen Trennschärfe nahe. Die übrigen T-Werte lassen keine weiteren Auffälligkeiten erkennen. Die mittlere Trennschärfe der abschließenden Testversion liegt bei $0,45$ und hat sich damit nach der Itementfernung von

Tabelle 5.22: Trennschärfen der abschließenden Testversion

Item-Nr.	Item	Trennschärfe	MNSQ	T-Wert
1	Bro 1	0,47	1,02	0,6
2	Bro 2	0,53	0,93	-1,8
3	Bro 3	0,51	0,94	-1,9
4	Hur 1	0,37	1,07	1,8
5	Hur 2	0,44	1,03	0,9
6	Hur 3	0,41	1,06	2,1
7	Sch 1	0,53	0,95	-1,6
8	Sch 2	0,42	1,03	0,9
9	Sch 3	0,37	1,01	0,1
10	Kli 1	0,44	1,03	1,0
11	Kli 2	0,38	1,07	2,2
12	Kli 3	0,38	1,07	1,9
13	Ros 1	0,53	0,92	-2,4
14	Ros 2	0,57	0,93	-2,4
15	Ros 3	0,49	1,01	0,2
16	Sko 1	0,44	0,99	-0,3
17	Sko 2	0,54	0,92	-2,3
18	Sko 3	0,43	1,00	0,1
19	Sol 1	**0,30**	1,11	**3,5**
20	Sol 2	0,49	0,99	-0,3
21	Sol 3	0,49	0,96	-1,4

MNSQ = Infit Mean Square T-Wert = Prüfgröße des Infit-Maßes

ursprünglich $0,42$ noch ein wenig weiter verbessert.

Die Infit Mean Squares der abschließenden Testversion liegen bei allen Items innerhalb der Grenzen von $0,75$ und $1,33$. Im Testdurchschnitt liegt der Mean Square bei 1.

Um zu prüfen, ob der auffällige T-Wert des Items *Sol1* (Item 19) wirklich durch eine abweichende Trennschärfe begründet ist, wird anhand des in Tabelle 5.17 dargestellten beispielhaften ICC-Ausschnitts der Items 19 bis 21 noch eine visuelle Prüfung der Items vorgenommen. Einen Gesamtüberblick über alle ICCs der abschließenden Testversion bieten die Abbildungen D.6 bis D.8 ab S. 260.

Die Abbildung stellt den modellgemäßen Verlauf der einzelnen Items des Sets Solarzelle (Model Probability) dem empirisch ermittelten Verlauf gegenüber. Es wird deutlich, dass Item 19 nicht modellkonform verläuft und sogar die ICC des Items 21 kreuzt. Die Voraussetzung gleicher Trennschärfen des einfachen Rasch-Modells, also der parallele Verlauf der ICCs, wird demnach verletzt. Da es sich jedoch nur noch um

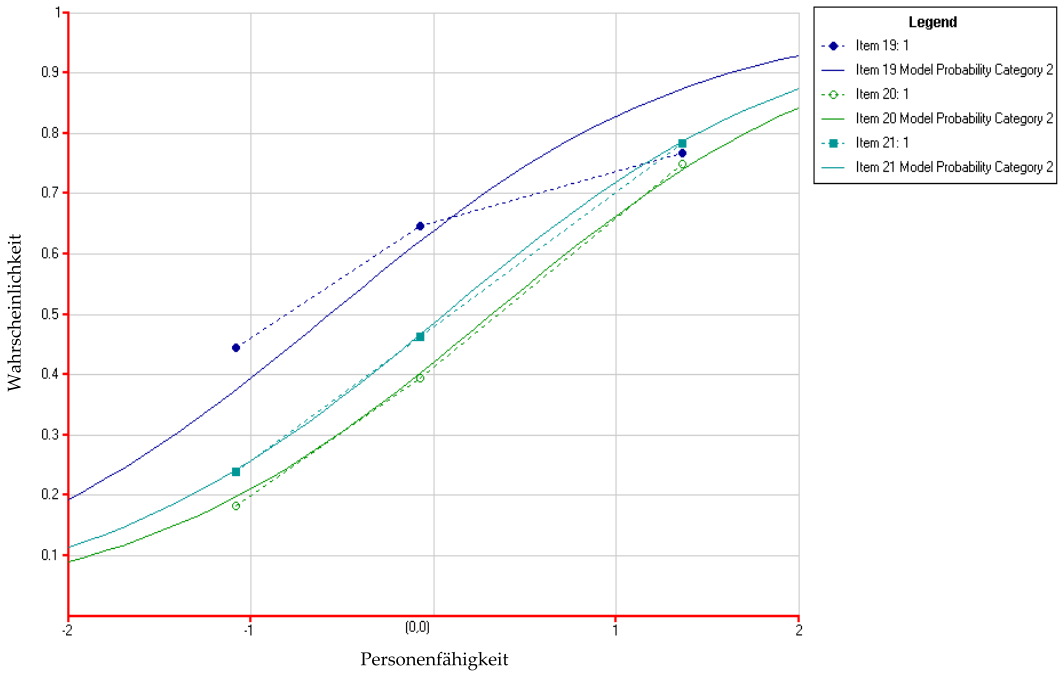


Abbildung 5.17: ICCs der Items 19 bis 21 (*Sol1-3*)

ein einziges Item handelt, das eine solche Abweichung zeigt und die übrigen Items des Sets sehr gute Trennschärfen sowie einen sehr guten Infit aufweisen, wird diese Abweichung in diesem Fall toleriert.

AD 2: PRÜFUNG VON GRUPPENUNTERSCHIEDEN

Zunächst werden hier die Unterschiede zwischen den Geschlechterstichproben geprüft. Zu diesem Zweck wurde bei gegebener Normalverteilung und vorliegender Varianzhomogenität ein einseitiger t-Test durchgeführt, der mit einem t-Wert von $-0,352$ ($p = 0,725$) allerdings keine signifikanten Gruppenunterschiede ergab. Dieses Ergebnis bestätigt die Hypothese, dass sich Mädchen und Jungen hinsichtlich ihrer *prozessbezogenen naturwissenschaftlichen Grundbildung* nicht signifikant unterscheiden. Tabelle 5.23 gibt Auskunft über die verglichenen deskriptiven Werte.

Anders fällt der Vergleich zwischen den Schulniveaus aus. Zum Zweck des Gruppenvergleichs wurde hier eine Varianzanalyse (ANOVA) mit Prüfung eines a-priori-Kontrasts durchgeführt. Durch den Kontrast wurde die Annahme geprüft, dass die Schülerinnen und Schüler des Gymnasiums eine signifikant höhere Ausprägung der gemessenen Kompetenz besitzen als die Realschülerinnen und Realschüler und diese wiederum eine höhere Ausprägung besitzen als die Hauptschülerinnen und Haupt-

Tabelle 5.23: Untersuchung von Geschlechterunterschieden hinsichtlich der gemessenen Kompetenz

		n	M	SD
Kompetenz	m	373	0,143	1,090
	w	375	0,172	1,160

n = Stichprobengröße M = Mittelwert SD = Standardabweichung

schüler. Wichtig zu beachten ist an dieser Stelle, dass die zur Durchführung einer Varianzanalyse notwendige Voraussetzung homogener Varianzen hier nicht gegeben ist. Gemäß Bortz (1999) ist die Varianzanalyse gegenüber dieser Verletzung robust, wenn die Stichproben ausreichend groß und vor allem *gleich groß* sind. Der F-Test wird in diesem Fall nur unerheblich beeinflusst. Da diese Bedingungen im Fall der Niveaustichproben erfüllt sind, fand die Varianzanalyse Anwendung.

Sie ergibt zunächst bei der allgemeinen Testung auf Gruppenunterschiede einen F-Wert von $218,52$ $(p < 0,001)$, also ein hoch signifikantes Ergebnis. Die Prüfung der einzelnen Gruppenunterschiede in Form des beschriebenen Kontrasts resultiert in einem t-Wert von $21,11$ $(p < 0,001)$. Die Hypothese, der angenommenen Gruppenunterschiede kann also aufgrund der Datenlage bestätigt werden. Tabelle 5.24 gibt Auskunft über die deskriptiven Werte der Stichprobe.

Tabelle 5.24: Untersuchung von Schulniveau-Unterschieden hinsichtlich der gemessenen Kompetenz

		n	M	SD
Kompetenz	Hauptschule	250	-0,614	0,666
	Realschule	250	0,041	0,923
	Gymnasium	250	1,045	1,048

n = Stichprobengröße M = Mittelwert SD = Standardabweichung

AD 3: PRÜFUNG DER PERSONENHOMOGENITÄT: DIF-ANALYSEN

Die erste Analyse des *Differential Item Functioning (DIF)* erfolgt anhand der nach Geschlecht gebildeten Teilstichproben. Im Anschluss daran werden die nach Schulniveau zusammengestellten Unterstichproben bezüglich möglicher *DIF*-Anzeichen untersucht.

Der Überprüfung des *DIF* dient gemäß Abschnitt 3.2.1 zunächst der *graphische Modelltest* (Rost, 2004a) der pro Untergruppe geschätzten Itemschwierigkeiten. Im Anschluss werden die Schätzwert-Intervalle der Itemschwierigkeiten betrachtet, bevor die Logit-Differenzen abschließend anhand der folgenden Kriterien von Wilson (2005) bewertet werden:

- Werte $< 0,43$: vernachlässigbare Differenz
- Werte zwischen $0,43$ und $0,64$: mittelmäßige Differenz
- Werte $> 0,64$: große Differenz.

Abbildung 5.18 präsentiert die graphische Prüfung möglicher DIF-Effekte zwischen der Gruppe der Schülerinnen und der Gruppe der Schüler. Der Bereich zwischen den gestrichelten Linien wird durch die maximal zulässige Differenz von $0,43$ zu beiden Seiten der linearen Geraden aufgespannt.

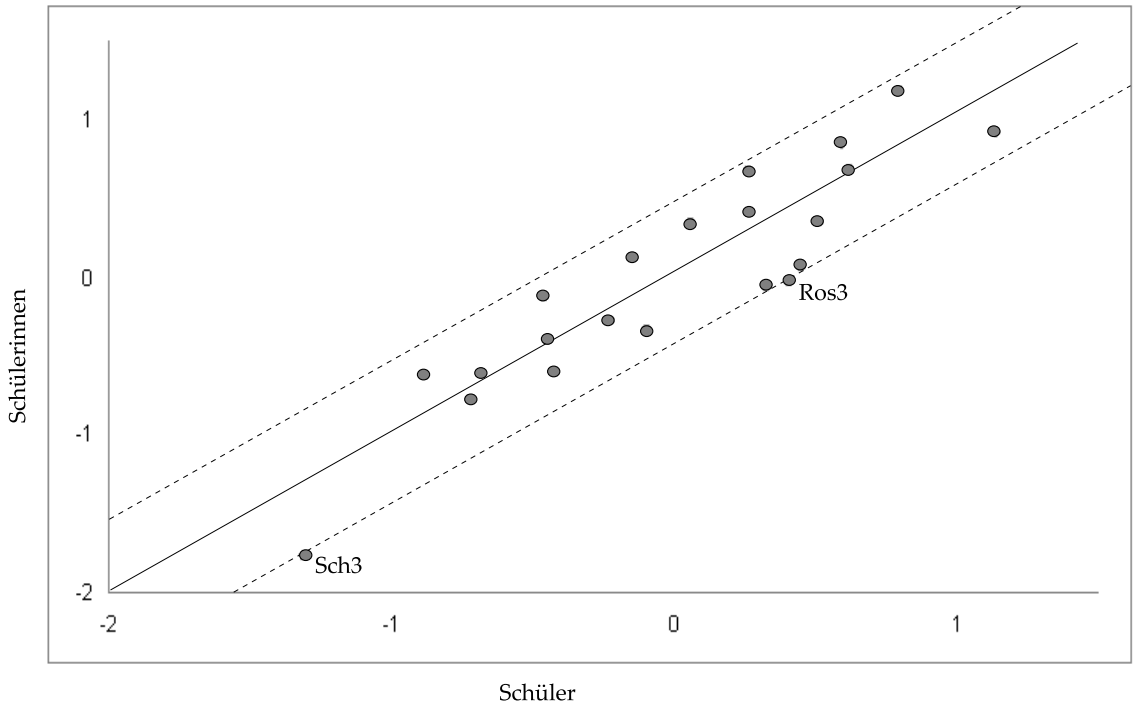


Abbildung 5.18: Prüfung von DIF-Effekten zwischen den Gruppen der Schülerinnen und Schüler

Es ist zu erkennen, dass die Items *Sch3* $(-1,31; -1,77)$ und *Ros3* $(0,41; -0,03)$ Werte aufweisen, die in den Grenzbereich dessen fallen, was als Abweichung bezeichnet werden kann. Gemäß Tabelle D.7 (s. Anhang, S. 263) zeigt sich, dass sich die Konfidenzintervalle der jeweiligen Wertepaare nicht überschneiden. Die Logit-Differenzen

sind im Falle des Items *Sch3* $(0,46)$ und *Ros3* $(0,44)$ als knapp in den mittelmäßigen Bereich fallend zu klassifizieren. Die Abweichungen der genannten Items fallen insgesamt lediglich marginal aus, so dass hier nicht von DIF-Effekten gesprochen werden kann.

Ein etwas anderes Bild ergibt sich im Rahmen der Suche nach möglichen DIF-Effekten bei einem Vergleich der nach Schulniveau unterteilten Gruppen. Der in Abbildung 5.19 angestellte Vergleich der an Haupt- und Realschulstichprobe geschätzten Itemschwierigkeiten ergibt zwei Abweichungen: Item *Sko2* $(-0,39;-0,96)$ und Item Kli3 $(0,31;0,88)$. Die Konfidenzintervalle der jeweiligen Itemschwierigkeitspaare überschneiden sich nicht. Die in Tabelle D.8 (s. Anhang, S.264) ausgewiesene Logit-Differenz des Items *Sko2* beträgt $0,58$ und die des Items *Kli3* liegt bei $0,58$. Damit fallen beide Werte in den Bereich der mittelmäßigen Differenz.

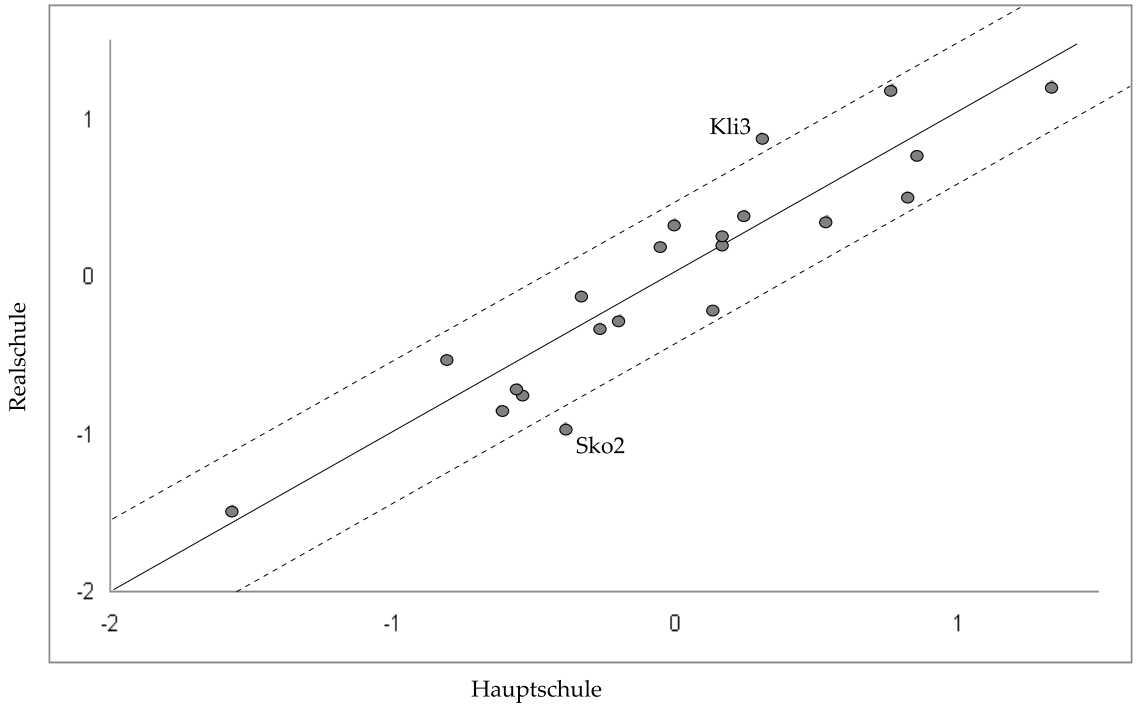


Abbildung 5.19: Graphische Darstellung der DIF-Prüfung (Hauptschule und Realschule)

Der Vergleich der geschätzten Itemschwierigkeiten von Haupt- und Gymnasialstichprobe ergibt gemäß Abbildung 5.20 im Falle von sechs Items mittelmäßige bis starke Abweichungen. In Tabelle D.9 (s. Anhang, S. 265) werden die Items *Bro2*, *Kli2*, *Ros2*, *Sko2* und *Sol1* als abweichend ausgewiesen. Die Konfidenzintervalle der Items überschneiden sich nicht. Die Logit-Differenz des Items *Bro2* $(0,63)$ ist als mittelmäßig und die Differenzen der Items *Kli2* $(0,70)$, *Ros2* $(0,85)$, *Sko2* $(0,77)$ und *Sol1* $(1,02)$

sind als groß zu bewerten sind.

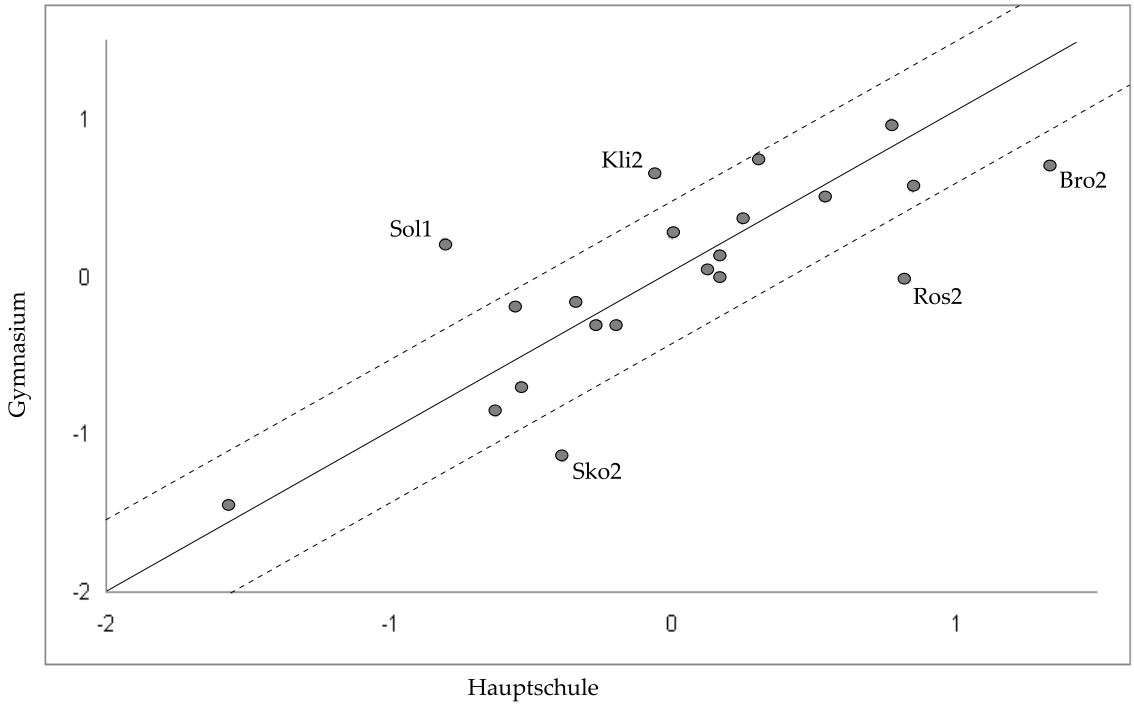


Abbildung 5.20: Graphische Darstellung der DIF-Prüfung (Hauptschule und Gymnasium)

Der Vergleich der Schwierigkeitsschätzung von Realschul- und Gymnasialstichprobe in Abbildung 5.21 weist fünf Items als abweichend aus. Keines der in Tabelle D.10 (s. Anhang, S. 266) dargestellten Konfidenzintervalle der Itemschwierigkeitspaare zeigt eine Überschneidung. Anders als im Vergleich von Haupt- und Gymnasialstichprobe ist allerdings nur eine der auffälligen Differenzen (*Sol1*) mit einem Wert von $0,75$ als groß zu bewerten. Die Logit-Differenz des Items *Kli2* $(0,46)$ fällt gerade in den als mittelmäßig zu bewertenden Bereich und die Items *Bro2* $(0,51)$, *Ros2* $(0,53)$ und *Sko3* $(0,52)$ liegen klar in diesem Bereich.

Anders als im Vergleich der Schwierigkeitsschätzungen von Schülerinnen und Schülern werden im Vergleich der Schulniveau-Teilstichproben im Falle folgender Items DIF-Effekte offenbar:

- *Bro2*: (Hauptschule-Gymnasium, Realschule-Gymnasium)
- *Kli2*: (Hauptschule-Gymnasium, Realschule-Gymnasium)
- *Kli3*: (Hauptschule-Realschule)
- *Ros2*: (Hauptschule-Gymnasium, Realschule-Gymnasium)

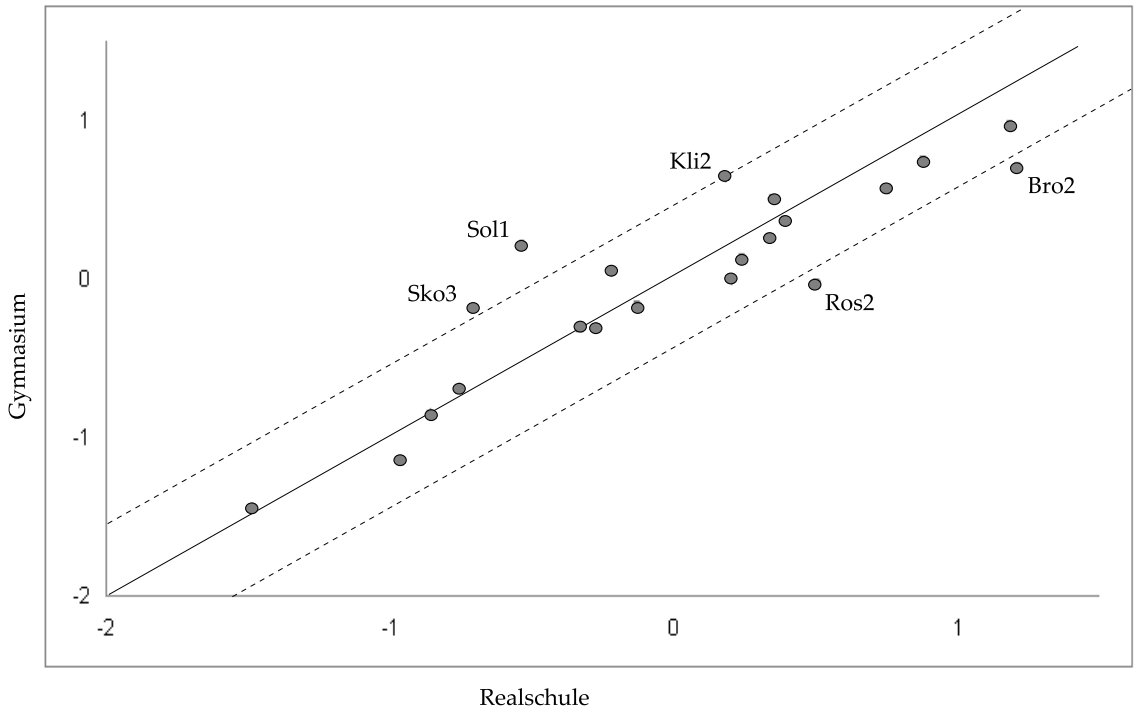


Abbildung 5.21: Graphische Darstellung der DIF-Prüfung (Realschule und Gymnasium)

- *Sko2*: (Hauptschule-Gymnasium, Hauptschule-Realschule)
- *Sko3*: (Realschule-Gymnasium)
- *Sol1*: (Hauptschule-Gymnasium, Realschule-Gymnasium)

In Klammern sind hier zur Übersicht noch einmal die Niveauvergleiche angegeben, innerhalb derer das Item *DIF* gezeigt hat. Mögliche Gründe für das *Differential Item Functioning* und die Konsequenzen dieser Ergebnisse im Hinblick auf den Umgang mit dem Testinstrument werden im Rahmen der Diskussion thematisiert.

AD 4: GLOBALE MODELLPRÜFUNG

Um die abschließende Testversion im Sinne einer internen Validierung hinsichtlich ihrer Dimensionalität zu prüfen, werden noch einmal die CAICs und BICs der eindimensionalen und der dreidimensionalen Testversion gegenübergestellt. Tabelle 5.25 gibt einen Überblick über die Werte. Im Rahmen der Modellprüfungen des Feld- und Haupttests konnte keine klare Entscheidung für eines der Modelle getroffen werden. Einen ähnlichen Eindruck vermitteln die Ergebnisse der abschließenden Testversion, wobei die Unterschiede zwischen beiden Modellen hier doch deutlich geringer

ausfallen als vor der Eliminierung der Itemsets. Die CAIC-Werte unterscheiden sich kaum und der Vergleich der BIC-Werte weist das dreidimensionale Modell knapp als das geeignetere aus. Weiteren Aufschluss über die Dimensionalität der Daten kön-

Tabelle 5.25: Globale Modellprüfung anhand des CAIC und BIC

Modell	unabh. Paramter	Deviance	CAIC	BIC
Eindimensional	22	19220,8	19388,4	19366,4
Dreidimensional	27	19182,6	19388,3	19361,3

Deviance = globale Fit-Statistik CAIC = Consistent Akaike´s Information Criterion
BIC = Bayes Information Criterion

nen die latenten Korrelationen der drei Dimensionen bieten, die in Tabelle 5.26 dargestellt sind. Die Korrelationen zwischen den einzelnen Dimensionen sind als sehr

Tabelle 5.26: Latente Korrelationen des dreidimensionalen Modells

	Dimension 1	Dimension 2
Dimension 1		
Dimension 2	0,92	
Dimension 3	0,96	0,93

hoch zu bewerten, so dass es auch hier angemessen erscheint, die drei Dimensionen zu einer zusammenzufassen und von einem eindimensionalen Modell auszugehen. In Kombination mit den geringen Unterschieden, die im Modellvergleich gefunden wurden und die im Fall des CAIC gemäß einer Klassifizierung von Burnham and Anderson (2002) als vernachlässigbar bezeichnet werden können, fällt der Vergleich der beiden Modelle insgesamt noch weniger klar aus als im Feld- und Haupttest. Das eindimensionale Modell kann hier ebenso Geltung beanspruchen wie das dreidimensionale Modell. Mögliche Gründe für die nicht eindeutige Dimensionalität der Testdaten werden in der Diskussion der Haupttest-Ergebnisse (s. Abschnitt 6.1.3, S. 226) betrachtet.

EXTERNE VALIDIERUNG

Im Rahmen der Prüfung der externen Validität werden die Ergebnisse des zu validierenden Tests in Beziehung zu externen Validitätskriterien gesetzt. Sie geschieht

in Form von Korrelationen der Testleistung mit verschiedenen Außenkriterien (Leistungs- und Motivationskriterien) und gliedert sich in die Prüfung der konvergenten und diskriminanten Validität.

PRÜFUNG DER KONVERGENTEN VALIDITÄT

Die Prüfung der konvergenten Validität erfolgt anhand von Korrelationen der Kompetenz (WLEs) mit den *Schulnoten*. Tabelle 5.27 zeigt, dass die einseitig berechneten Spearman-Korrelationen bis auf die Korrelation mit der Chemienote in allen Fällen auf einem Niveau von $p < 0,01$ signifikant sind. Auffällig ist hier, dass die Korrelation zwischen Kompetenz und Chemienote deutlich geringer ausfällt als die übrigen Korrelationen. Diesbezüglich ist zu bemerken, dass im Bereich der Chemienote die Stichprobe sehr klein ausfällt. Dies hängt damit zusammen, dass bei der Ermittlung der Daten nach den Noten im letzten Halbjahreszeugnis gefragt wurde. Viele Schülerinnen und Schüler befanden sich zum Zeitpunkt der Testung im ersten Halbjahr ihres Chemieunterrichts und konnten daher hier keine Notenangaben machen. Hinzu kommt, dass diejenigen, die hier Angaben machen konnten zum größten Teil Schülerinnen und Schüler der Hauptschule (n=174) oder der Realschule (N=201) waren, während Schülerinnen und Schüler des Gymnasiums hier kaum vertreten sind (n=13). Die Stichprobe ist im Fall der Chemienote demnach doppelt eingeschränkt.

Tabelle 5.27: Korrelationen zwischen gemessener Kompetenz und Schulnoten

Noten	Deutsch	Englisch	Mathematik	Physik	Chemie	Biologie
Kompetenz	0,156**	0,231**	0,273**	0,195**	0,101*	0,210**
n	741	734	740	727	388	722

** = $p < 0,01$ (einseitig) * = $p < 0,05$ (einseitig)

Folgende Hypothesen werden geprüft:

1. Vor dem Hintergrund der Ergebnisse aus Metaanalysen zum Zusammenhang zwischen Leistungskonstrukten sollten gemessene Kompetenz und *Naturwissenschaftsnoten* eine signifikant positive Korrelation erreichen, die größer ist als $0,3$ (vgl. Abschnitt 2.6.1, S. 72):

$$r_{(Kompetenz,Naturwissenschaftsnoten)} > 0,3.$$

Die Hypothese wird durch die Datenlage nicht gestützt. Die Korrelationen zwischen Kompetenz und Physik- sowie Biologienote liegen gerundet bei $0,2$ und damit deutlich unter der angestrebten Höhe. Die Korrelation mit der Chemienote weicht sogar noch stärker ab. Führt man, die im Theorieteil (vgl. Abschnitt 2.6.1) beschriebenen Abhängigkeiten der *Schulnoten* vom schulischen Referenzrahmen berücksichtigend, getrennte Berechnungen der Korrelationen zwischen *Naturwissenschaftsnoten* und gemessener Kompetenz durch, so ergeben sich im Gymnasium Werte bis $r = 0,3$ (Korrelation zwischen Physiknote und Kompetenz). Hier wird also die in der Hypothese geforderte Korrelationshöhe erreicht.

2. Die Korrelation zwischen Kompetenz und den *Naturwissenschaftsnoten* liegt signifikant über der Korrelation der Kompetenz mit den *nicht-naturwissenschaftlichen Noten*:

$$r_{(Kompetenz,Naturwissenschaftsnoten)} > r_{(Kompetenz,nicht-naturw.Noten)}.$$

Ohne dass eine weitere statistische Prüfung durchgeführt werden muss, ist es augenscheinlich, dass diese Hypothese verworfen werden muss. Die Korrelation der Kompetenz mit den *nicht-naturwissenschaftlichen Noten* übersteigt sogar im Fall der Englisch- und Mathematiknote die Korrelation zwischen Kompetenz und *Naturwissenschaftsnoten*.

PRÜFUNG DER DISKRIMINANTEN VALIDITÄT

Die Prüfung der diskriminanten Validität erfolgt anhand von Korrelationen der gemessenen Kompetenz mit verschiedenen Interessenskalen (Motivationskonstrukten). Hier sind zum einen das *Fachinteresse* und zum anderen das *Interesse an naturwissenschaftsbezogenen Aktivitäten* sowie das *Interesse an naturwissenschaftlichen Tätigkeiten* zu nennen. Tabelle 5.28 gibt einen Überblick über die Korrelationen zwischen Kompetenz und einzelnen Fachinteressen. Die Fachinteressen wurden inhaltlich folgendermaßen zusammengefasst und skaliert: *Sprachinteresse* (Deutsch, Englisch), *nichtnaturwissenschaftliches Interesse* (Deutsch, Englisch, Mathematik), *Mathematik- und Naturwissenschaftsinteresse* sowie *Naturwissenschaftsinteresse* (Physik, Chemie, Biologie). Es werden also an dieser Stelle die WLEs der Kompetenz mit den WLEs der Fachinteressen korreliert.

Folgende Hypothesen werden geprüft:

Tabelle 5.28: Übersicht der WLE-Korrelationen zwischen gemessener Kompetenz und Fachinteressen

Fachinteresse	Gesamt	D/E	D/E/M	N/M	N
Kompetenz	0,146**	0,025	0,096**	0,148**	0,122**

** = $p < 0,01$ (einseitig)

D = Deutsch E = Englisch M = Mathematik N = Naturwissenschaften

1. Die Korrelation zwischen Kompetenz und *Naturwissenschaftsnoten* (Korrelation zwischen zwei Leistungskonstrukten) fällt signifikant höher aus als die Korrelation zwischen Kompetenz und naturwissenschaftlichem Fachinteresse (Korrelation zwischen einem Leistungs- und einem Motivationskonstrukt):

   $$r_{(Kompetenz,Naturwissenschaftsnoten)} > r_{(Kompetenz,naturw.Fachinteresse)}.$$

   Zur Prüfung dieser Hypothese wurden die *Schulnoten* entsprechend den Fachinteressen inhaltlich zusammengefasst und in dieser zusammengefassten Form mit der Kompetenz korreliert.

   Die Korrelation zwischen Kompetenz und *Naturwissenschaftsnote* ($r = 0,243$) und die Korrelation zwischen Kompetenz und *naturwissenschaftlichem Fachinteresse* ($r = 0,122$) wurden einseitig auf einem α-Niveau von $0,05$ anhand der Fishers Z-Transformation (vgl. Abschnitt 3.2.1) getestet. Da $z_{emp} = 3,15$ größer ausfällt als die statistische Prüfgröße ($1,645$), kann die aufgestellte Hypothese bestätigt werden. Die Korrelation zwischen den beiden Leistungskonstrukten fällt also signifikant höher aus als die Korrelation zwischen dem Leistungs- und dem Motivationskonstrukt.

2. Die Korrelation zwischen Kompetenz und *naturwissenschaftlichem Fachinteresse* fällt zum einen signifikant höher aus als die Korrelation zwischen Kompetenz und *Sprachinteresse*. Zum anderen fällt sie höher aus als die Korrelation mit dem *nichtnaturwissenschaftlichen Fachinteresse*:

   $$r_{(Kompetenz,naturw.Fachinteresse)} > r_{(Kompetenz,FachinteresseD/E)}$$
   $$r_{(Kompetenz,naturw.Fachinteresse)} > r_{(Kompetenz,nicht-naturw.Fachinteresse)}.$$

Der einseitige Test auf dem α-Niveau von $0,05$ zur Prüfung der erstgenannten Hypothese ergab einen $z_{emp} = 1,80$, was bei dem weiter oben genannten Annahmebereich bedeutet, dass die Korrelation zwischen Kompetenz und *naturwissenschaftlichem Fachinteresse* ($r = 0,122$) statistisch signifikant höher ausfällt als die Korrelation zwischen Kompetenz und dem *Sprachinteresse* ($r = 0,025$). Die angenommene Hypothese kann damit aufgrund der Datenlage bestätigt werden.

Die Prüfung der zweitgenannten Hypothese fällt bei Annahme des gleichen Signifikanzniveaus und einseitigem Test mit $z_{emp} = 0,55$ deutlich *nicht* signifikant aus. Die Korrelation zwischen Kompetenz und *naturwissenschaftlichem Fachinteresse* fällt *nicht* signifikant höher aus als die Korrelation zwischen Kompetenz und *nichtnaturwissenschaftlichem Fachinteresse*. Die Hypothese muss demnach verworfen werden.

Die Darstellung der Korrelation zwischen Kompetenz und *Interesse an naturwissenschaftsbezogenen Aktivitäten* und zwischen Kompetenz und *Interesse an naturwissenschaftlichen Tätigkeiten* in Tabelle 5.29 zeigt, dass beide Skalen signifikant mit der gemessenen Kompetenz korrelieren.

Tabelle 5.29: Übersicht der WLE-Korrelationen zwischen gemessener Kompetenz und Interessensskalen

Interessensskalen	Nawi-Aktivitäten	Nawi-Arbeitsweisen
Kompetenz	0,139**	0,206**

** = $p < 0,01$ (einseitig)

Folgende Hypothese wird anhand dieser Daten geprüft:

1. Die Korrelation zwischen Kompetenz und *Interesse naturwissenschaftlichen Tätigkeiten* sollte aufgrund der inhaltlichen Nähe zum Testinstrument signifikant höher ausfallen als die Korrelation zwischen Kompetenz und *Interesse an naturwissenschaftsbezogenen Aktivitäten*:

$$r_{(Kompetenz, naturwissenschaftlichen Tätigkeiten)} > r_{(Kompetenz, naturwissenschaftsbezogene Aktivitäten)}.$$

Der einseitige Test auf dem α-Niveau von $0,05$ zur Prüfung der Hypothese ergibt einen $z_{emp} = 1,89$. Da dieser Wert damit größer ist als der kritische Wert von $1,645$,

kann die Hypothese aufgrund der Datenlage bestätigt werden: die Korrelation zwischen Kompetenz und dem *Interesse an naturwissenschaftlichen Tätigkeiten* ($r = 0,206$) fällt signifikant höher aus als die Korrelation zwischen Kompetenz und dem Interesse an naturwissenschaftsbezogenen Aktivitäten ($r = 0,139$).

5.5.2 PRÜFUNG DER RELIABILITÄT

Wie bereits weiter oben erwähnt, werden bei der Prüfung der Reliabilität neben der probabilistischen WLE-Reliabilität weiterhin auch die klassische Reliabilität in Form von Cronbachs Alpha dargestellt, um eine Vergleichbarkeit zu bisherigen Multiple-Choice-Verfahren im Bereich der Erfassung naturwissenschaftlicher Kompetenzen sicherzustellen, die vor dem Hintergrund der klassischen Testtheorie konstruiert wurden.

Die folgende Tabelle 5.30 gibt einen Überblick über die genannten Reliabilitäten, die gemäß einer Klassifizierung von Bortz (1999) als hoch bewertet werden können. Im Rahmen der Diskussion werden die Reliabilitätskoeffizienten interpretiert und zu

Tabelle 5.30: Übersicht über probabilistische und klassische Reliabilitäten

Reliabilität	
WLE	0,77
Cronbachs Alpha	0,81

den Reliabilitäten ähnlicher Testverfahren in Beziehung gesetzt.

6 Diskussion

Im Rahmen der Diskussion werden die Haupttestergebnisse und insbesondere die nach Entfernung der qualitativ unzureichenden Itemsets ermittelten Kennwerte der abschließenden Testversion interpretiert. Die Schlussfolgerungen aus Pilotierung und Feldtest sind nur in geringem Maß Teil dieser Betrachtungen, da sie bereits im Ergebnisteil diskutiert wurden, um die Überarbeitung von Items begründen und darstellen zu können.

Die Diskussion gliedert sich somit in die Interpretation und kritische Bewertung der Haupttestergebnisse nach Itementfernung und die Prüfung der Gütekriterien. Den Abschluss der Diskussion bildet der Vergleich des Tests mit bisher bestehenden Verfahren.

6.1 Interpretation der Haupttestergebnisse

Im Rahmen der Interpretation der Haupttestergebnisse wird zunächst eine qualitative Einordnung der anhand der abschließenden Testversion geschätzten Item- und Personenparameter vorgenommen. Im Anschluss folgt eine kurze Betrachtung einiger weniger Auffälligkeiten, die sich im Hinblick auf fehlende Werte ergeben haben. Den Abschluss dieses Abschnitts bilden die Schlussfolgerungen aus der Validitäts- und Reliabilitätsprüfung. Hier steht vor allem der Umgang mit den Ergebnissen der DIF-Analysen im Vordergrund.

6.1.1 Einordnung der statistischen Kennwerte

Zu Beginn der Einordnung der statistischen Kennwerte werden zunächst die Trennschärfen betrachtet. Die Bewertung der Aufgabenschwierigkeiten erfolgt in Relation zur Verteilung der Personenfähigkeiten.

Trennschärfen

Die Trennschärfen der abschließenden Testversion können bis auf eine Ausnahme (Sol1) nach dem Kriterium von Kelava und Moosbrugger (2007) mit Werten $\geq 0,4$

als gut bezeichnet werden. Die durchschnittlichen Trennschärfen der Itemsets liegen vollständig in diesem Bereich. Hinsichtlich der Trennschärfen kann die Testentwicklung somit als gelungen bezeichnet werden.

AUFGABENSCHWIERIGKEITEN

Das Kriterium zur Beurteilung der Aufgabenschwierigkeiten ist eine an die Verteilung der Personenfähigkeiten angepasste Verteilung. Dem liegt zu Grunde, dass es zur optimalen Differenzierung aller Fähigkeitsniveaus Items in jedem Fähigkeitsbereich geben sollte. Ein erstes Indiz für eine dementsprechend angepasste Testschwierigkeit ist der Vergleich der mittleren Aufgabenschwierigkeit mit der mittleren Personenfähigkeit, die hier auf 0 fixiert wurde. Es wird deutlich, dass die mittlere Schwierigkeit der Items mit einem Wert von $-0,14$ in der Nähe der mittleren Personenfähigkeit liegt. Der Test kann damit für die vorliegende Stichprobe in der Schwierigkeit als angemessen bezeichnet werden, auch wenn er ein wenig zu leicht erscheint.

Ein Blick auf die Verteilung der Itemschwierigkeiten anhand der in Abbildung 5.16 (S. 199) dargestellten Wright-Map erlaubt ein differenzierteres Urteil. Hier ist zu erkennen, dass sich die Aufgaben gut um den mittleren Fähigkeitsbereich der Testpersonen verteilen. Allerdings muss kritisch angemerkt werden, dass nicht genügend Items vorhanden sind, welche die Randbereiche, also das obere und untere Fähigkeitsdrittel, abdecken. Die Folge daraus ist, dass der Test zwar gut zwischen Personen im mittleren Fähigkeitsbereich diskriminiert, aber Personen im oberen und unteren Fähigkeitsbereich aufgrund der dort fehlenden Items nicht differenziert genug erfasst. Dieser Umstand wird bei einer Betrachtung der Verteilung der Personenfähigkeiten in den einzelnen Schulstichproben deutlich und ist bei der Anwendung des Tests zu beachten (vgl. Anhang, Abbildungen D.9-D.11, ab S. 267). Will man über eine Stichprobe Aussagen treffen, für die der Test eine optimale Schwierigkeit aufweist, wie es in der vorliegenden Stichprobe der Fall ist, so ist die Anwendung des Tests unproblematisch, weil er in diesem Fall ausreichend Aussagekraft besitzt. Dies trifft auch auf die Teilstichprobe der Realschülerinnen und Realschüler zu. Setzt man den Test allerdings allein in einer Gymnasial- oder Hauptschulstichprobe ein, so sind Einschränkungen in Kauf zu nehmen. Wie Tabelle 5.21 (S. 200) gezeigt hat, ist der Test für die Gymnasialstichprobe zu leicht, es gibt einen Deckeneffekt. Dementsprechend diskriminiert der Test gut im mittleren und unteren Leistungsbereich, kann aber ab einer gewissen oberen Leistungsgrenze keine Aussagen mehr im Hinblick auf die Kompetenz treffen.

Umgekehrt verhält es sich im Fall der Hauptschulstichprobe. Hier kommt es zum Bodeneffekt. Für die Hauptschülerinnen und Hauptschüler ist der Test zu schwer, so dass er lediglich im mittleren und oberen Leistungsbereich zu differenzieren vermag. Im unteren Leistungsbereich fehlt es an leichten Items, um auch extrem niedrige Leistungsniveaus feststellen zu können.

Für den Einsatz dieses Testinstruments bedeutet dies, dass die Einschränkungen in der Beurteilung von Gruppenunterschieden beachtet werden sollten, wenn es um Gruppen extremer Leistungsbereiche geht. Gruppen mittlerer Leistungsbereiche sind anhand des Tests im Hinblick auf ihre *prozessbezogene naturwissenschaftliche Grundbildung* optimal beurteilbar.

6.1.2 BEURTEILUNG FEHLENDER WERTE

Wie bereits in Abschnitt 5.4.2 beschrieben, kann der Anteil fehlender Werte in der Hauptstudie insgesamt als sehr gering angesehen werden und bietet insofern keinen Anlass zur Diskussion. Auffällig bleibt dennoch, dass sowohl hinsichtlich der drei Fähigkeitskomplexe als auch hinsichtlich der Schulniveaus keine Gleichverteilung der fehlenden Werte vorliegt. Um für zukünftige Item- und Testentwicklungen mögliche Fehlerquellen, zumindest aber zu beachtende Punkte zu identifizieren, werden diese beiden Abweichungen noch einmal genauer betrachtet.

Der Fähigkeitskomplex *H* (*Identifizieren wissenschaftlicher Hypothesen*) weicht deutlich von den beiden anderen Komplexen *P* (*Planen einer wissenschaftlichen Untersuchung*) und *N* (*Nutzen wissenschaftlicher Ergebnisse*) ab, auf die annähernd gleich viele fehlende Werte entfallen.

Ähnlich verhält es sich mit der ungleichen Verteilung fehlender Werte auf die Schulniveaustichproben. Hier weist die Hauptschulstichprobe im Vergleich zu den anderen beiden Schulformen deutlich mehr fehlende Werte auf. Im Folgenden werden drei Gründe diskutiert, die für diese Ergebnisse verantwortlich sein können.

1. Itemschwierigkeit
   - *in Bezug auf den Fähigkeitskomplex*: Ein Grund, warum der Fähigkeitskomplex *H* eine erhöhte Anzahl fehlender Werte aufweist, könnte darin bestehen, dass er im Durchschnitt eventuell eine höhere Itemschwierigkeit besitzt als die beiden anderen Komplexe. Diese Möglichkeit konnte nach einem Vergleich der klassischen Itemschwierigkeiten der einzelnen Komplexe verworfen werden. Im Mittel liegt die klassische Itemschwierigkeit (relative Lösungshäufigkeit) des Bereichs *H* bei $0,49$. Sie fällt damit zwar etwas höher aus als die des Bereichs *P* mit

$0,48$, liegt aber unter der Itemschwierigkeit des Bereichs *N* mit einem Wert von $0,50$.

- *in Bezug auf die Hauptschulstichprobe*: Im Durchschnitt liegen die Hauptschülerinnen und Hauptschüler mit einem Logit-Wert von $-0,614$ hinsichtlich ihrer Fähigkeiten unter der durchschnittlichen Itemschwierigkeit des Tests, die im Rahmen der Schätzung der Personenparameter auf Null fixiert wurde. Damit ist der Test für sie im Mittel deutlich zu schwer. Es erscheint also möglich, dass die Schülerinnen und Schüler Items aufgrund einer Überforderung auslassen. Sie könnte zu einem Nachlassen der Motivation und schließlich zum Auslassen einzelner Items führen (vgl. Motivationstheorien:Weiner, 1986; Heckhausen & Heckhausen, 2006). Da diese Überforderung weniger die Realschul- und kaum die Gymnasialstichprobe betrifft, für die der Test in der Schwierigkeit angemessen bzw. zu leicht ist, stellt dieser Ansatz eine mögliche Erklärung der größeren Anzahl fehlender Werte dar.

2. Leseaufwand
   - *in Bezug auf den Fähigkeitskomplex*: Ein weiterer Erklärungsansatz für die im Vergleich zu den anderen Fähigkeitskomplexen höhere Anzahl fehlender Werte könnte darin liegen, dass der Bereich H einen erhöhten Leseaufwand bietet. Auch dieser Erklärungsansatz lässt sich bei einer Überprüfung der durchschnittlichen Wörteranzahl pro Item nicht halten. Die Items des Bereichs H besitzen eine durchschnittliche Anzahl von 144 Wörtern. Zwar liegt der Bereich P mit durchschnittlich 120 Wörtern pro Item noch darunter, doch der Bereich E liegt mit 165 Wörtern pro Item darüber.
   - *in Bezug auf die Hauptschulstichprobe*: Für die Hauptschulstichprobe stellt der Leseaufwand eine nicht unwesentliche Schwierigkeit dar. Die PISA-Ergebnisse 2006 (Drechsel & Artelt, 2007) haben gezeigt, dass die Lesekompetenz der Hauptschülerinnen und Hauptschüler deutlich unter der Lesekompetenz der Realschülerinnen und Realschüler und der Gymnasiastinnen und Gymnasiasten liegt. Lange Texte in Testaufgaben stellen allgemein ein Problem dar, da sie die Aufgabenschwierigkeit erhöhen können und Schülerinnen und Schüler mit geringerer Lesekompetenz benachteiligen. Vor diesem Hintergrund erscheint es möglich, dass hoher Leseaufwand die Schülerinnen und Schüler der Hauptschule in höherem Maße als die Schülerinnen und Schüler der Real- und Hauptschulstichprobe dazu veranlasst, Items auszulassen.

3. Positionierung der Items

- *in Bezug auf den Fähigkeitskomplex*: Wie bereits im Theoriekapitel (s. Abschnitt 4.3) ausgeführt, wurde darauf geachtet, dass die zehn Items, die pro Fähigkeit entwickelt wurden, über die Itemsets hinweg gleichmäßig auf die drei innerhalb eines Sets möglichen Positionen verteilt wurden. Weiterhin wurde die Reihenfolge der Itemsets für jedes Testheft per Zufall bestimmt. Aus diesem Grund kann eine ungünstige Positionierung der Items des Fähigkeitsbereichs H nicht der Grund für den höheren Anteil fehlender Werte sein.

Insgesamt ergeben sich aus den Ausführungen, dass die unterschiedliche Verteilung fehlender Werte nicht auf die hier überprüfbaren Merkmale der Fähigkeitskomplexe zurückzuführen sind. Es konnte also nicht abschließend geklärt werden, worin hier die Ursache liegen könnte. Hinsichtlich der Hauptschulstichprobe kann die Ursache einer erhöhten Anzahl fehlender Werte an dieser Stelle ebenfalls nicht abschließend geklärt werden. Die Vermutung, dass die Ursache in einer allgemeinen Überforderung und im Leseaufwand zu suchen ist, erscheint allerdings sehr wahrscheinlich.

Hinsichtlich zukünftiger Itementwicklungen, die schulniveau-übergreifend erfolgen sollen, kann geschlussfolgert werden, dass die Texte noch kürzer gehalten sein sollten, um Schülerinnen und Schüler mit geringerer Lesekompetenz nicht zu benachteiligen.

6.1.3 BEDEUTUNG DER VALIDIERUNG

Im Folgenden werden die Ergebnisse der Validierung diskutiert und eingeordnet. Die interne Validierung steht hier im Vordergrund, da sie der Prüfung des angenommenen Testmodells, des eindimensionalen Rasch-Modells, dient. Ergäbe die Prüfung eine grundsätzliche Verletzung der Validität, so müsste die Testentwicklung als gescheitert betrachtet werden.

Aufgaben- bzw. Testanalyse und Validierung fanden anhand der gleichen Stichprobe statt und die Validierung gründete sich auf den anhand dieser Daten reduzierten Itempool. Wie bereits im Ergebnisteil ausgeführt, ist eine unter diesen Bedingungen durchgeführte Validitätsprüfung lediglich als erster Ansatz zur Validierung zu verstehen. Abschließende Aussagen können erst im Rahmen einer Kreuzvalidierung anhand einer neuen Stichprobe getroffen werden (Lienert & Raatz, 1998).

INTERNE VALIDIERUNG

Die Prüfung der internen Validität bestand aus der Prüfung des Infits, der Untersuchung von Gruppenunterschieden, DIF-Analysen und der globalen Modellprüfung.

Diese Bereiche werden auf der Grundlage der Ergebnisse und in Bezug auf bestehende Forschungsergebnisse eingeordnet und einzeln diskutiert.

Prüfung des Infits

Die Prüfung des Infits ergab letztlich ein Item (*Sol1*), dessen Trennschärfe gemäß der ICC-Betrachtungen eine Abweichung und damit eine Verletzung des Rasch-Modells zeigte. Da diese Abweichung jedoch nur ein Item von insgesamt 21 Items betrifft und die übrigen Testitems gute bis sehr gute Kennwerte zeigen, kann der Test insgesamt als trennschärfenhomogen und im Sinne des angelegten Testmodells als raschmodellkonform bezeichnet werden.

Prüfung von Gruppenunterschieden

Sollte der Test tatsächlich eine Kompetenz erfassen, die einen speziellen Bereich naturwissenschaftlicher Grundbildung darstellt, so sollten die Testergebnisse der Schülerinnen und Schüler den Testergebnissen der Schülerinnen und Schüler der nationalen PISA-Untersuchung ähneln. Die Daten konnten diese Annahme bestätigen. Die anhand des Tests erfasste eindimensionale Kompetenz scheint ebenso wie die naturwissenschaftliche Kompetenz bei PISA (Rönnebeck et al., 2008) in Bezug auf die Schülerinnen und Schüler in Schleswig-Holstein - eindimensional und nicht auf Ebene der Teilkompetenzen betrachtet - gleichermaßen stark ausgeprägt zu sein.

Hinsichtlich der Unterschiede zwischen den Schulniveaus wurde angenommen, dass sie sich aufgrund der unterschiedlich differenzierten naturwissenschaftlichen Ausbildung und in Anlehnung an die nationalen PISA-Ergebnisse 2006 für Schleswig-Holstein (Rönnebeck et al., 2008) im Hinblick auf die *prozessbezogene naturwissenschaftliche Grundbildung* ihrer Schülerinnen und Schüler signifikant unterscheiden. Die Ergebnisse zeigen, dass dies tatsächlich der Fall ist: Gymnasiastinnen und Gymnasiasten schneiden im Test signifikant besser ab als Realschülerinnen und Realschüler und diese wiederum zeigen signifikant höhere Kompetenzwerte als die Hauptschülerinnen und Hauptschüler. Anhand des Tests können also unterschiedliche Kompetenzniveaus abgebildet werden.

Die Unterschiede zwischen den Schulniveaus stellen im Vergleich zu den anderen, in diesem Abschnitt dargestellten Ergebnissen lediglich einen schwachen Hinweis auf die Validität des Verfahrens dar, da sie zwar eine notwendige, aber keine hinreichende Bedingung darstellen, um von Validität sprechen zu können. Da hier keine abschließende Aussage darüber getroffen werden kann, inwiefern die im Schulniveau-

vergleich ermittelten Unterschiede nur auf die gemessene *prozessbezogene naturwissenschaftliche Grundbildung* zurückzuführen ist oder ob noch andere Einflüsse zu diesen Unterschieden geführt haben, wurden DIF-Analysen durchgeführt, die nachfolgend betrachtet werden.

DIF-Analysen

Als *Differential Item Functioning (DIF)* wird der Umstand bezeichnet, dass ein Item in unterschiedlichen Gruppen unterschiedliche statistische Eigenschaften zeigt. Es stellt im Sinne der Item-Response-Theorie einen Hinweis auf Items dar, die etwas anderes messen als intendiert und ist damit eine Möglichkeit zur Prüfung der internen Validität.

Abschnitt 5.5.1 gibt Auskunft darüber, dass im Vergleich der Geschlechterstichproben zwar Abweichungen auftreten, diese aber so marginal ausfallen, dass hier nicht von DIF-Effekten gesprochen werden kann. Auffälliger sind hier die Unterschiede, die sich zwischen den Schulniveaustichproben ergeben haben. Die Items Bro2, Kli2+3, Ros2, Sko2+3 und Sol1 weisen DIF auf. Die meisten DIFs konnten zwischen Hauptschul- und Gymnasialstichprobe sowie zwischen Realschul- und Gymnasialstichprobe identifiziert werden. Lediglich im Falle der Items Sko2 und Kli3 zeigen sich als mittelmäßig zu klassifizierende DIF-Effekte zwischen Haupt- und Realschulstichprobe. Als groß zu bewertende DIFs werden im Rahmen des Vergleichs zwischen Haupt- und Gymnasialstichprobe in Form der Items Kli2, Ros2, Sko2 und Sol1 offenbar.

An dieser Stelle wird versucht, eine Antwort auf die Frage zu finden, wie mit der Feststellung derartiger Abweichungen umgegangen werden kann und welche Konsequenzen sich für die Testentwicklung oder für den Umgang mit den Testergebnissen ergeben.

Die Analyse des DIF kann für Testentwickler eine sehr wertvolle Methode darstellen. Dennoch ist es schwierig, die Ursachen des DIF zweifelsfrei und abschließend zu klären. Es gibt meist viele interne und externe Faktoren, die herangezogen werden können, um festzustellen, inwiefern die Unterschiede in den Itemschwierigkeiten unterschiedlicher Personengruppen in einem von der Testintention abweichenden Zusammenhang stehen. Und selbst wenn keine oder nur geringe DIF-Effekte festgestellt werden, bedeutet dies nicht, dass der Test kein DIF enthält bzw. fair für alle Gruppen von Testpersonen ist (Camilli, 1993). Schmitt et al. (1993) nennen drei Gründe, warum sich die Identifikation von DIF-Ursachen so schwierig gestaltet:

- Bisher liegt der Fokus der DIF-Forschung noch immer auf Methoden zur Identifi-

kation von DIF.

- Die Identifikation von Faktoren, die DIF verursachen, erfordert Theorien über differentielle Itemschwierigkeiten.
- Die Identifikation von Faktoren ist komplex. Mehr als ein Faktor kann für das DIF eines Items verantwortlich sein.

Als Folge von Identifikations- und Interpretationsschwierigkeiten möglicher Faktoren sind auch Maßnahmen für den Umgang mit DIF nicht eindeutig zu bestimmen. Verschiedene Autoren machen Vorschläge, die einen vernünftigen Umgang mit der Identifikation von DIF zum Ziel haben.

- Zunächst sollte festgestellt werden, dass die DIF-Effekte kein Ergebnis von Zufallsschwankungen sind. Sie sollten also anhand einer weiteren Stichprobe erneut überprüft werden (Du, 1995; Wilson, 2005).
- Es sollte geprüft werden, inwiefern das DIF bedeutsame Auswirkungen auf den Test bzw. auf die Messintention des Tests hat (Du, 1995).
- Weiterhin müssen Hypothesen darüber aufgestellt werden, welche Faktoren für das DIF bestimmter Items verantwortlich sein könnten (Schmitt et al., 1993; Wilson, 2005).
- In einem weiteren Schritt werden diese Hypothesen durch speziell entwickelte Items getestet (Schmitt et al., 1993). In diesem Zuge müssen zunächst alle Items bezüglich ihrer Anforderungen und Merkmale beschrieben werden. Im Anschluss daran werden jeweils zwei Itemversionen erstellt, die sich lediglich in dem hypothetisch für den DIF-Effekt angenommenen Faktor unterscheiden und in Schwierigkeit, Trennschärfe, Distraktoren usw. gleich sind.

Die Vorschläge der Autoren machen deutlich, dass die differenzierte Untersuchung von DIF-Effekten ein komplexes Vorhaben darstellt, das den Rahmen dieser Arbeit sprengt. Aus diesem Grund können an dieser Stelle lediglich Vorschläge zur weiteren Behandlung dieser Befunde gemacht sowie mögliche Hypothesen über Ursachen des DIF aufgestellt werden.

In Anlehnung an das vorgeschlagene Vorgehen wird im Rahmen einer ohnehin notwendigen Kreuzvalidierung zur abschließenden Prüfung der Testwerte zunächst festzustellen sein, inwiefern die identifizierten DIFs lediglich zufällig auftreten oder aber systematischer Natur sind. Treten sie systematisch auf, so müssen Hypothesen

hinsichtlich möglicher Faktoren formuliert werden, die in Abhängigkeit von Aufgabeneigenschaften und in Abhängigkeit vom Schulniveau zu Item-DIFs führen. Hypothesen, die geprüft werden könnten, können sich auf folgende Faktoren beziehen:

- *Aufgabenkomplexität*: Die DIF-Items könnten sich aufgrund bestimmter Komplexitätsmerkmale von den übrigen Items unterscheiden und für die Schülerinnen und Schüler dadurch unterschiedlich beherrschbar sein. Vier der sieben DIF-Items erfordern die Fähigkeit, Schlussfolgerungen aus einer Graphik zu ziehen. Allerdings gibt es auch unter den Items, die kein DIF aufweisen, solche, für deren Lösung diese Fähigkeit notwendig ist. Es scheint, dass die DIF-Items ein spezielles Merkmal enthalten, das die Aufgabe komplexer werden lässt und ihre Lösung erschwert.

- *Textlänge*: Alle DIF-Items liegen über der durchschnittlichen Wörteranzahl pro Aufgabe und unterscheiden sich in dieser Hinsicht von dem Rest der Items. Hier scheint die Lesekompetenz in der Lösung der Aufgabe zu großen Raum einzunehmen. Da die Schülerinnen und Schüler der einzelnen Schularten sich in ihrer Lesekompetenz deutlich unterscheiden (Drechsel & Artelt, 2007), kann der Faktor Textlänge hier ein Grund für die DIF-Effekte sein.

- *Interesse*: Iteminhalte könnten für Schülerinnen und Schüler der unterschiedlichen Schulniveaus unterschiedlich interessant sein.

Erste Ansätze zur Prüfung dieser Hypothesen bietet die Befragung einer Physik- und Mathematiklehrerin, die gebeten wurde, mögliche Ursachen dafür zu finden, dass bestimmte Items bestimmten Schulniveaugruppen unterschiedlich schwer erscheinen. Dabei ergaben sich im Hinblick auf die DIF-Items folgende Ideen:

- *Bro2, Ros2*: Die Items sind für Schülerinnen und Schüler der Haupt- und Realschule schwerer zu lösen als für Gymnasiastinnen und Gymnasiasten. Die hier dargestellten Graphiken erfordern ein höheres Abstraktionsniveau als die Graphiken der übrigen Items. Im Fall von Bro2 müssen statt der sonst zu berücksichtigenden zwei Variablen drei Variablen berücksichtigt werden und im Fall von Ros2 müssen die Schülerinnen und Schüler spezielle Eigenschaften der Darstellung in Balkendiagrammen verstehen.

- *Sko2*: Dieses Item erscheint den Hauptschülerinnen und Hauptschülern schwerer als der restlichen Stichprobe. Hier kann ähnlich argumentiert werden wie im vorangehenden Abschnitt, denn auch hier müssen drei anstatt zwei Variablen betrachtet werden.

- *Kli2, Kli3, Sko3, Sol1*: Diese Items erscheinen den Schülerinnen und Schülern des Gymnasiums schwerer als den Schülerinnen und Schülern der Haupt- und/ oder Realschule. In diese Items könnten Gymnasiastinnen und Gymnasiasten mehr hineinlesen als letztlich nötig ist, um das Item zu lösen. Es könnte sein, dass ihnen die Lösung zu einfach erscheint, sie deshalb nach einem tieferen Sinn suchen und dadurch die falsche Antwortalternative wählen.

Diese ersten Ideen bezüglich möglicher Ursachen für das DIF beziehen sich auf den oben genannten Punkt der Aufgabenkomplexität. Können mehrere solcher Kommentare zusammengetragen werden, so lassen sich nach oben genanntem Vorgehen kontrollierte DIF-Analysen realisieren.

Grundsätzlich muss unter Berücksichtigung des Punktes, inwiefern das DIF bedeutsame Auswirkungen auf den Test bzw. auf die Messintention des Tests hat (Du, 1995), an dieser Stelle noch einmal darauf hingewiesen werden, dass das Testinstrument nicht vorrangig dem Vergleich zwischen Schulniveaus dient. Soll es in dieser Intention genutzt werden, so sollten zunächst tiefergehende Untersuchungen der DIF-Ursachen durchgeführt werden. Sollten die DIFs sich in diesem Zuge als nicht zufällig erweisen und somit darauf hinweisen, dass die identifizierten Items über die Messintention hinaus andere Konstrukte messen, so muss vor diesem Hintergrund eine Entscheidung getroffen werden, die auch in der Eliminierung von Items (bzw. Itemsets) bestehen kann (Orlando & Marshall, 2002).

Liegt die Testintention im einfachen Screening einzelner Stichproben ohne den Vergleich über unterschiedliche Schulniveaus hinweg, so sollte der Test unter Berücksichtigung möglicher Boden- und Deckeneffekte gemäß aktuellem Stand problemlos einsetzbar sein, eine vorangehende Kreuzvalidierung vorausgesetzt.

Globale Modellprüfung

Der *Test zur Erfassung prozessbezogener naturwissenschaftlicher Grundbildung* wurde von Beginn an mit dem Ziel entwickelt, die genannte Grundbildung als eindimensionale Kompetenz zu erfassen. Da sie sich als sehr umfangreich erwies, war es notwendig, sie anhand dreier Fertigkeiten zu operationalisieren. Darauf aufbauend wurden schließlich Items entwickelt, die zum einen spezifisch auf die einzelnen Fertigkeiten zugeschnitten und zugleich gemeinsamer Bestandteil *eines* latenten Konstrukts sein sollten.

Durch diese Art der Grundstruktur ergab sich die Notwendigkeit, im Rahmen einer globalen Modellprüfung zu untersuchen, inwiefern die Datenlage angemessener

durch ein eindimensionales oder ein dreidimensionales Modell zu erklären ist. Zu diesem Zweck wurden im Feld- und Haupttest sowie anhand der abschließenden Testversion CAICs und BICs der beiden Testmodelle verglichen. In allen drei Fällen konnte aufgrund der Modellprüfung kein klares Urteil darüber gefällt werden, welches Testmodell angemessener ist. Der CAIC- bzw. BIC-Vergleich sprach zwar in allen Fällen für das dreidimensionale Modell, doch die Unterschiede fielen im Feldtest und insbesondere in der abschließenden Testversion so gering aus, dass zusätzlich die latenten Korrelationen der drei Dimensionen betrachtet werden, um letzten Endes auf dieser Basis eine Entscheidung für eines der Modelle treffen zu können.

In allen Modellvergleichen wiesen die Dimensionen des dreidimensionalen Modells hohe ($> 0,80$) bis sehr hohe ($> 0,90$) Korrelationen auf, so dass die drei Dimensionen nicht als unabhängig bezeichnet werden konnten. Dieser Befund sprach dafür, doch von einem eindimensionalen Modell auszugehen.

Die Frage danach, wie diese Befunde zusammenzufassen bzw. einzuordnen sind, kann im Grunde durch die Darstellung der Operationalisierung erklärt werden. Gerade vor diesem Hintergrund erscheinen diese zunächst nicht eindeutigen Ergebnisse logisch. Die drei Fertigkeiten, die der Test misst und anhand derer die *prozessbezogene naturwissenschaftliche Grundbildung* operationalisiert wurde, sind zwar Teil einer zu Grunde liegenden Kompetenz, wurden aber jeweils zur Messung einer speziellen Fertigkeit entwickelt und weisen somit gewisse Unterschiede auf. Diese Unterschiede sind jedoch nicht so groß, dass die drei Fertigkeiten eigene unabhängige Dimensionen bilden. Genau diesen Umstand, also die ursprüngliche Konstruktionsabsicht, spiegeln die vorliegenden Ergebnisse der Modellprüfung wider. Insofern kann hier durchaus davon gesprochen, dass der Test diesbezüglich inhaltlich valide ist.

Die Konsequenz, die während der einzelnen Testentwicklungsphase aus den beschriebenen Ergebnissen gezogen wurde, nämlich die Daten zusammengefasst eindimensional zu betrachten, ist durch die Ergebnisse der PISA-Untersuchung begründet. Hier weisen die naturwissenschaftlichen Teilkompetenzen Korrelationen zwischen $0,90$ und $0,93$ auf (OECD, 2009). Sie wurden neben einer gesonderten Auswertung ebenfalls als Gesamtskala ausgewertet. Ebenso wurde aufgrund der hohen Korrelationen der Dimensionen hier vorgegangen. Anders als bei PISA war es hier jedoch gemäß der Testintention weder beabsichtigt noch anhand der abschließenden Testversion möglich, die Kompetenz der Schülerinnen und Schüler im Hinblick auf die drei Fertigkeiten gesondert festzustellen: Sie wiesen lediglich WLE-Reliabilitäten zwischen $0,39$ und $0,49$ auf, die für einen solchen Zweck als unzureichend bezeichnet werden müssen.

EXTERNE VALIDIERUNG

Im Rahmen der externen Validierung wurden die Ergebnisse des zu validierenden Tests in Beziehung zu externen Validitätskriterien gesetzt. Der Test wurde mit verschiedenen Außenkriterien korreliert. Dazu gehörten als Leistungskriterium die *Schulnoten* und als Motivationskriterien das *Schulfachinteresse*, das *Interesse an naturwissenschaftsbezogenen Aktivitäten* und das *Interesse an naturwissenschaftlichen Arbeitsweisen*.

Prüfung der konvergenten Validität

Bei der Prüfung der *konvergenten Validität* konnte die Annahme, dass die Korrelation zwischen gemessener Kompetenz und den Naturwissenschaftsnoten eine Höhe $> 0,3$ erreichen sollte, nicht bestätigt werden. Zwar fielen die Korrelationen im Falle aller Naturwissenschaftsnoten signifikant aus, allerdings blieb selbst die am stärksten ausgeprägte Korrelation mit der Biologienote ($r = 0,21$) weit unter der in Abschnitt 2.6.1 begründeten, erwartbaren Korrelation. Auffällig ist hier weiterhin, dass sowohl die Sprachnoten als auch die Mathematiknote signifikant und im Falle der Englisch- und Mathematiknote höher ausfallen als die Korrelationen zwischen Naturwissenschaftsnoten und Kompetenz. Somit musste auch die zweite Hypothese, die Korrelation zwischen Kompetenz und Naturwissenschaftsnoten falle höher aus als die Korrelation zwischen Kompetenz und nicht-naturwissenschaftlichen Noten, verworfen werden.

Ursachen für diese Ergebnisse können zum einen in dem Kriterium Schulnote selbst liegen. Wie bereits im Theorieteil ausgeführt (vgl. Abschnitt 2.6.1), erfassen Schulnoten nicht allein fachliche Kompetenzen, sondern sind ein Produkt diverser Einflüsse: Der Referenzrahmen der Klasse, Schülermerkmale, Erwartungen, Voreinstellungen und Diagnosekompetenz des Lehrers spielen dabei eine bedeutsame Rolle (Schrader & Helmke, 2001). Das oft klasseninterne und nicht in Form objektiver Kriterien bestehende Bezugssystem hat einen so großen Einfluss, dass vergleichbare Leistungen in unterschiedlichen Klassen nicht zur gleichen Beurteilung führen. Erreichen in diesem Sinne die Schülerinnen und Schüler eines Schulniveaus leichter eine akzeptable naturwissenschaftliche Beurteilung als Schülerinnen und Schüler eines anderen Schulniveaus, so ist es möglich, dass die Korrelationen zwischen Kompetenzausprägung und Naturwissenschaftsnote geringer als angenommen ausfallen. Für diese Erklärung spricht das nach Schulniveau getrennte Korrelationsergebnis zwischen Naturwissenschaftsnoten und Kompetenz, das im Gymnasium ein Wert von $r = 0,3$ (Physiknote - Kompetenz) erreicht.

Einen weiteren Hinweis auf die Begründung der geringen Korrelationen zwischen Naturwissenschaftsnoten und Kompetenz können auch die TIMSS-Ergebnisse liefern (Köller, Baumert & Neubrand, 2000). Hier zeigt sich, dass Personen mit einem guten metatheoretischen Verständnis der Physik als Wissenschaft, also Personen mit angemessenen epistemologischen Überzeugungen nicht die besten Schulleistungen im Fach Physik zeigen. Es besteht ein negativer korrelativer Zusammenhang zwischen Metaverständnis und Schulleistung. Dieses Ergebnis könnte, übertragen auf die *prozessbezogene naturwissenschaftliche Grundbildung*, darauf hinweisen, dass diese Kompetenz nicht in die anhand von Schulnoten erfassten Naturwissenschaftleistungen eingeht.

In die gleiche Richtung zielt der folgende Ansatz, der versucht, die im Vergleich zu den Naturwissenschaftsnoten hohen Korrelationen der Kompetenz mit den Sprachnoten zu erklären. Sie könnten dadurch begründet sein, dass die gemessene Kompetenz neben der eigentlichen Messintention in geringem Maße auch Kompetenzen erfasst, die gewichtiger in die Schulnoten einfließen als es die Elemente tun, die im Rahmen des *Tests zur Erfassung prozessbezogener naturwissenschaftlicher Grundbildung* erfasst werden. Dazu könnte eben der Einfluss sprachlicher Kompetenzen gehören, die klarer Bestandteil der Deutsch- und Englischnote sind und die zweifellos benötigt werden, um die Testaufgaben inhaltlich erfassen und im Anschluss richtig beantworten zu können. Die Lesekompetenz ist nicht das Ziel der Kompetenzmessung, doch die ermittelten Kompetenzwerte enthalten ihren Einfluss. Da anzunehmen ist, dass sprachliche Kompetenzen größerer Bestandteil der Sprachnoten sind als die prozessbezogene naturwissenschaftliche Grundbildung Bestandteil der Naturwissenschaftsnoten ist, lassen sich die vergleichsweise hohen Korrelationen mit den Sprachnoten bzw. die geringen Korrelationen mit den Naturwissenschaftsnoten erklären. Für diese Argumentation sprechen die Ergebnisse der PISA-Untersuchung 2006 im Hinblick auf das Experimentieren im Unterricht (Seidel, Prenzel, Wittwer & Schwindt, 2007). Die im Rahmen des Experimentierens im Vordergrund stehende Fertigkeit, die berichtet wird und gleichzeitig Teil des entwickelten Tests ist, ist das *Ziehen von Schlüssen*. Schülerinnen und Schüler führen in den meisten Fällen Experimente lediglich nach Anweisung durch oder schauen den Demonstrationsexperimenten der Lehrkraft zu. Eigenständiges Experimentieren ist kaum Teil des Unterrichts. Dies bedeutet, dass mit dem Experimentieren zusammenhängende Fähigkeiten wie das Identifizieren wissenschaftlicher Hypothesen oder das Planen einer wissenschaftlichen Untersuchung kaum Teil der Naturwissenschaftsnote sein können.

Vor dem Hintergrund dieser Befunde erscheint die Erklärung der geringen Korre-

lationen zwischen prozessbezogener naturwissenschaftlicher Kompetenz und Naturwissenschaftsnoten plausibel.

Für die hohe Korrelation der Mathematiknote mit der Kompetenznote können Gemeinsamkeiten der für beide Leistungen notwendigen kognitiven Grundlagen verantwortlich sein. Wie bereits in Abschnitt 2.3 festgestellt, besteht eine wichtige kognitive Grundlage der *prozessbezogenen naturwissenschaftlichen Grundbildung* im logisch-schlussfolgernden Denken und Problemlösen. Diese Grundlagen sind ebenfalls für die Bewältigung mathematischer Probleme notwendig. Analog zum Fall der Sprachnoten können diese kognitiven Grundlagen in größerem Zusammenhang mit der Mathematiknote stehen als die *prozessbezogene naturwissenschaftliche Grundbildung* mit den Naturwissenschaftsnoten.

Zusammenfassend kann festgehalten werden, dass die geringe Korrelation zwischen Kompetenz und Schulnoten durchaus in der Unangemessenheit der Noten als Prädiktoren für das Abschneiden im *Test zur Erfassung prozessbezogener naturwissenschaftlicher Grundbildung* liegen kann. Darüber hinaus muss kritisch betrachtet werden, dass die hohe Korrelation mit den Sprachnoten zwar logisch zu erklären ist, aber dennoch auf eine eingeschränkte Validität des Testinstruments hinweist, da es nicht nur die definierte naturwissenschaftliche Kompetenz sondern auch Lesekompetenz zu erfassen scheint. Allerdings ist dieser Umstand bei Paper-und-Pencil-Verfahren nicht ganz zu vermeiden bzw. wäre nur dadurch teilweise zu vermeiden, dass die Aufgaben vorgelesen werden oder z.B. eine computerbasierte audio-visuelle Darstellungsform gefunden wird.

Prüfung der diskriminanten Validität

Bei der Prüfung der diskriminanten Validität konnte zunächst die Annahme, die Korrelation zwischen zwei Leistungskonstrukten (Kompetenz - Naturwissenschaftsnoten) falle signifikant höher aus als die Korrelation zwischen einem Leistungs- und einem Motivationskonstrukt (Kompetenz - naturwissenschaftliches Fachinteresse) bestätigt werden. Dieses Ergebnis kann als Hinweis darauf gewertet werden, dass das durch den Test erfasste Konstrukt tatsächlich ein Leistungskonstrukt darstellt.

Ein anderes Bild ergab sich bei der Überprüfung der Hypothesen, die Korrelation der Kompetenz mit dem *naturwissenschaftlichen Fachinteresse* falle zum einen signifikant höher aus als die Korrelation mit dem *Sprachinteresse* und zum anderen ebenfalls signifikant höher als die Korrelation mit dem *nicht-naturwissenschaftlichen Fachinteresse*. Hinter diesen Hypothesen stand die Annahme, dass ein ausgeprägtes naturwissenschaftliches Interesse im Sinne einer epistemischen Orientierung zu einer

tieferen Beschäftigung mit Naturwissenschaften und damit unter anderem zu einem differenzierten Verständnis naturwissenschaftlicher Forschungsprozesse führen sollte, das sich wiederum in höheren Testwerten niederschlagen sollte.

Die erste Hypothese ließ sich anhand der Datenlage bestätigen. Der Zusammenhang zwischen *naturwissenschaftlichem Fachinteresse* und gemessener Kompetenz fällt zum einen hoch signifikant und zum anderen signifikant höher aus als der Zusammenhang zwischen *Sprachinteresse* und gemessener Kompetenz, der nicht signifikant ist. Dieses Ergebnis kann im Sinne der epistemischen Orientierung als indirektes Anzeichen für die Validität des Tests gewertet werden.

Die zweite Hypothese ließ sich anhand der Daten nicht bestätigen. Allerdings zeigt sich eine Tendenz, dass die Korrelation zwischen Kompetenz und *naturwissenschaftlichem Fachinteresse* höher ausfällt als die Korrelation zwischen Kompetenz und *nichtnaturwissenschaftlichem Interesse*. Dieses Ergebnis spricht für die bereits zu Beginn dieses Abschnitts im Rahmen der Prüfung der konvergenten Validität vermuteten Gründe für die hohe Korrelation zwischen Mathematiknote und Kompetenz. Umfasst das Interesse an Mathematik ein Interesse am logisch-schlussfolgernden Denken und am Problemlösen, so sind dies Bereiche, die auch im naturwissenschaftlichen Fachinteresse eine Rolle spielen und damit dazu führen können, dass keine signifikanten Unterschiede sichtbar werden. Das Interesse an Mathematik müsste in diesem Sinne eher dem Interesse an Naturwissenschaften zugeordnet werden.

Den Abschluss der Prüfung der diskriminanten Validität bildete der Korrelationsvergleich zwischen Kompetenz und *Interesse an naturwissenschaftsbezogenen Aktivitäten* und zwischen Kompetenz und *Interesse an naturwissenschaftlichen Arbeitsweisen*. In beiden Fällen fielen die Korrelationen signifikant positiv aus. Die Ergebnisse konnten zeigen, dass die Korrelation mit dem *Interesse an naturwissenschaftlichen Arbeitsweisen*, wie vermutet, signifikant höher ausfällt als die Korrelation mit dem *Interesse an naturwissenschaftsbezogenen Aktivitäten*. Dieses Ergebnis deutet darauf hin, dass die gemessene Kompetenz größere inhaltliche Überschneidungen mit dem *Interesse an den naturwissenschaftlichen Arbeitsweisen* zeigt als mit dem *Interesse an naturwissenschaftsbezogenen Aktivitäten*. Diese inhaltliche Nähe ist anhand der in Kapitel 4 abgebildeten Skalen (s. Tabelle 4.3, S. 150 und Tabelle 4.4, S. 151) nachvollziehbar. Vergleicht man die Fertigkeiten, die anhand des Tests erfasst werden mit dem *Interesse an naturwissenschaftlichen Tätigkeiten*, so weisen insbesondere folgende Items eine inhaltliche Nähe auf:

- Wie viel Interesse hast du an folgenden naturwissenschaftlichen Dingen?
  - Einen Versuch aufbauen,

- Sich ausdenken, wie man eine bestimmte Vermutung durch einen Versuch prüfen könnte,
- Den Wert oder Nutzen einer neuen naturwissenschaftlichen Erkenntnis beurteilen.

Somit spricht das vorliegende Ergebnis für die Validität des Tests.

ZUSAMMENFASSENDE BEWERTUNG

Nach der Darstellung und Einordnung der einzelnen Validierungsergebnisse soll nun eine Bewertung der Validierung als Ganzes erfolgen. Es werden zunächst die Punkte dargestellt, die darauf andeuten, dass der *Test zur Erfassung prozessbezogener naturwissenschaftlicher Grundbildung* in valider Weise misst, was er zu messen beansprucht. Im Anschluss wird auf Fragen eingegangen, die in Bezug auf die Validierung noch nicht abschließend geklärt werden konnten.

Die im Folgenden dargestellten Punkte können als erste Anzeichen eines validen Testinstruments bezeichnet werden.

- *Interne Validierung*:
  - Die Trennschärfen sind bis auf ein abweichendes Item als homogen zu bezeichnen und erfüllen damit eine wichtige Voraussetzung, um von der Geltung des eindimensionalen Raschmodells ausgehen zu können.
  - Die Prüfung der Gruppenunterschiede (nach Geschlecht bzw. Schulniveau) zeigen die auf Grundlage bestehender Forschungsergebnisse zur naturwissenschaftlichen Grundbildung erwartbaren Ergebnisse und sprechen dafür, dass der Test valide ist und es ermöglicht, zwischen unterschiedlichen Leistungsniveaus zu differenzieren.
  - Die DIF-Analysen weisen, betrachtet man die nach Geschlecht aufgeteilte Stichprobe, keine gravierenden Auffälligkeiten auf.
  - Die Ergebnisse der Modellprüfungen geben die theoretische Struktur der prozessbezogenen naturwissenschaftlichen Grundbildung wieder.
- *Externe Validierung*:
  - Die Korrelation zwischen Kompetenz und *Naturwissenschaftsnoten* (Korrelation zwischen zwei Leistungskonstrukten) fallen höher aus als die Korrelation zwischen Kompetenz und *naturwissenschaftlichem Fachinteresse* (Korrelation zwischen

einem Leistungs- und einem Motivationskonstrukt). Dieses Ergebnis spricht dafür, dass der als Leistungstest angelegte Test die Leistung und nicht das Interesse der Schülerinnen und Schüler misst.

– Die Korrelation der Kompetenz mit dem *Interesse an naturwissenschaftlichen Arbeitsweisen* fällt signifikant höher aus als die Korrelation der Kompetenz mit dem *Interesse an naturwissenschaftlichen Aktivitäten*. Dieses Ergebnis wird als Anzeichen für die inhaltliche Nähe der Inhalte des Kompetenztests und der Interessenskala und damit als Zeichen von Validität gewertet.

Als offene Punkte der Testvalidierung müssen mit Blick auf die interne Validierung der Umgang mit dem *DIF* zwischen den Schulniveaustichproben sowie in Bezug auf die externe Validierung die Suche nach einem geeigneten externen Kriterium gesehen werden.

Die DIF-Analysen zeigten im Vergleich der Schulniveaus Auffälligkeiten, die darauf hindeuten, dass in die Erfassung der Kompetenz noch weitere Einflüsse eingehen. Erste Vorschläge zum Umgang mit diesem Ergebnis wurden gemacht. Das Testinstrument kann relativ unkompliziert eingesetzt werden, wenn ein Vergleich der Schulniveaus nicht angestrebt ist. Da jedoch dieser Vergleich keine besonders abwegige Absicht darstellt, sollte sich die unbedingt notwendige Kreuzvalidierung einer genaueren Prüfung der *DIFs* widmen. Wie bereits im Abschnitt 6.1.3 erwähnt, gilt es allerdings zunächst zu prüfen, ob es sich hier um zufällig auftretende DIF-Effekte handelt oder ob sie systematischer Natur sind.

In den Rahmen einer Kreuzvalidierung gehört weiterhin das Heranziehen eines Außenkriteriums, das besser zur Validierung geeignet ist als die Schulnoten. Interessant wären hier Performanztests, die zusätzlich zum Kompetenztest mit einem Teil der Stichprobe durchgeführt werden könnten, oder aber die in Abschnitt 2.4.4 beschriebenen Portfolios. Ein Performanztest könnte so aufgebaut sein, dass seine Aufgaben strukturell und hinsichtlich der angesprochenen kognitiven Dimensionen den Testaufgaben ähneln, aber offen formuliert und mit praktisch gestaltbaren Elementen versehen sind. Portfolios könnten genutzt werden, indem ein Forschungsprojekt dokumentiert wird, das einen kompletten wissenschaftlichen Forschungsprozess umfasst. Beide Methoden erfordern exakte Auswertungskriterien, um als gute Validierungskriterien fungieren zu können. Dennoch ist fraglich, inwieweit die im Vergleich zu Schulnoten inhaltlich größere Nähe, aber auch die größere Komplexität dieser Kriterien sowie die besonderen Erhebungsmethoden eine angemessenere Art der Validierung darstellen. Es ist unter Berücksichtigung dieser Punkte und in Anlehnung an bestehende Forschungsergebnisse zu Zusammenhängen zwischen Multiple-Choice-

Tests und Performanztests eine gewisse Bandbreite an Korrelationen zu erwarten: Berichtet werden Korrelationen zwischen $0,3$ und $0,6$ (Shavelson & Ruiz-Primo, 1999; Ayala, Shavelson, Yin & Schultz, 2002). Dennoch wäre es hilfreich, sich der Validierung auf diese Weise zu nähern, da die naturwissenschaftlichen Fähigkeiten, die dieser Test misst, als prozessbezogen definiert wurden und somit in Verbindung mit der Messung ihrer praktischen Umsetzung stehen sollten.

Die genannten Ansätze für eine weitergehende Validierung können lediglich als Ausblick dienen. Die ersten, in diesem Rahmen möglichen, Validierungsschritte sind gemacht und müssen nun an einer weiteren Stichprobe mit den genannten Erweiterungen fortgesetzt werden. Dies ist insbesondere deshalb wichtig, da Haupttest und erste Validierung hier anhand der gleichen Stichprobe durchgeführt wurden. Obwohl Validierungen aufgrund ihres Prozesscharakters nie abgeschlossen sind, würde bereits eine einfache Kreuzvalidierung anhand einer neuen und strukturell gleich zusammengesetzten Stichprobe, die ähnliche Ergebnisse zeigt, die Aussagekraft des Instruments stützen.

6.1.4 BEDEUTUNG DER RELIABILITÄTSPRÜFUNG

Bezüglich der Reliabilitätsprüfung wurde darauf geachtet, dass zum einen die probabilistische WLE-Reliabilität und zum anderen die klassische Reliabilität (Cronbachs Alpha) berichtet wird. Die WLE-Reliabilität folgt aus der probabilistischen Testentwicklung und die Darstellung von Cronbachs Alpha dient der Vergleichbarkeit mit bereits bestehenden Testverfahren, die sich mit ähnlichen Inhalten beschäftigen.

Die in Abschnitt 3.2.2 dargestellten Mindestwerte, die der Test zur Erfassung prozessbezogener naturwissenschaftlicher Grundbildung erreichen sollte, wurden sowohl im Falle der WLE-Reliabilität ($0,77$) als auch im Falle von Cronbachs Alpha ($0,81$) erreicht. Es wurde im Methodenteil (Abschnitt 3.2.2) argumentiert, dass der als ökonomisches Verfahren für Gruppen-Screenings angelegte *Test zur Erfassung prozessbezogener Grundbildung* eine Reliabilität erreichen sollte, die unter der von Intelligenz- und Individualdiagnostik, jedoch über der von Persönlichkeitstests liegen sollte. Es sollten also Werte zwischen $0,70$ und $0,90$ erreicht werden. Dies ist hier deutlich der Fall. Der Test erreicht eine angemessene Reliabilität.

Auffällig ist, dass die WLE-Reliabilität leicht unter der Reliabilität nach Cronbachs Alpha liegt. Dieser Unterschied ist dadurch begründet, dass die WLE-Reliabilität sensibler für den Einfluss von Messfehlern ist als Cronbachs Alpha. Dieser Umstand ist letzten Endes auf die Schätzung der WLE-Reliabilität durch die Andrich-Methode zurückzuführen (vgl. Abschnitt 5.5.2), die nur asymptotisch genau schätzt, wodurch es

zu einer Überschätzung der Messfehlervarianzen kommt. Diese Überschätzung spielt insbesondere bei einer Testlänge < 20 Items eine Rolle. Je geringer die Itemanzahl, desto größer die Überschätzung der Fehlervarianz und die Unterschätzung der WLE-Reliabilität. Da die Endversion des *Tests zur Erfassung prozessbezogener naturwissenschaftlicher Grundbildung* 21 Items umfasst, sollte die Überschätzung der Messfehlervarianz allerdings allenfalls gering ausfallen. Einen vertiefenden Blick auf Reliabilitäts- und Personenparameterschätzungen im Zusammenhang mit der Testlänge bietet Walter (2005).

6.2 DAS INSTRUMENT IM VERGLEICH ZU BISHER ENTWICKELTEN INSTRUMENTEN

Zum Abschluss der Diskussion wird eine Verbindung zum Beginn der Arbeit hergestellt, indem der *Test zur Erfassung prozessbezogener naturwissenschaftlicher Grundbildung* anhand seiner Kennwerte zu den dort genannten Multiple-Choice-Tests in Beziehung gesetzt wird. Hier steht ein Vergleich mit den beiden deutschsprachigen Verfahren im Vordergrund, dem Kompetenztest von Phan (Phan, 2007) und dem NAW-Test von Klos et al. (Klos et al., 2008).

Was den im Rahmen dieser Arbeit entwickelten Test von den in Abschnitt 2.4.5 (S. 57) dargestellten Multiple-Choice-Tests zunächst ganz allgemein unterscheidet ist die Tatsache, dass er auf Grundlage der probabilistischen Testtheorie entwickelt wurde. Anders als bei Testverfahren, die nach der klassischen Testtheorie entwickelt werden, gibt es eine theoretische Vorstellung über die Verbindung zwischen manifestem Testverhalten und latenter Kompetenz. Der Vorteil liegt hier insbesondere darin, dass anhand der probabilistischen Testtheorie geprüft werden kann, inwiefern es gerechtfertigt ist, die Reaktionen auf einzelne Testitems zusammenzufassen. Es ist in diesem Rahmen empirisch prüfbar, inwiefern die Testitems das gleiche Merkmal messen und inwiefern die Modellparameter stichprobenunabhängig sind. Diese Punkte stellen für die Validierung des Testinstruments einen großen Vorteil gegenüber klassisch konstruierten Verfahren dar und können als Qualitätsmerkmal betrachtet werden. In der genannten Validierung liegt ein weiterer allgemeiner Vorteil des neu entwickelten Verfahrens gegenüber den älteren Verfahren: Zur Prüfung des neuen Verfahrens wurden erste Schritte der Validierung unternommen, während für die meisten älteren Verfahren zwar eine Reliabilitätsprüfung durchgeführt wurde, es aber oft keine Validierung gab oder aber von keiner berichtet wurde.

Im speziellen Vergleich der Qualitätsmaße fällt auf, dass der *Test zur Erfassung pro-*

zessbezogener naturwissenschaftlicher Grundbildung mit seiner Reliabilität (Cronbachs Alpha) von $0,81$ gut abschneidet. Gleiches kann für die mittlere Trennschärfe des Tests von $0,45$ gesagt werden. Im Vergleich zu den beiden deutschsprachigen Verfahren ist zu bemerken, dass der neu entwickelte Test über dem Cronbachs Alpha des Kompetenztests von Phan (*r=0,77*), über dem NAW-Test für Schülerinnen und Schüler der siebten Klasse (*r=0,73*) und leicht unter dem NAW-Test für Schülerinnen und Schüler der zwölften Jahrgangsstufe (*r=0,82*) liegt. Ein weiterer Vorteil gegenüber dem Kompetenztest von Phan liegt in der durchgeführten Validierung und gegenüber beiden deutschsprachigen Verfahren in der Unabhängigkeit der Aufgabenantworten.

Das grundlegende Merkmal, anhand dessen sich der vorliegende Test von bestehenden Verfahren unterscheidet, ist der umfangreiche Entwicklungsprozess, der aktuellen Standards der Item- und Testentwicklung entspricht. Zunächst wurde ein Kompetenzmodell erstellt und zu messende Fertigkeiten definiert, die theoretisch begründet als Indikatoren für die Ausprägung der *prozessbezogenen naturwissenschaftlichen Grundbildung* herangezogen wurden. Auf messtheoretischer Ebene wurde mit dem eindimensionalen Rasch-Modell ein probabilistisches Testmodell zu Grunde gelegt, das eine Verbindung zwischen manifestem Testverhalten und latenter Kompetenz ermöglicht. In drei Testphasen wurden die entwickelten Items an insgesamt über 1300 Schülerinnen und Schülern getestet. Dabei kamen nicht nur quantitative Verfahren, sondern auch qualitative Verfahren wie *Cognitive-Lab-Interviews* und *Expertenpanel* zum Einsatz, jeweils angepasst an die unterschiedlichen Testphasen. Die Ergebnisse dieser Tests wurden in die Überarbeitung und Entfernung von Testitems umgesetzt und mündeten schließlich in einer ersten Validierung, welche die Qualität des Testverfahrens bestätigen konnte. Ökonomische Durchführung, Auswertung und Interpretation des Testverfahrens machen einen weiteren Teil der Qualität dieses Verfahrens aus.

Setzt man abschließend das Testverfahren in Beziehung mit der Erfassung der naturwissenschaftlichen Grundbildung bei PISA, so zeigt sich eine Parallele in den Korrelationen der drei Fähigkeitsskalen, die mit Werten $> 0,90$ ebenso hoch ausfallen wie die Korrelationen zwischen den PISA-Teilkompetenzen. Ein Unterschied zu PISA besteht unter anderem darin, dass die WLE-Reliabilität des Tests mit $0,77$ in jedem Fall deutlich unter der WLE-Reliabilität der naturwissenschaftlichen Kompetenz bei PISA von $0,83$ (OECD, 2009) liegt, was sich unter anderem durch die größere Itemanzahl erklären lässt. Inhaltlich lässt sich das vorliegende Verfahren insofern abgrenzen, als es sich auf einen speziellen, nämlich den prozessbezogenen Anteil naturwissenschaf-

tlicher Grundbildung, konzentriert, der sich gemäß wissenschaftlicher Befunde als besonders bedeutsam und in Bezug auf weitere Aspekte naturwissenschaftlicher Bildung als einflussreich erwiesen hat (vgl. Abschnitt 2.2.2).

Die Entwicklung eines *Multiple-Choice-Tests zur Erfassung prozessbezogener naturwissenschaftlicher Grundbildung* kann insgesamt als gelungen bezeichnet werden. Die Ergebnisse der Testentwicklung und der geschilderte Vergleich mit bestehenden Testverfahren machen seine besonderen Qualitäten deutlich. Erste Ansätze, in welchen Bereichen diese Qualitäten in Zukunft genutzt werden können und auf welche Weise die praktische Anwendbarkeit des Tests geprüft werden kann, werden im nachfolgenden Abschnitt erläutert.

7 Ausblick

Der *Test zur Erfassung prozessbezogener naturwissenschaftlicher Grundbildung* wurde vor dem Hintergrund entwickelt, dass es bisher vor allem im deutschsprachigen Raum und außerhalb der PISA-Studie an Verfahren mangelte, die eine Erfassung naturwissenschaftlicher Grundbildung in ökonomischer und valider Weise ermöglichen. Dabei wäre es in unterschiedlichen Bildungskontexten, schulisch wie auch außerschulisch, interessant, ein Verfahren für ein schnelles Screening der *prozessbezogenen naturwissenschaftlichen Grundbildung* nutzen zu können.

Im Rahmen des Theorieteils dieser Arbeit wurde das Multiple-Choice-Verfahren für den beschriebenen Zweck als das Geeignete identifiziert und in Anlehnung an aktuelle *Scientific-Literacy*-Konzeptionen wie die aus *PISA* (OECD, 2006), *21st Century Science* (University of York Science Education Group, 2006) und dem *Project 2061* (American Association of the Advancement of Science, 2001) auf der Grundlage der probabilistischen Testtheorie entwickelt. Vor diesem Hintergrund kann das entstandene Testverfahren als zeitgemäßes und vor dem Hintergrund der Ergebnisse als gelungenes Verfahren bezeichnet werden.

Um die Validität des Verfahrens abzusichern, steht noch eine Kreuzvalidierung anhand einer neuen Stichprobe aus, die unbedingt durchgeführt werden sollte, bevor der Test genutzt werden kann, um Aussagen über die *prozessbezogene naturwissenschaftliche Grundbildung* von Schülerinnen und Schülern der neunten Klasse zu machen. Dies ist dann möglich, wenn sich die ermittelten Itemparameter anhand einer Kreuzvalidierung replizieren lassen. Im Zuge der Validierung sollte zur Prüfung der externen Validität über den Einsatz der vorgeschlagenen Verfahren Performanztest bzw. Portfolios nachgedacht werden, die von einem Teil der Stichprobe bearbeitet werden könnten. Wenn die Validierung noch in Richtung der Grundlagenforschung weitergedacht werden soll, so würden sich hier auch Möglichkeiten zur weiteren Erforschung von DIF-Effekten ergeben. Da anhand der neuen Stichprobe ohnehin geprüft werden sollte, ob die identifizierten DIF-Items sich erneut zeigen, könnte diese Gelegenheit genutzt werden, um im Sinne des in Abschnitt 6.1.3 vorgeschlagenen Vorgehens DIF-Ursachen systematisch zu untersuchen.

Einen weiteren offenen Punkt stellt die Festlegung von Kompetenzstufen dar, um das Niveau, das die Schülerinnen und Schüler erreichen, auch inhaltlich beschreiben

zu können. Wie bereits in Abschnitt 3.1.3 ausgeführt, sollte diese Festlegung in pragmatischer Weise durch die Unterteilung des Fähigkeitskontinuums in maximal drei Niveaustufen erfolgen, um noch genügend Items zur Beschreibung des jeweiligen Niveaus zur Verfügung zu haben. Offen bleibt noch, ob dieses Ziel vor dem Hintergrund der in den Randbereichen in geringerer Zahl vorhandenen Items möglich ist.

Betrachtet man die weiteren Verwendungsmöglichkeiten des Instruments, so bieten sich sowohl schulische als auch außerschulische Anwendungsbereiche an. Obwohl das entwickelte Verfahren nicht zur Interventionsmessung, sondern vielmehr mit der Absicht eines schnellen und ökonomischen Screenings entwickelt wurde, besteht die Möglichkeit, es mit einem ausreichenden Abstand zwischen Prä- und Postmessung und in einem Experimentaldesign mit Kontrollgruppe zur Messung von Interventionseffekten einzusetzen. Dies könnte für langfristige, auf den naturwissenschaftlichen Prozess bezogene, Unterrichtsprojekte oder für langfristige Kooperationen zwischen Schulen und naturwissenschaftlich orientierten außerschulischen Lernorten eine Option darstellen. Die optimale Variante würde hier sicher darin bestehen, zwei parallele Testversionen für die beiden Messzeitpunkte zu besitzen, doch dafür reicht die vorliegende Itemanzahl nicht aus.

Die aktuell interessanteste Möglichkeit des Aufgabeneinsatzes stellt das Nationale Bildungspanel (NEPS) dar, das in Form eines Längsschnittes über unterschiedliche Kohorten Bildungsprozesse und -verläufe sowie die Entfaltung von Kompetenzen über die Lebensspanne verfolgt. Teil der in diesem Zusammenhang erhobenen Kompetenzen ist unter anderem das Wissen über die Naturwissenschaften, also das auf den Prozess naturwissenschaftlicher Erkenntnisgewinnung bezogene Wissen. Da sich in der Konzeptionalisierung dieses Bereichs Überschneidungen mit der Konzeptionalisierung der vorliegenden Testentwicklung zeigen, ergibt sich die Möglichkeit, die entwickelten Items in bestehender oder leicht abgewandelter Form in den Itementwicklungsprozess des NEPS für Schülerinnen und Schüler der neunten Klasse (im Alter von 15-16 Jahren) einfließen zu lassen. Im Falle des Einsatzes unveränderter Items würde sich hier zusätzlich die Möglichkeit der Validierung einzelner Items ergeben, da alle Items, die Teil der Bildungspanelerhebungen sind, noch unterschiedliche Pilotierungen durchlaufen.

Zusammenfassend bleibt festzustellen, dass es für den *Test zur Erfassung prozessbezogener naturwissenschaftlicher Grundbildung* als Ganzes oder in Form einzelner Itemsets für die Zukunft unterschiedliche Anwendungsfelder geben kann. Gerade die Aktualität des Verfahrens und die inhaltliche Nähe zu Kompetenzkonzeptionen des Nationalen Bildungspanels stellen Stärken dar, die sich sehr gut nutzen lassen.

Anhang A

Das Testinstrument

A.1 Testmanual

A.2 Testheft

Das Testheft ist auf Anfrage erhältlich. Bei Interesse wenden Sie sich bitte an:

Inga Glug
Leibniz-Institut für die Pädagigik der Naturwissenschaften (IPN)
Olshausenstraße 62
24098 Kiel

Tel.: 0431 / 880-3149
Mail: glug@ipn.uni-kiel.de

Testmanual zur Durchführung des *Tests zur Erfassung prozessbezogener naturwissenschaftlicher Grundbildung*

Vorstellen: Grund des Besuchs und Zweck der Untersuchung erklären

- Vorstellen: Mein Name ist Inga Glug, ich arbeite an der Universität Kiel und ich bin heute hier, weil ich eure Mithilfe benötige.

- Grund des Besuchs: Im Rahmen meiner Doktorarbeit habe ich Aufgaben zum naturwissenschaftlichen Forschen entwickelt. Anhand dieser Aufgaben soll herausgefunden werden, ob Schülerinnen und Schüler der 9. Klasse eine Idee davon haben, wie naturwissenschaftliche Forschung abläuft, also welche Fragen sich Wissenschaftlerinnen und Wissenschaftler stellen und wie sie diese beantworten.

- Zweck der Untersuchung: Da sich die Aufgaben noch in der Entwicklungsphase befinden und noch nicht klar ist, ob sie für Schülerinnen und Schüler der 9. Klasse geeignet sind, müssen sie geprüft werden.

- Aus diesem Grund benötige ich die Mithilfe von Schülerinnen und Schülern. Es geht also nicht darum, dass ihr getestet werden sollt, sondern darum zu schauen, ob die Aufgaben zu leicht, zu schwer oder vielleicht schon ganz gut sind. Ohne eure Hilfe kann diese Aufgabenentwicklung und damit auch meine Doktorarbeit nicht erfolgreich sein.

- Deshalb ist eine sorgfältige Bearbeitung der Aufgaben sehr wichtig!

- Wichtig: Es ist kein Test, der für die Schulleistung zählt. Die Lehrerinnen und Lehrer werden die bearbeiteten Aufgaben nicht zu sehen bekommen und die Bearbeitung erfolgt vollkommen anonym. Die Ergebnisse können später nicht mehr zu einzelnen Personen zurückverfolgt werden. Allerdings wird es für jede Schulklasse eine Auswertung geben, anhand derer sie sich mit den Parallelklassen vergleichen kann.

Aufgabenhefte austeilen und zunächst den Validierungsbogen bearbeiten lassen

- Den Validierungsbogen mit den Schülerinnen und Schülern durchgehen:

    - Zunächst die allgemeinen Daten: Schule, Klasse, Geschlecht
    - Danach folgt die Einschätzung des Interesses für die einzelnen Schulfächer und die Angabe der Schulnoten des letzten Halbjahreszeugnisses.
    - Bei der Einschätzung des Interesses an naturwissenschaftlichen Tätigkeiten und naturwissenschaftsbezogenen Interessen darauf hinweisen, dass die Schülerinnen und Schüler pro Zeile nur ein Kreuz machen dürfen
    - Die Schülerinnen und Schüler sollen noch nicht umblättern, wenn sie mit der Bearbeitung des Validierungsbogens fertig sind, sondern den Stift hinlegen und nach vorne schauen

Aufgabenbearbeitung erklären

- Die ersten beiden Seiten des Testheftes inklusive der Beispielaufgabe durchgehen.

- nach den Erklärungen und unmittelbar bevor sie die Bearbeitung der Aufgaben beginnen, müssen die Schülerinnen und Schüler noch darüber informiert werden, dass sie nach Fertigstellung der Aufgaben...

    - noch einmal kontrollieren sollen, ob sie wirklich alle Aufgaben beantwortet haben
    - in der Klasse bleiben und sich ruhig beschäftigen sollen (evtl. den Lehrer/ die Lehrerin bitten, eine kleine Aufgabe vorzubereiten).

- Nachfragen, ob die Schülerinnen und Schüler noch Fragen haben. Sie können zwischendurch Fragen zu Wörtern stellen, die sie nicht verstehen. Bei der Lösung der Aufgaben wird nicht geholfen.

- Anfangen, wenn es keine Fragen mehr gibt und alle Formalitäten geklärt sind.

Beachten

- Die Schülerinnen und Schüler können keinen Fragebogen mitnehmen, da das Ganze ja Teil der Doktorarbeit ist und vorher nicht in Umlauf gebracht werden darf.
- Die Lehrerinnen und Lehrer können gern in den Fragebogen reinschauen, sollen ihn dann aber nach ausreichendem „Studium“ wieder zurückgeben.

Anhang B

Interesse an Tätigkeiten, die im Physikunterricht vorkommen

Im folgenden findest du einige Tätigkeiten, die auch im Physikunterricht vorkommen. Gib bitte an,

a) wie groß dein Interesse ist, im Physikunterricht überhaupt oder bei einzelnen Themen eine bestimmte Tätigkeit allein oder zusammen mit anderen auszuüben und,
b) wie oft in den letzten Monaten im Unterricht Gelegenheit dazu war.

1. Beobachten, wie der Lehrer oder andere Schüler einen physikalischen Versuch durchführen

a) mein Interesse daran?	b) wie oft im Unterricht?
☐ sehr groß	☐ sehr oft
☐ groß	☐ oft
☐ mittel	☐ manchmal
☐ gering	☐ selten
☐ sehr gering	☐ nie

2. einen Physiktext lesen

a) mein Interesse daran?	b) wie oft im Unterricht?
☐ sehr groß	☐ sehr oft
☐ groß	☐ oft
☐ mittel	☐ manchmal
☐ gering	☐ selten
☐ sehr gering	☐ nie

3. einem Vortrag über Physik (Lehrer oder Schüler) zuhören

a) mein Interesse daran?	b) wie oft im Unterricht?
☐ sehr groß	☐ sehr oft
☐ groß	☐ oft
☐ mittel	☐ manchmal
☐ gering	☐ selten
☐ sehr gering	☐ nie

---

4. etwas bauen, einen Versuch aufbauen oder ein Gerät konstruieren

a) mein Interesse daran?	b) wie oft im Unterricht?
☐ sehr groß	☐ sehr oft
☐ groß	☐ oft
☐ mittel	☐ manchmal
☐ gering	☐ selten
☐ sehr gering	☐ nie

---

5. einen Versuch selber durchführen, Messungen machen

a) mein Interesse daran?	b) wie oft im Unterricht?
☐ sehr groß	☐ sehr oft
☐ groß	☐ oft
☐ mittel	☐ manchmal
☐ gering	☐ selten
☐ sehr gering	☐ nie

6. etwas ausprobieren, ein Gerät auseinandernehmen oder zusammensetzen

a) mein Interesse daran?	b) wie oft im Unterricht?
☐ sehr groß	☐ sehr oft
☐ groß	☐ oft
☐ mittel	☐ manchmal
☐ gering	☐ selten
☐ sehr gering	☐ nie

---

7. sich ausdenken, wie man eine bestimmte Vermutung durch einen Versuch prüfen könnte

a) mein Interesse daran?	b) wie oft im Unterricht?
☐ sehr groß	☐ sehr oft
☐ groß	☐ oft
☐ mittel	☐ manchmal
☐ gering	☐ selten
☐ sehr gering	☐ nie

---

8. etwas berechnen, den Ausgang eines Versuchs exakt vorhersagen, Aufgaben lösen

a) mein Interesse daran?	b) wie oft im Unterricht?
☐ sehr groß	☐ sehr oft
☐ groß	☐ oft
☐ mittel	☐ manchmal
☐ gering	☐ selten
☐ sehr gering	☐ nie

9. etwas erfinden, sich ein bestimmtes Gerät ausdenken

a) mein Interesse daran?

- ☐ sehr groß
- ☐ groß
- ☐ mittel
- ☐ gering
- ☐ sehr gering

b) wie oft im Unterricht?

- ☐ sehr oft
- ☐ oft
- ☐ manchmal
- ☐ selten
- ☐ nie

---

10. mit anderen über eine bestimmte technische Neuerung diskutieren

a) mein Interesse daran?

- ☐ sehr groß
- ☐ groß
- ☐ mittel
- ☐ gering
- ☐ sehr gering

b) wie oft im Unterricht?

- ☐ sehr oft
- ☐ oft
- ☐ manchmal
- ☐ selten
- ☐ nie

---

11. sich eine eigene Meinung zu Fragen aus Physik und Technik bilden

a) mein Interesse daran?

- ☐ sehr groß
- ☐ groß
- ☐ mittel
- ☐ gering
- ☐ sehr gering

b) wie oft im Unterricht?

- ☐ sehr oft
- ☐ oft
- ☐ manchmal
- ☐ selten
- ☐ nie

12. den Wert oder Nutzen einer physikalisch-technischen Neuerung beurteilen

a) mein Interesse daran?

- ☐ sehr groß
- ☐ groß
- ☐ mittel
- ☐ gering
- ☐ sehr gering

b) wie oft im Unterricht?

- ☐ sehr oft
- ☐ oft
- ☐ manchmal
- ☐ selten
- ☐ nie

Anhang C

Leitfaden zur Aufgabenbeurteilung

Leitfaden zur Itembeurteilung

Überprüft die vorliegende Aufgabe Ihrer Meinung nach das „Identifizieren wissenschaftlicher Hypothesen“, „Planen einer wissenschaftlichen Untersuchung“ oder das „Nutzen wissenschaftlicher Nachweise“? (bitte nur eine Alternative ankreuzen):

Identifizieren wissenschaftlicher Hypothesen:
unter Berücksichtigung gegebener Hintergrundinformationen wissenschaftlich prüfbare Fragen erkennen ☐

Planen einer wissenschaftlichen Untersuchung:
unter Berücksichtigung gegebener Hintergrundinformationen bezüglich einer zu prüfenden Annahme erkennen, welche Variablen zu variieren und welche konstant zu halten sind ☐

Nutzen wissenschaftlicher Nachweise:
unter Berücksichtigung gegebener Hintergrundinformationen zur untersuchten Vermutung Schlussfolgerungen aus erhobenen Daten bzw. Daten-Graphiken ziehen ☐

Welcher Fachwissenschaft kann die Aufgabe am ehesten zugeordnet werden? (nur ein Kreuz)

Biologie ☐ Chemie ☐ Physik ☐

Szenario

Die Komplexität des Szenarios ist für 15- bis 16-Jährige angemessen.	Ja / Nein
Das Szenario ist für 15- bis 16-Jährige interessant.	Ja / Nein
Das Szenario ist für 15- bis 16-Jährige verständlich formuliert.	Ja / Nein
Die im Szenario dargestellten Inhalte sind sachlich richtig.	Ja / Nein

---

Aufgaben

Die Aufgaben-Schwierigkeit ist für 15- bis 16-Jährige angemessen.	Ja / Nein
Die Aufgabe ist für 15- bis 16-Jährige interessant.	Ja / Nein
Die Aufgabe ist für 15- bis 16-Jährige verständlich formuliert.	Ja / Nein
Die in der Aufgabe dargestellten Inhalte sind sachlich richtig.	Ja / Nein

---

Allgemeine Einschätzung

Das Material (der Text, die Tabelle o.ä.) und die dazugehörige Aufgabe bilden eine in sich geschlossene, sinnvolle Einheit.	Ja / Nein

Abbildung C.1: Experten-Leitfaden zur Itembeurteilung

Anhang D

Ergebnisse

D.1 Feldtest

D.1.1 Analyse der Antwortalternativen

D.1.2 Bewertung der Feldtest-Items

Tabelle D.1: Analyse der Antwortalternativen (Teil 1)

Item	Label	Score	Count	% of tot	Pt Bis	PV1Avg:1
Bro1	a	1.00	127	40.32	0.45	-0.02
	b	0.00	36	11.43	-0.15	-0.58
	c	0.00	142	45.08	-0.31	-0.53
	d	0.00	6	1.90	-0.09	-0.48
	9	0.00	4	1.27	-0.07	-0.98
Bro2	a	0.00	32	10.16	-0.11	-0.56
	b	0.00	117	37.14	-0.06	-0.36
	c	0.00	72	22.86	-0.16	-0.51
	d	1.00	90	28.57	0.30	-0.08
	9	0.00	4	1.27	-0.05	-0.68
Bro3	a	0.00	37	11.75	-0.22	-0.63
	b	0.00	56	17.78	-0.29	-0.76
	c	0.00	24	7.62	-0.23	-0.85
	d	1.00	193	61.27	0.50	-0.10
	9	0.00	5	1.59	-0.02	-0.36
Hur1	a	0.00	100	31.75	-0.15	-0.45
	b	0.00	50	15.87	-0.18	-0.36
	c	1.00	115	36.51	0.19	-0.18
	d	0.00	47	14.92	0.13	-0.20
	9	0.00	3	0.95	-0.06	-0.37
Hur2	a	0.00	26	8.25	-0.12	-0.66
	b	1.00	202	64.13	0,46	-0.14
	c	0.00	41	13.02	-0.29	-0.77
	d	0.00	39	12.38	-0.25	-0.65
	9	0.00	7	2.22	-0.05	-0.62
Hur3	a	0.00	50	15.87	-0.21	-0.64
	b	1.00	143	45.40	0.40	-0.08
	c	0.00	64	20.32	-0.16	-0.53
	d	0.00	51	16.19	-0.10	-0.50
	9	0.00	7	2.22	-0.13	-0.69
Sch1	a	0.00	76	24.13	-0.29	-0.67
	b	0.00	61	19.37	-0.19	-0.59
	c	1.00	132	42.90	0.53	0.07
	d	0.00	44	13.97	-0.16	-0.63
	9	0.00	2	0.63	-0.08	-0.86
Sch2	a	0.00	106	33.65	-0.10	-0.37
	b	0.00	7	2.22	-0.10	-0.60
	c	0.00	47	14.92	-0.16	-0.58
	d	1.00	142	45.08	0.29	-0.18
	9	0.00	13	4.13	-0.12	-0.81
Sch3	a	0.00	68	21.59	-0.24	-0.65
	b	0.00	29	9.21	-0.21	-0.80
	c	0.00	25	7.94	-0.28	-0.95
	d	1.00	189	60.00	0.48	-0.08
	9	0.00	4	1.27	0.01	-0.17
Sti1	a	0.00	68	21.59	-0.17	-0.53
	b	1.00	181	57.46	0.36	-0.19
	c	0.00	29	9.21	-0.24	-0.75
	d	0.00	32	10.16	-0.09	-0.38
	9	0.00	5	1.59	-0.08	-0.62
Sti2	a	0.00	27	8.57	-0.13	-0.46
	b	1.00	130	41.27	0.18	-0.25
	c	0.00	30	9.52	-0.01	-0.29
	d	0.00	123	39.05	-0.01	-0.43
	9	0.00	5	1.59	-0.11	-0.09
Sti3	a	1.00	121	38.41	0.35	-0.08
	b	0.00	105	33.33	-0.15	-0.49
	c	0.00	27	8.57	-0.13	-0.61
	d	0.00	56	17.78	-0.14	-0.49
	9	0.00	6	1.90	-0.06	-0.39

Tabelle D.2: Analyse der Antwortalternativen (Teil 2)

Item	Label	Score	Count	% of tot	Pt Bis	PV1Avg:1
KLe1	a	1.00	110	34.92	0.32	-0.11
	b	0.00	61	19.37	-0.09	-0.41
	c	0.00	77	24.44	-0.20	-0.56
	d	0.00	60	19.05	-0.06	-0.41
	9	0.00	7	2.22	-0.05	-0.36
KLe2	a	0.00	169	53.65	0.29	-0.17
	b	0.00	67	21.27	-0.25	-0.66
	c	0.00	46	14.60	-0.20	-0.65
	d	1.00	26	8.25	0.11	-0.05
	9	0.00	7	2.22	-0.01	-0.35
KLe3	a	0.00	103	32.70	0.11	-0.22
	b	0.00	64	20.32	-0.18	-0.61
	c	1.00	113	35.87	0.14	-0.22
	d	0.00	25	7.94	-0.14	-0.61
	9	0.00	10	3.17	-0.03	-0.53
Kli1	a	0.00	63	20.00	-0.07	-0.41
	b	0.00	63	20.00	-0.06	-0.38
	c	0.00	76	24.13	-0.22	-0.53
	d	1.00	108	34.29	0.33	-0.12
	9	0.00	5	1.59	-0.11	-0.84
Kli2	a	1.00	131	41.59	0.31	-0.12
	b	0.00	67	21.27	-0.11	-0.50
	c	0.00	66	20.95	-0.14	-0.50
	d	0.00	44	13.97	-0.14	-0.51
	9	0.00	7	2.22	-0.02	-0.39
Kli3	a	0.00	55	17.46	-0.17	-0.50
	b	0.00	74	23.49	-0.13	-0.50
	c	1.00	120	38.10	0.30	-0.11
	d	0.00	62	19.68	-0.04	-0,42
	9	0.00	4	1.27	-0.06	-0.79
Reg1	a	0.00	82	26.03	-0.11	-0.46
	b	0.00	43	13.65	-0.22	-0.59
	c	0.00	38	12.06	-0.13	-0.53
	d	1.00	149	47.30	0.34	-0.15
	9	0.00	3	0.95	-0.03	-0.91
Reg2	a	1.00	78	24.76	0.22	-0.09
	b	0.00	54	17.14	-0,24	-0.65
	c	0.00	53	16.38	-0.29	-0.75
	d	0.00	123	39.05	0.24	-0.17
	9	0.00	7	2.22	-0.09	-0.60
Reg3	a	0.00	63	20.00	-0,28	-0.64
	b	1.00	124	39.37	0.25	-0.18
	c	0.00	78	24.76	-0.02	-0.37
	d	0.00	45	14.29	0.02	-0.33
	9	0.00	5	1.59	-0.03	-0.22
Ros1	a	0.00	40	12.70	-0.24	-0.64
	b	0.00	35	11.11	-0.21	-0.74
	c	1.00	200	63.49	0.48	-0.11
	d	0.00	30	9.52	-0.26	-0.89
	9	0.00	10	3.17	-0.07	-0.66
Ros2	a	0.00	58	18.41	-0.16	-0.50
	b	0.00	42	13.33	-0,20	-0.74
	c	0.00	87	27.62	-0.21	-0.55
	d	1.00	124	39.37	0.47	0.02
	9	0.00	4	1.27	-0.05	-0.55
Ros3	a	0.00	39	12.38	-0.13	-0.57
	b	1.00	174	55.24	0.42	-0.09
	c	0.00	73	23.17	-0.33	-0.72
	d	0.00	17	5.40	-0.03	-0.33
	9	0.00	12	3.81	-0.10	-0.72

Tabelle D.3: Analyse der Antwortalternativen (Teil 3)

Item	Label	Score	Count	% of tot	Pt Bis	PV1Avg:1
Sko1	a	1.00	165	52.38	0.34	-0.15
	b	0.00	43	13.65	-0.21	-0.56
	c	0.00	20	6.35	-0.15	-0.88
	d	0.00	83	26.35	-0.12	-0.45
	9	0.00	4	1.27	-0.06	-0.83
Sko2	a	0.00	24	7.62	-0.25	-0.91
	b	0.00	40	12.70	-0.23	-0.60
	c	1.00	204	64.76	0.49	-0.12
	d	0.00	44	13.97	-0.27	-0.79
	9	0.00	3	0.95	-0.01	-0.89
Sko3	a	0.00	145	46.03	-0.10	-0.42
	b	0.00	36	11.43	-0.20	-0.65
	c	1.00	108	34.29	0.35	-0.01
	d	0.00	18	5.71	-0.16	-0.77
	9	0.00	8	2.54	-0.11	-0.89
Sol1	a	1.00	176	55.87	0.34	-0.19
	b	0.00	31	9.84	-0.21	-0.73
	c	0.00	71	22.54	-0.25	-0.55
	d	0.00	36	11.43	0.01	-0.17
	9	0.00	1	0.32	-0.10	-1.19
Sol2	a	1.00	129	40.95	0.46	-0.02
	b	0.00	72	22.86	-0.09	-0.40
	c	0.00	56	17.78	-0.33	-0.83
	d	0.00	54	17.14	-0.15	-0.49
	9	0.00	4	1.27	-0.07	-0.91
Sol3	a	0.00	105	33.33	0.05	-0.27
	b	1.00	111	35.24	0.24	-0.15
	c	0.00	51	16.19	-0.23	-0.67
	d	0.00	35	11.11	-0.12	-0.63
	9	0.00	13	4.13	-0.07	-0.45

Tabelle D.4: Bewertung der Feldtest-Items (Teil 1)

Item (Trennschärfe)	Bewertung
Bro1 (0,45) δ = 0,12	- Trennschärfe gut - die richtige Antwort a wird von Personen mit der durchschnittl. höchsten Fähigkeit gewählt - der höchste Anteil der Personen wählt die falsche Alternative c! - die Distraktoren sind nicht ausgewogen besetzt *Bewertung*: besonders Distraktoren c (sollte seltener gewählt werden) und d (sollte häufiger gewählt werden) sollten überarbeitet werden
Bro2 (0,30) δ = 0,69	- die richtige Antwort d wird von Personen mit der durchschnittl. höchsten Fähigkeit gewählt - Distraktor b zieht die meisten Antworten auf sich - Distraktoren a und c sehen gut aus *Bewertung*: Distraktor b sollte überarbeitet werden
Bro3 (0,50) δ = -0,83	- gutes Item (Trennschärfe sehr gut) - die richtige Antwort d wird von Personen mit der durchschnittl. höchsten Fähigkeit gewählt - Die Distraktoren sind ausgewogen besetzt *Bewertung*: keine Überarbeitung notwendig
Hur1 (0,19) δ = 0,29	- schlechte Trennschärfe - die falsche Antwortalternative d wird von Personen mit der durchschnittl. höchsten Fähigkeit gewählt - der falsche Distraktor a wird am häufigsten gewählt (107/105) - auf die richtige Antwortalternative entfallen die zweitmeisten Antworten *Bewertung*: auf Distraktor a sollten weniger Antworten entfallen; Distraktor d muss überarbeitet werden, da die Fähigsten diese Alternative wählen
Hur2 (0,46) δ = -0,96	- gutes Item (Trennschärfe gut) - die richtige Antwort b wird von Personen mit der durchschnittl. höchsten Fähigkeit gewählt - Die Distraktoren werden ausgewogen gewählt *Bewertung*: keine Überarbeitung notwendig
Hur3 (0,40) δ = -0,12	- gutes Item (Trennschärfe gut - die richtige Antwort b wird von Personen mit der durchschnittl. höchsten Fähigkeit gewählt - die Distraktoren werden ausgewogen gewählt *Bewertung*: keine Überarbeitung notwendig
Sch1 (0,53) δ = 0,04	- gutes Item (sehr gute Trennschärfe) - die richtige Antwort c wird von Personen mit der durchschnittl. höchsten Fähigkeit gewählt - Die Distraktoren werden ausgewogen gewählt *Bewertung*: keine Überarbeitung nötig
Sch2 (0,29) δ = -0,10	- unzureichende Trennschärfe - die richtige Antwort d wird von Personen mit der durchschnittl. höchsten Fähigkeit gewählt - Distraktoren werden nicht ausgewogen gewählt: b wird zu selten und a zu häufig gewählt *Bewertung*: Distraktoren a und b sollten überarbeitet werden
Sch3 (0,48) δ = -0,77	- gutes Item (gute Trennschärfe) - die richtige Antwort d wird von Personen mit der durchschnittl. höchsten Fähigkeit gewählt - Distraktor a wird etwas häufiger gewählt als die übrigen *Bewertung*: Distraktor a noch mal anschauen. Überarbeitung ist nicht unbedingt notwendig
Sti1 (0,36) δ = -0,66	- Trennschärfe nicht ganz ausreichend - die richtige Antwort b wird von Personen mit der durchschnittl. höchsten Fähigkeit gewählt - Distraktor a wird etwas häufiger gewählt als die übrigen *Bewertung*: Distraktor a noch mal anschauen. Überarbeitung ist nicht unbedingt notwendig. Das Item ist eher leicht und könnte Eisbrecher-Funktion besitzen.
	Rot: Überarbeitung dringend notwendig Grün: Überarbeitung nicht unbedingt notwendig Orange: Überarbeitung notwendig Schwarz: keine Überarbeitung notwendig

Tabelle D.5: Bewertung der Feldtest-Items (Teil 2)

Item (Trennschärfe)	Bewertung
Sti2 (0,18) δ = 0,07	- schlechte Trennschärfe - die richtige Antwort b wird von Personen mit der durchschnittl. höchsten Fähigkeit gewählt - Distraktoren a und c werden ausgewogen gewählt - Distraktor d wird zu häufig gewählt *Bewertung*: Distraktor d muss überarbeitet werden: soll in der Konsequenz weniger häufig gewählt werden
Sti3 (0,35) δ = 0,20	- unzureichende Trennschärfe - die richtige Antwort a wird von Personen mit der durchschnittl. höchsten Fähigkeit gewählt - Distraktor b wird zu häufig gewählt *Bewertung*: Distraktor b muss auf jeden Fall überarbeitet werden: soll in der Konsequenz weniger häufig gewählt werden
KLe1 (0,32) δ = 0,37	- unzureichende Trennschärfe - die richtige Antwort a wird von Personen mit der durchschnittl. höchsten Fähigkeit gewählt - Die Distraktoren werden relativ gleich häufig gewählt *Bewertung*: Trennschärfe könnte besser sein
KLe2 (0,11) δ = 2,29	- schlechte Trennschärfe - Item ist sehr schwer (schwerstes Item, geht weit über die durchschnittliche Personenfähigkeit hinaus - die richtige Antwort d wird von Personen mit der durchschnittl. höchsten Fähigkeit gewählt - Die meisten Antworten entfallen auf den falschen Distraktor a (wird von Personen mit der zweitbesten durchschnittl. Fähigkeit gewählt) *Bewertung*: Distraktor a muss überarbeitet werden: soll in der Konsequenz weniger häufig gewählt werden; Item oder aber die richtige Antwortalternative sollte leichter werden
KLe3 (0,14) δ = 0,32	- schlechte Trennschärfe - Personen mit der durchschnittl. höchsten Fähigkeit können entscheiden sich entweder für die richtige Antwortalternative oder aber für Distraktor a - Distraktor a wird zu häufig gewählt *Bewertung*: Distraktor a muss überarbeitet werden: soll in der Konsequenz weniger häufig gewählt werden
Kli1 (0,33) δ = 0,40	- unzureichende Trennschärfe - die richtige Antwort d wird von Personen mit der durchschnittl. höchsten Fähigkeit gewählt - Die Distraktoren werden relativ gleich häufig gewählt *Bewertung*: Trennschärfe könnte besser sein
Kli2 (0,31) δ = 0,05	- unzureichende Trennschärfe - die richtige Antwort a wird von Personen mit der durchschnittl. höchsten Fähigkeit gewählt - Die Distraktoren werden relativ gleich häufig gewählt *Bewertung*: Trennschärfe könnte besser sein
Kli3 (0,30) δ = 0,22	- unzureichende Trennschärfe - die richtige Antwort c wird von Personen mit der durchschnittl. höchsten Fähigkeit gewählt - Die Distraktoren werden gleich häufig gewählt *Bewertung*: Trennschärfe könnte besser sein
Reg1 (0,34) δ = -0,20	- unzureichende Trennschärfe - die richtige Antwort d wird von Personen mit der durchschnittl. höchsten Fähigkeit gewählt - Distraktor a wird etwas häufiger als die anderen gewählt *Bewertung*: Trennschärfe könnte besser sein; Distraktor a noch einmal prüfen
Reg2 (0,22) δ = 0,91	- schlechte Trennschärfe - die richtige Antwort a wird von Personen mit der durchschnittl. höchsten Fähigkeit gewählt - Distraktor d wird häufiger gewählt als die richtige Alternative a *Bewertung*: Distraktor d muss überarbeitet werden: soll in der Konsequenz seltener gewählt werden oder aber Distraktor a muss leichter werden (lieber d überarbeiten, da a als richtige Antwort von den fähigsten Personen gewählt wird)

Tabelle D.6: Bewertung der Feldtest-Items (Teil 3)

Item (Trennschärfe)	Bewertung
Reg3 (0,25) δ = 0,16	- schlechte Trennschärfe - die richtige Antwort b wird von Personen mit der durchschnittl. höchsten Fähigkeit gewählt - Die Distraktoren werden relativ gleich häufig gewählt *Bewertung*: Trennschärfe sollte besser sein
Ros1 (0,48) δ = -0,93	- gutes Item (gute Trennschärfe) - die richtige Antwort c wird von Personen mit der durchschnittl. höchsten Fähigkeit gewählt - Die Distraktoren werden ausgewogen gewählt *Bewertung*: keine Überarbeitung nötig
Ros2 (0,47) δ = 0,16	- gutes Item (gute Trennschärfe) - die richtige Antwort d wird von Personen mit der durchschnittl. höchsten Fähigkeit gewählt - Die Distraktoren werden ausgewogen gewählt *Bewertung*: keine Überarbeitung nötig
Ros3 (0,42) δ = -0,55	- gutes Item (gute Trennschärfe) - die richtige Antwort b wird von Personen mit der durchschnittl. höchsten Fähigkeit gewählt - Distraktor d wird zu selten gewählt *Bewertung*: keine Überarbeitung nötig, evtl. Distraktor d noch einmal prüfen
Sko1 (0,34) δ = -0,43	- unzureichende Trennschärfe - die richtige Antwort a wird von Personen mit der durchschnittl. höchsten Fähigkeit gewählt - Distraktor c wird zu selten gewählt *Bewertung*: Trennschärfe könnte besser sein; Distraktor c muss überarbeitet werden: soll in der Konsequenz häufiger gewählt werden
Sko2 (0,49) δ = -0,99	- gutes Item (gute Trennschärfe) - die richtige Antwort c wird von Personen mit der durchschnittl. höchsten Fähigkeit gewählt - Die Distraktoren werden relativ ausgewogen gewählt *Bewertung*: keine Überarbeitung nötig
Sko3 (0,35) δ = 0,40	- unzureichende Trennschärfe - die richtige Antwort c wird von Personen mit der durchschnittl. höchsten Fähigkeit gewählt - Distraktor a wird häufiger gewählt als die richtige Antwort *Bewertung*: Trennschärfe könnte besser sein; Distraktor a muss überarbeitet werden: soll in der Konsequenz seltener gewählt werden
Sol1 (0,34) δ = -0,59	- unzureichende Trennschärfe - die richtige Antwort a wird von Personen mit der durchschnittl. höchsten Fähigkeit gewählt - Distraktor c wird zu häufig gewählt *Bewertung*: Trennschärfe könnte besser sein; Distraktor c muss überarbeitet werden: soll in der Konsequenz seltener gewählt werden
Sol2 (0,46) δ = 0,09	- gutes Item (gute Trennschärfe) - die richtige Antwort a wird von Personen mit der durchschnittl. höchsten Fähigkeit gewählt - Die Distraktoren werden relativ ausgewogen gewählt *Bewertung*: keine Überarbeitung nötig
Sol3 (0,24) δ = 0,35	- schlechte Trennschärfe - die richtige Antwort b wird von Personen mit der durchschnittl. höchsten Fähigkeit gewählt - Distraktor a wird zu häufig gewählt *Bewertung*: Trennschärfe sollte besser sein; Distraktor a muss überarbeitet werden: soll in der Konsequenz seltener gewählt werden

D.2 Haupttest

D.2.1 Item-Charakteristik-Kurven (ICCs)

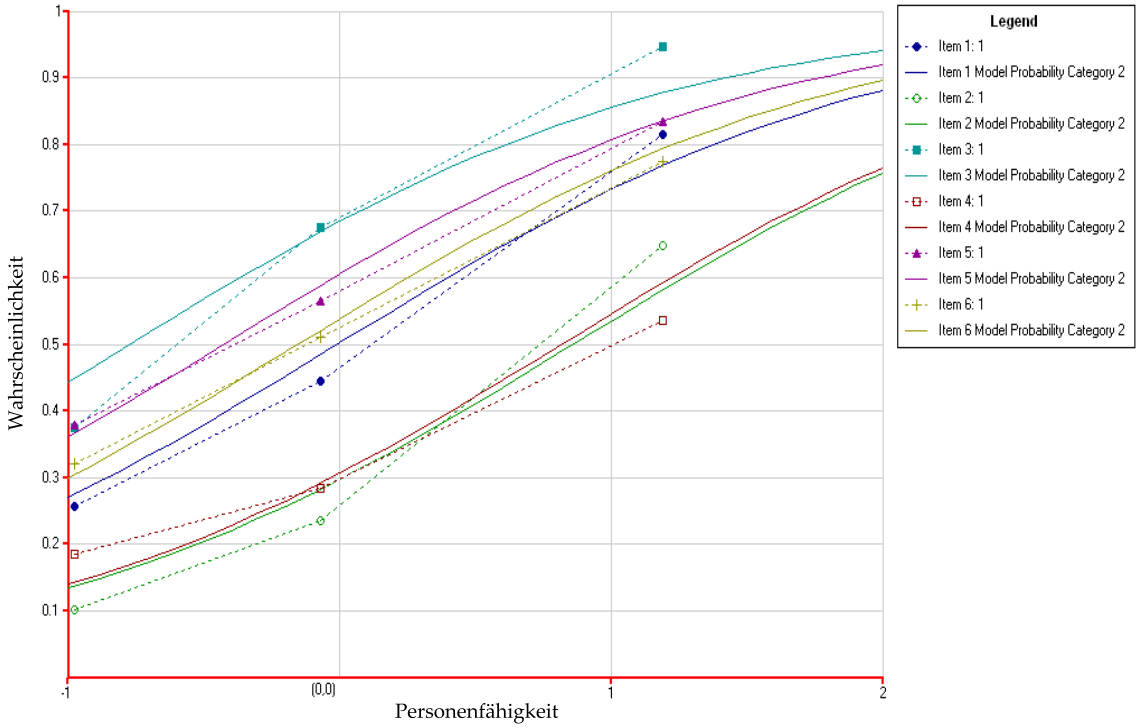


Abbildung D.1: ICCs der Items 1 bis 6 vor Itemauswahl

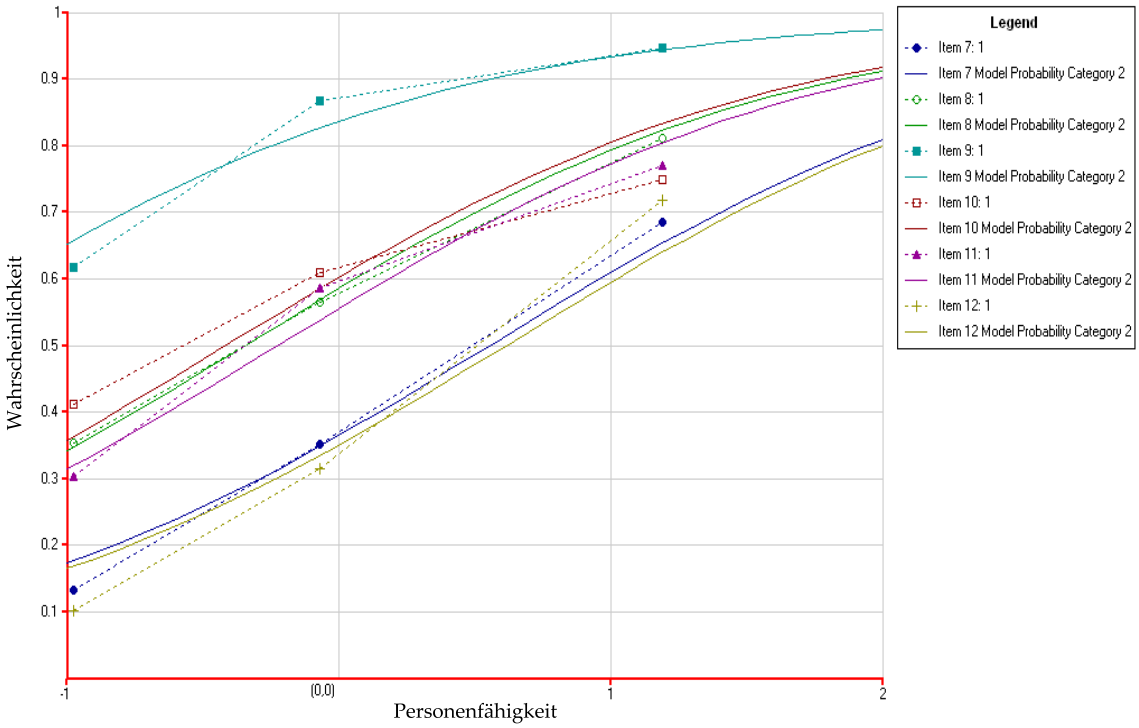


Abbildung D.2: ICCs der Items 7 bis 12 vor Itemauswahl

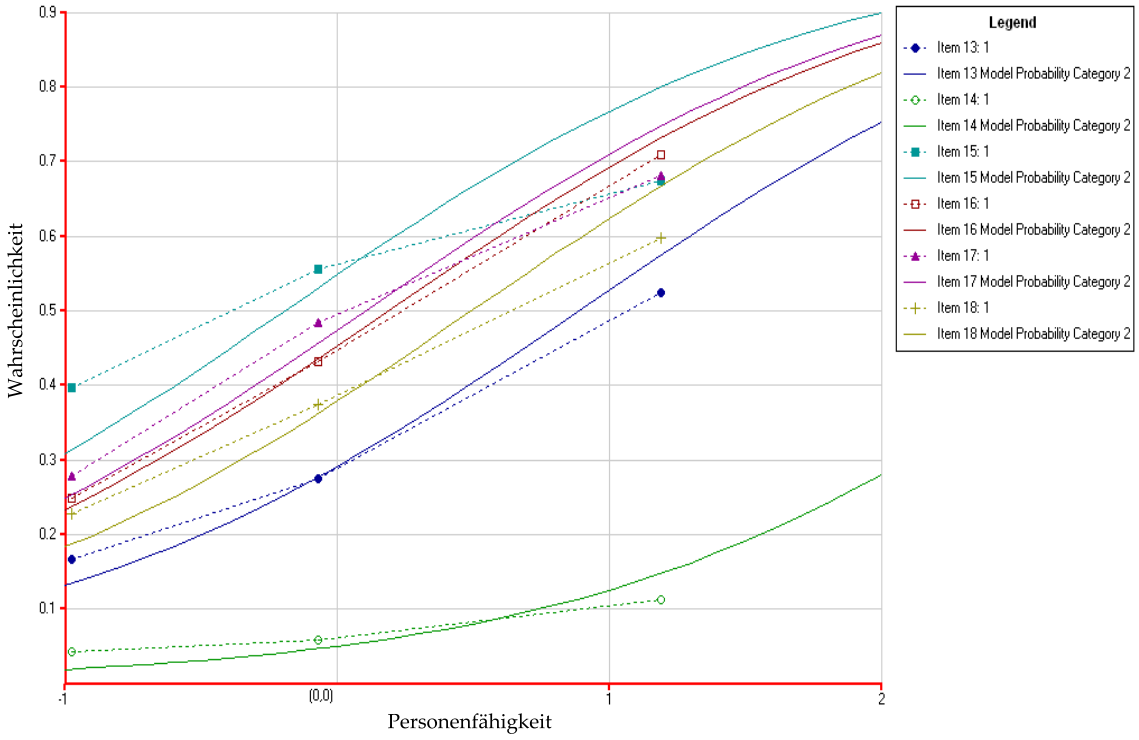


Abbildung D.3: ICCs der Items 13 bis 18 vor Itemauswahl

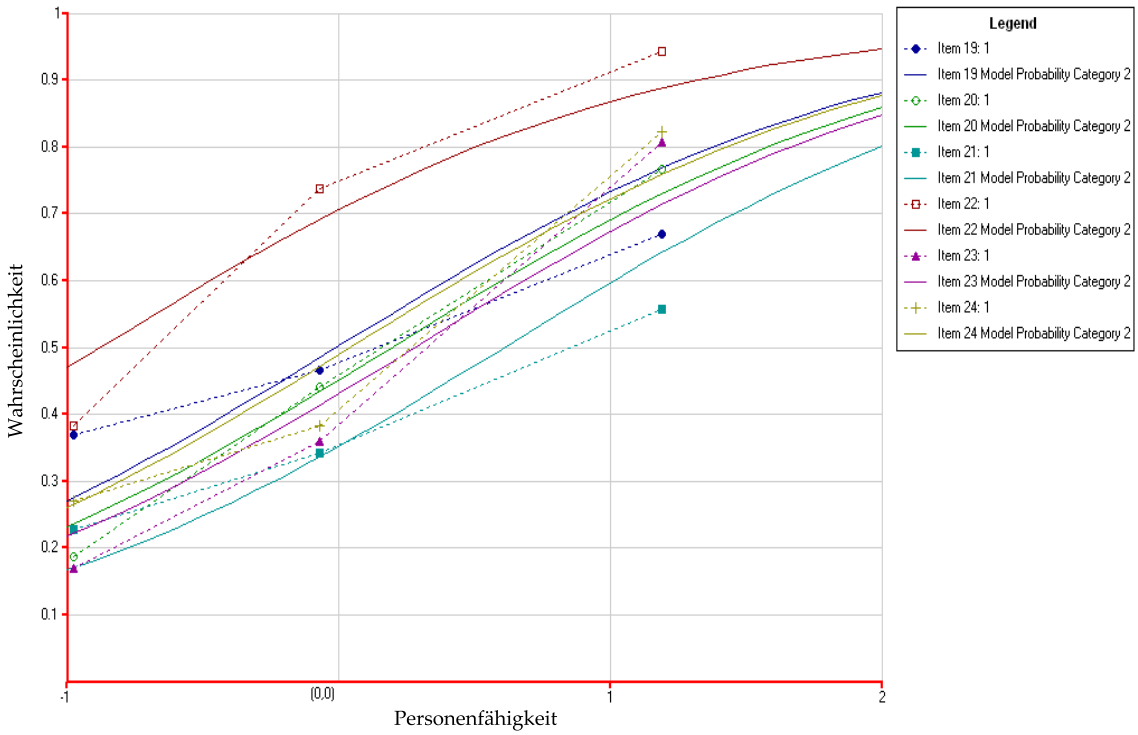


Abbildung D.4: ICCs der Items 19 bis 24 vor Itemauswahl

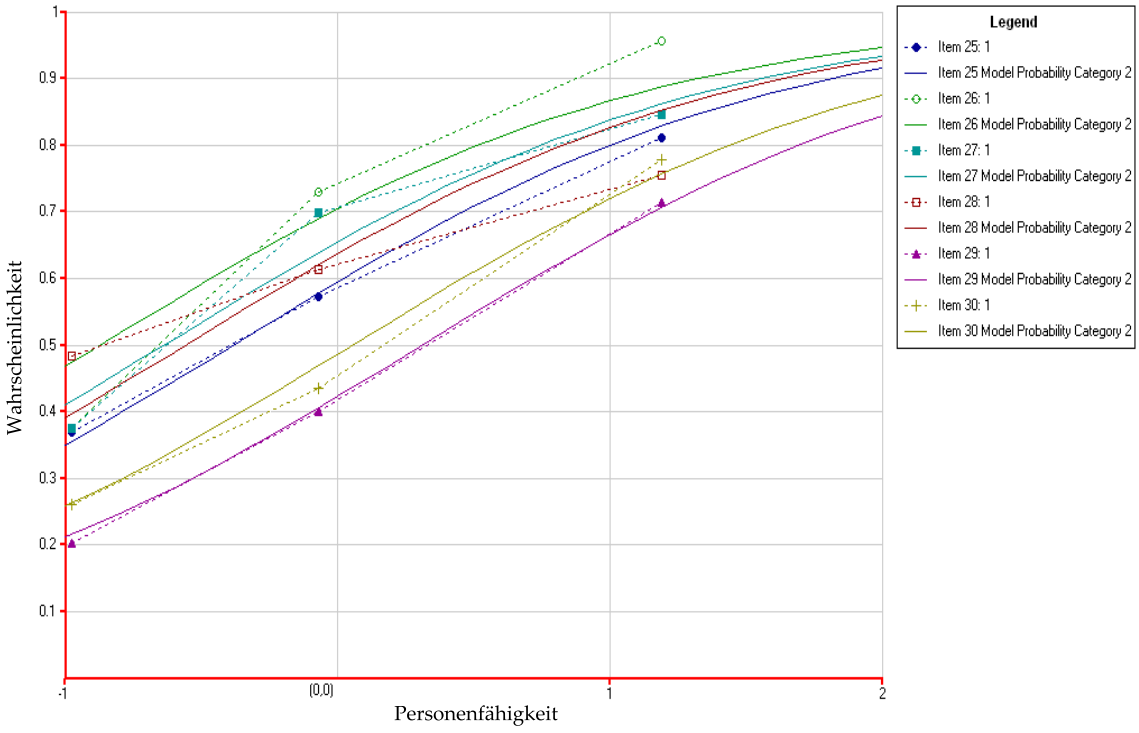


Abbildung D.5: ICCs der Items 25 bis 30 vor Itemauswahl

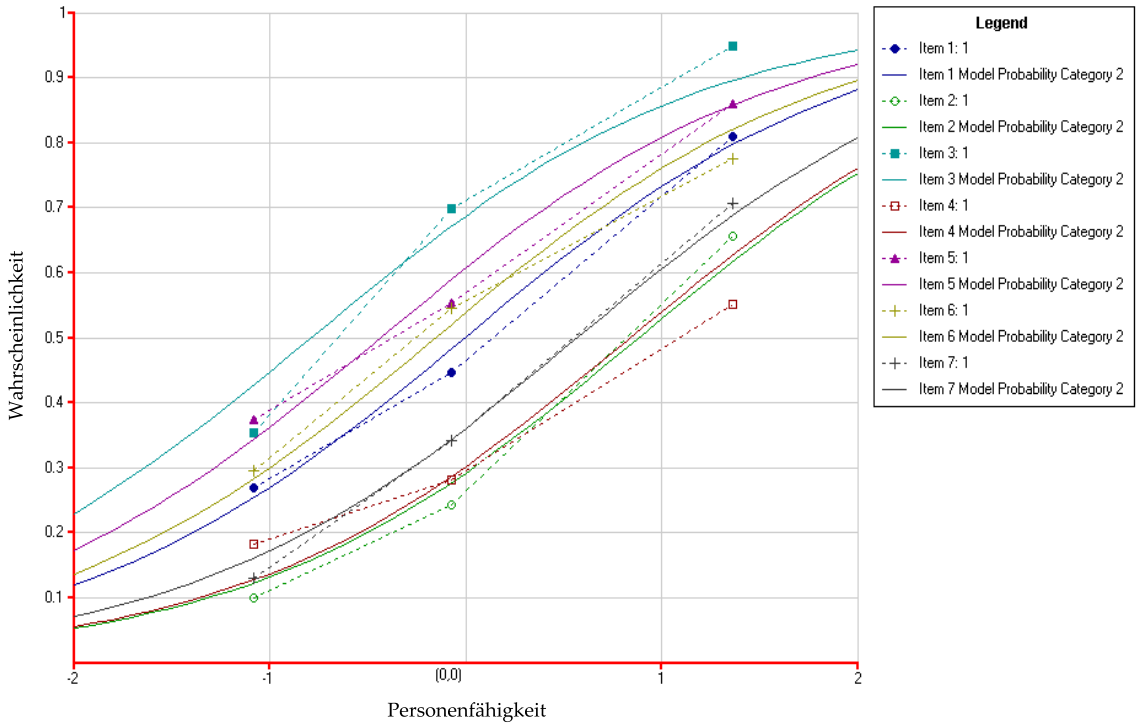


Abbildung D.6: ICCs der Items 1 bis 7

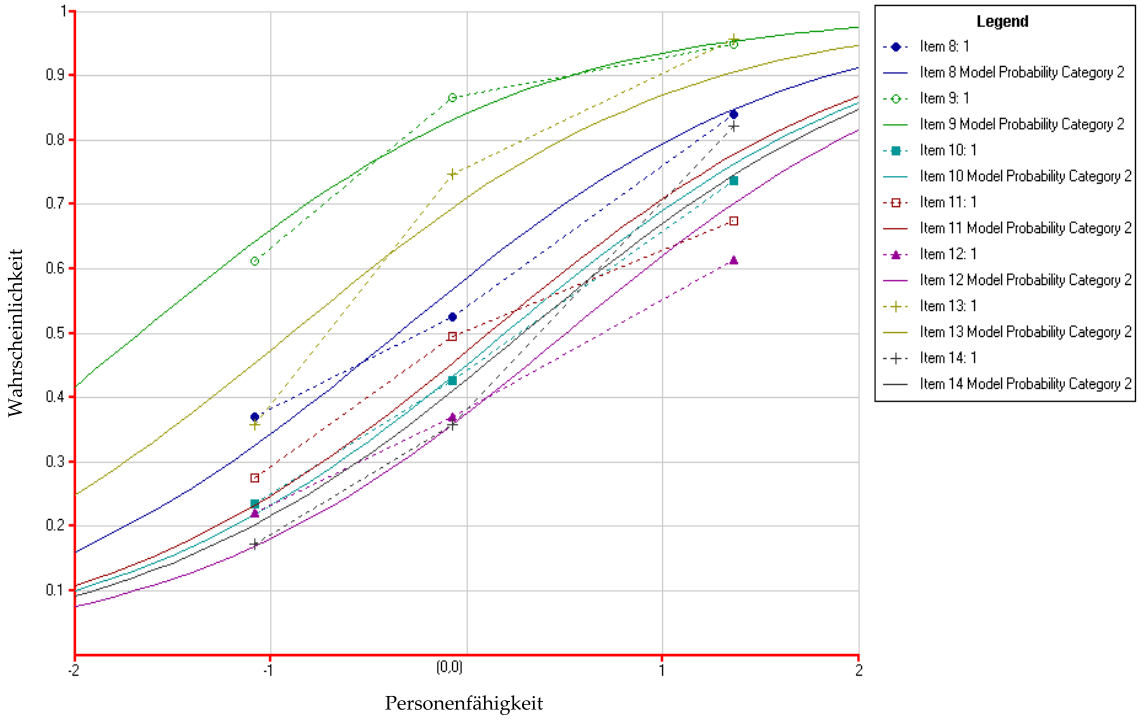


Abbildung D.7: ICCs der Items 8 bis 14

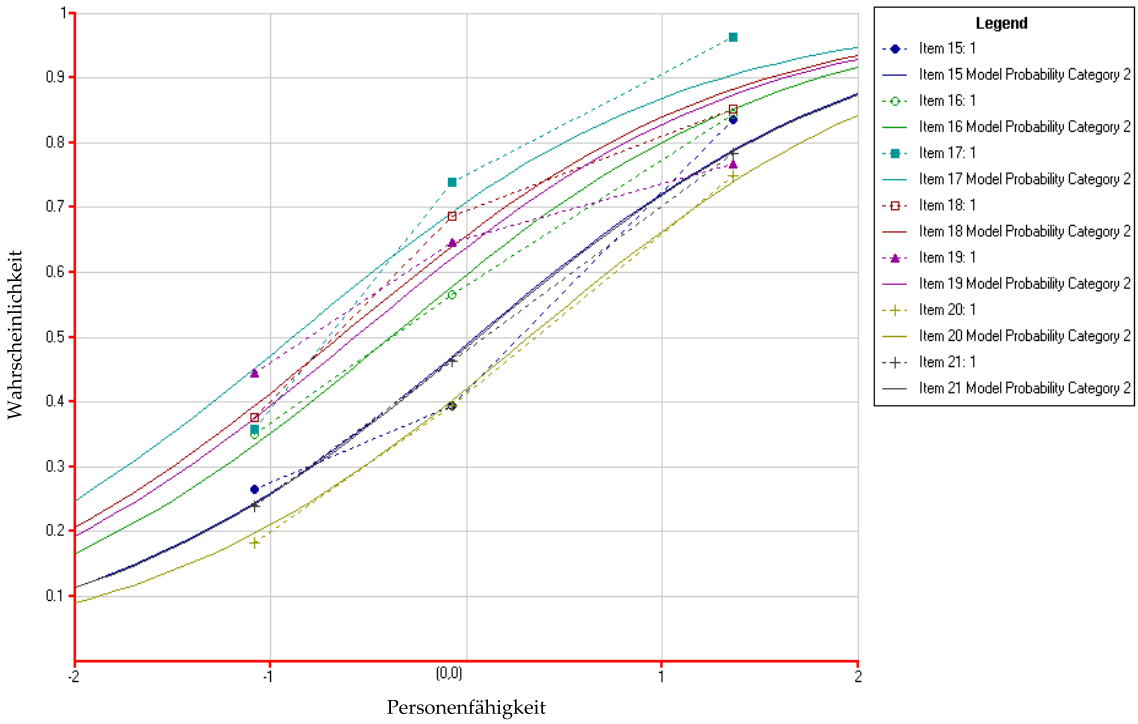


Abbildung D.8: ICCs der Items 15 bis 21

D.2.2 Differential-Item-Functioning

D.2.3 Wright-Maps der abschließenden Testversion

Tabelle D.7: Prüfung von DIF-Effekten anhand der Geschlechterstichproben

Schülerinnen					Logit-Differenz	Schüler				
			CI					CI		
Items	Schwierigkeit	Std. Err	untere Grenze	obere Grenze		Items	Schwierigkeit	Std. Err	untere Grenze	obere Grenze
Bro 1	-0,06	0,081	-0,219	0,099	0,39	Bro 1	0,33	0,080	0,170	0,484
Bro 2	0,92	0,083	0,753	1,079	0,22	Bro 2	1,14	0,083	0,976	1,302
Bro 3	-0,62	0,083	-0,778	-0,452	0,07	Bro 3	-0,68	0,082	-0,843	-0,521
Hur 1	1,19	0,085	1,020	1,354	0,39	Hur 1	0,79	0,082	0,632	0,954
Hur 2	-0,13	0,081	-0,286	0,032	0,34	Hur 2	-0,46	0,081	-0,622	-0,304
Hur 3	0,12	0,081	-0,035	0,283	0,27	Hur 3	-0,15	0,080	-0,305	-0,009
Sch 1	0,84	0,083	0,680	1,006	0,25	Sch 1	0,59	0,081	0,431	0,749
Sch 2	-0,33	0,082	-0,488	-0,166	0,23	Sch 2	-0,10	0,080	-0,253	0,061
Sch 3	-1,77	0,091	-1,948	-1,592	0,46 *	Sch 3	-1,31	0,086	-1,479	-1,141
Kli 1	0,42	0,082	0,255	0,577	0,15	Kli 1	0,26	0,080	0,106	0,420
Kli 2	0,06	0,081	-0,101	0,217	0,39	Kli 2	0,45	0,080	0,288	0,602
Kli 3	0,69	0,082	0,528	0,850	0,07	Kli 3	0,62	0,081	0,458	0,776
Ros 1	-0,63	0,083	-0,791	-0,465	0,25	Ros 1	-0,88	0,083	-1,045	-0,719
Ros 2	0,35	0,082	0,189	0,511	0,16	Ros 2	0,51	0,081	0,351	0,669
Ros 3	-0,03	0,081	-0,192	0,126	0,44 *	Ros 3	0,41	0,080	0,248	0,562
Sko 1	-0,27	0,082	-0,433	-0,111	0,05	Sko 1	-0,23	0,080	-0,383	-0,069
Sko 2	-0,77	0,083	-0,934	-0,608	0,05	Sko 2	-0,72	0,082	-0,885	-0,563
Sko 3	-0,60	0,083	-0,763	-0,437	0,18	Sko 3	-0,42	0,081	-0,583	-0,265
Sol 1	-0,41	0,082	-0,567	-0,245	0,04	Sol 1	-0,45	0,081	-0,609	-0,291
Sol 2	0,68	0,082	0,514	0,836	0,41	Sol 2	0,26	0,080	0,105	0,419
Sol 3	0,35	0,370	-0,375	1,075	0,29	Sol 3	0,06	0,363	-0,654	0,768

* Logit-Differenz 0,43 - 0,64 = mittelmäßig

Tabelle D.8: Prüfung von DIF-Effekten anhand der Haupt- und Realschulstichproben

Hauptschule			CI		**Logit-Differenz**	**Realschule**			CI	
Items	**Schwierigkeit**	**Std. Err**	**untere Grenze**	**obere Grenze**		**Items**	**Schwierigkeit**	**Std. Err**	**untere Grenze**	**obere Grenze**
Bro 1	0,17	0,097	-1,024	0,356	0,05	Bro 1	0,20	0,096	0,016	0,392
Bro 2	1,33	0,110	1,115	1,547	0,12	Bro 2	1,21	0,102	1,009	1,409
Bro 3	-0,54	0,093	-0,720	-0,356	0,22	Bro 3	-0,75	0,098	-0,946	-0,562
Hur 1	0,77	0,103	0,568	0,972	0,42	Hur 1	1,19	0,102	0,985	1,385
Hur 2	-0,27	0,094	-0,450	-0,082	0,06	Hur 2	-0,33	0,097	-0,519	-0,139
Hur 3	0,13	0,096	-0,060	0,316	0,35	Hur 3	-0,22	0,096	-0,406	-0,030
Sch 1	0,85	0,104	0,644	1,052	0,09	Sch 1	0,76	0,098	0,565	0,949
Sch 2	-0,33	0,094	-0,518	-0,150	0,21	Sch 2	-0,13	0,096	-0,314	0,062
Sch 3	-1,56	0,098	-1,756	-1,372	0,08	Sch 3	-1,48	0,105	-1,688	-1,276
Kli 1	0,25	0,097	0,055	0,435	0,15	Kli 1	0,39	0,097	0,200	0,580
Kli 2	-0,06	0,095	-0,241	0,131	0,24	Kli 2	0,19	0,096	-0,002	0,374
Kli 3	0,31	0,098	0,113	0,497	0,58 *	Kli 3	0,88	0,099	0,687	1,075
Ros 1	-0,62	0,093	-0,802	-0,438	0,23	Ros 1	-0,85	0,099	1,048	-0,660
Ros 2	0,82	0,103	0,621	1,025	0,32	Ros 2	0,50	0,097	0,314	0,694
Ros 3	0,17	0,097	-0,023	0,357	0,08	Ros 3	0,24	0,096	0,054	0,430
Sko 1	-0,20	0,094	-0,380	0,012	0,08	Sko 1	-0,27	0,096	-0,460	-0,084
Sko 2	-0,39	0,094	-0,569	-0,201	0,58 *	Sko 2	-0,96	0,100	-1,156	-0,764
Sko 3	-0,55	0,093	-0,736	-0,372	0,16	Sko 3	-0,71	0,098	-0,904	-0,520
Sol 1	-0,81	0,094	-0,989	-0,621	0,27	Sol 1	-0,54	0,097	-0,726	-0,346
Sol 2	0,54	0,100	0,339	0,731	0,18	Sol 2	0,35	0,097	0,162	0,542
Sol 3	0,00	0,436	-0,857	0,853	0,34	Sol 3	0,33	0,439	-0,526	1,194

* Logit-Differenz 0,426 - 0,638 = mittelmäßig

Tabelle D.9: Prüfung von DIF-Effekten zwischen Haupt- und Gymnasialstichprobe

Hauptschule					Logit-Differenz	Gymnasium				
			CI					CI		
Items	Schwierigkeit	Std.Err.	untere Grenze	obere Grenze		Items	Schwierigkeit	Std.Err.	untere Grenze	obere Grenze
Bro 1	0,17	0,097	-0,024	0,356	0,17	Bro 1	0,00	0,104	-0,205	0,203
Bro 2	1,33	0,110	1,115	1,547	0,63 *	Bro 2	0,70	0,100	0,506	0,898
Bro 3	-0,54	0,093	-0,720	-0,356	0,16	Bro 3	-0,70	0,112	-0,920	-0,480
Hur 1	0,77	0,103	0,568	0,972	0,20	Hur 1	0,97	0,100	0,774	1,166
Hur 2	-0,27	0,094	-0,450	-0,082	0,05	Hur 2	-0,31	0,107	-0,521	-0,101
Hur 3	0,13	0,096	-0,060	0,316	0,08	Hur 3	0,05	0,104	-0,158	0,250
Sch 1	0,85	0,104	0,644	1,052	0,28	Sch 1	0,57	0,101	0,368	0,764
Sch 2	-0,33	0,094	-0,518	-0,150	0,17	Sch 2	-0,16	0,106	-0,370	0,064
Sch 3	-1,56	0,098	-1,756	-1,372	0,11	Sch 3	-1,45	0,121	-1,688	-1,214
Kli 1	0,25	0,097	0,055	0,435	0,12	Kli 1	0,37	0,102	0,166	0,566
Kli 2	-0,06	0,095	-0,241	0,131	0,70 **	Kli 2	0,65	0,101	0,447	0,843
Kli 3	0,31	0,098	0,113	0,497	0,44 *	Kli 3	0,74	0,100	0,546	0,938
Ros 1	-0,62	0,093	-0,802	-0,438	0,24	Ros 1	-0,86	0,113	-1,079	-0,637
Ros 2	0,82	0,103	0,621	1,025	0,85 **	Ros 2	-0,02	0,105	-0,228	0,184
Ros 3	0,17	0,097	-0,023	0,357	0,03	Ros 3	0,13	0,103	-0,069	0,335
Sko 1	-0,20	0,094	-0,380	0,012	0,12	Sko 1	-0,31	0,107	-0,521	-0,101
Sko 2	-0,39	0,094	-0,569	-0,201	0,77 **	Sko 2	-1,15	0,117	-1,379	-0,921
Sko 3	-0,55	0,093	-0,736	-0,372	0,37	Sko 3	-0,19	0,106	-0,396	0,020
Sol 1	-0,81	0,094	-0,989	-0,621	1,02 **	Sol 1	0,22	0,103	0,016	0,420
Sol 2	0,54	0,100	0,339	0,731	0,03	Sol 2	0,51	0,101	0,307	0,703
Sol 3	0,00	0,436	-0,857	0,853	0,26	Sol 3	0,26	0,473	-0,666	1,188

* = Logit-Differenz 0,426 bis 0,638 = mittelmäßig

** = Logit-Differenz > 0,638 = groß

Tabelle D.10: Prüfung von DIF-Effekten anhand der Real- und Gymnasialstichproben

Realschule					Logit-Differenz	Gymnasium				
			CI					CI		
Items	Schwierigkeit	Std. Err	untere Grenze	obere Grenze		Items	Schwierigkeit	Std. Err	untere Grenze	obere Grenze
Bro 1	0,20	0,096	0,016	0,392	0,21	Bro 1	0,00	0,104	-0,205	0,203
Bro 2	1,21	0,102	1,009	1,409	0,51 *	Bro 2	0,70	0,100	0,506	0,898
Bro 3	-0,75	0,098	-0,946	-0,562	0,05	Bro 3	-0,70	0,112	-0,920	-0,480
Hur 1	1,19	0,102	0,985	1,385	0,22	Hur 1	0,97	0,100	0,774	1,166
Hur 2	-0,33	0,097	-0,519	0,139	0,02	Hur 2	-0,31	0,107	-0,521	-0,101
Hur 3	-0,22	0,096	-0,406	0,030	0,26	Hur 3	0,05	0,104	-0,158	0,250
Sch 1	0,76	0,098	-0,565	0,949	0,19	Sch 1	0,57	0,101	0,368	0,764
Sch 2	-0,13	0,096	-0,314	0,062	0,04	Sch 2	-0,16	0,106	-0,370	0,046
Sch 3	-1,48	0,105	-1,688	-1,276	0,03	Sch 3	-1,45	0,121	-1,688	-1,214
Kli 1	0,39	0,097	0,200	0,580	0,02	Kli 1	0,37	0,102	0,166	0,566
Kli 2	0,19	0,096	-0,002	0,374	0,46 *	Kli 2	0,65	0,101	0,447	0,843
Kli 3	0,88	0,099	0,687	1,075	0,14	Kli 3	0,74	0,100	0,546	0,938
Ros 1	-0,85	0,099	-1,048	-0,660	0,00	Ros 1	-0,86	0,113	-1,079	-0,637
Ros 2	0,50	0,097	0,314	0,694	0,53 *	Ros 2	-0,02	0,105	-0,228	0,184
Ros 3	0,24	0,096	0,054	0,430	0,11	Ros 3	0,13	0,103	-0,069	0,335
Sko 1	-0,27	0,096	-0,460	0,084	0,04	Sko 1	-0,31	0,107	-0,521	-0,101
Sko 2	-0,96	0,100	-1,156	-0,764	0,19	Sko 2	-1,15	0,117	-1,379	-0,921
Sko 3	-0,71	0,098	-0,904	-0,520	0,52 *	Sko 3	-0,19	0,106	-0,396	0,020
Sol 1	-0,54	0,097	-0,726	-0,346	0,75 **	Sol 1	0,22	0,103	0,016	0,420
Sol 2	0,35	0,097	-0,162	0,542	0,15	Sol 2	0,51	0,101	0,307	0,703
Sol 3	0,33	0,439	-0,526	1,194	0,07	Sol 3	0,26	0,473	-0,666	1,188

* Logit-Differenz 0,43 - 0,64 = mittelmäßig
** Logit-Differenz > 0,64 = groß

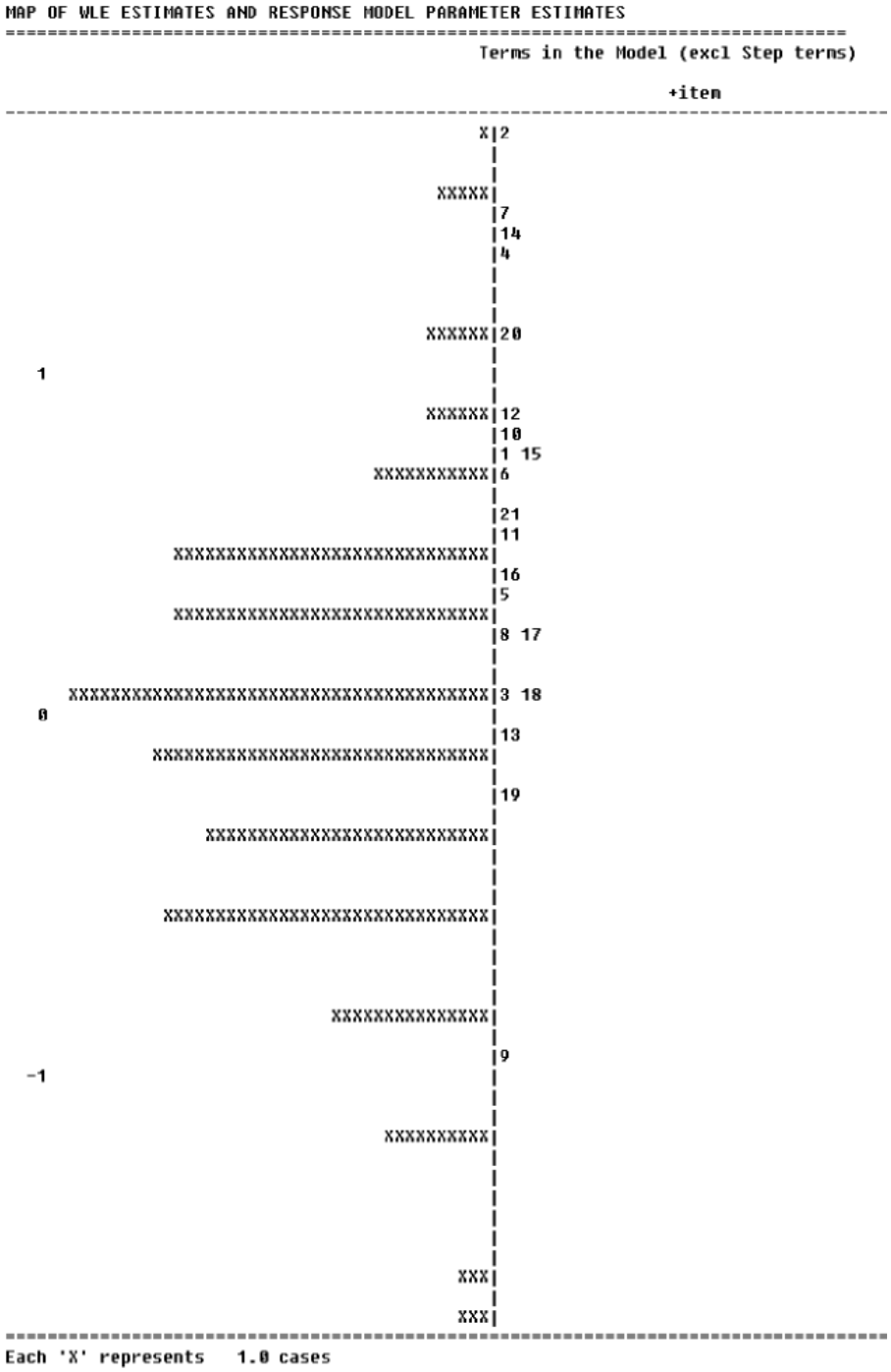


Abbildung D.9: Wright-Map Hauptschule

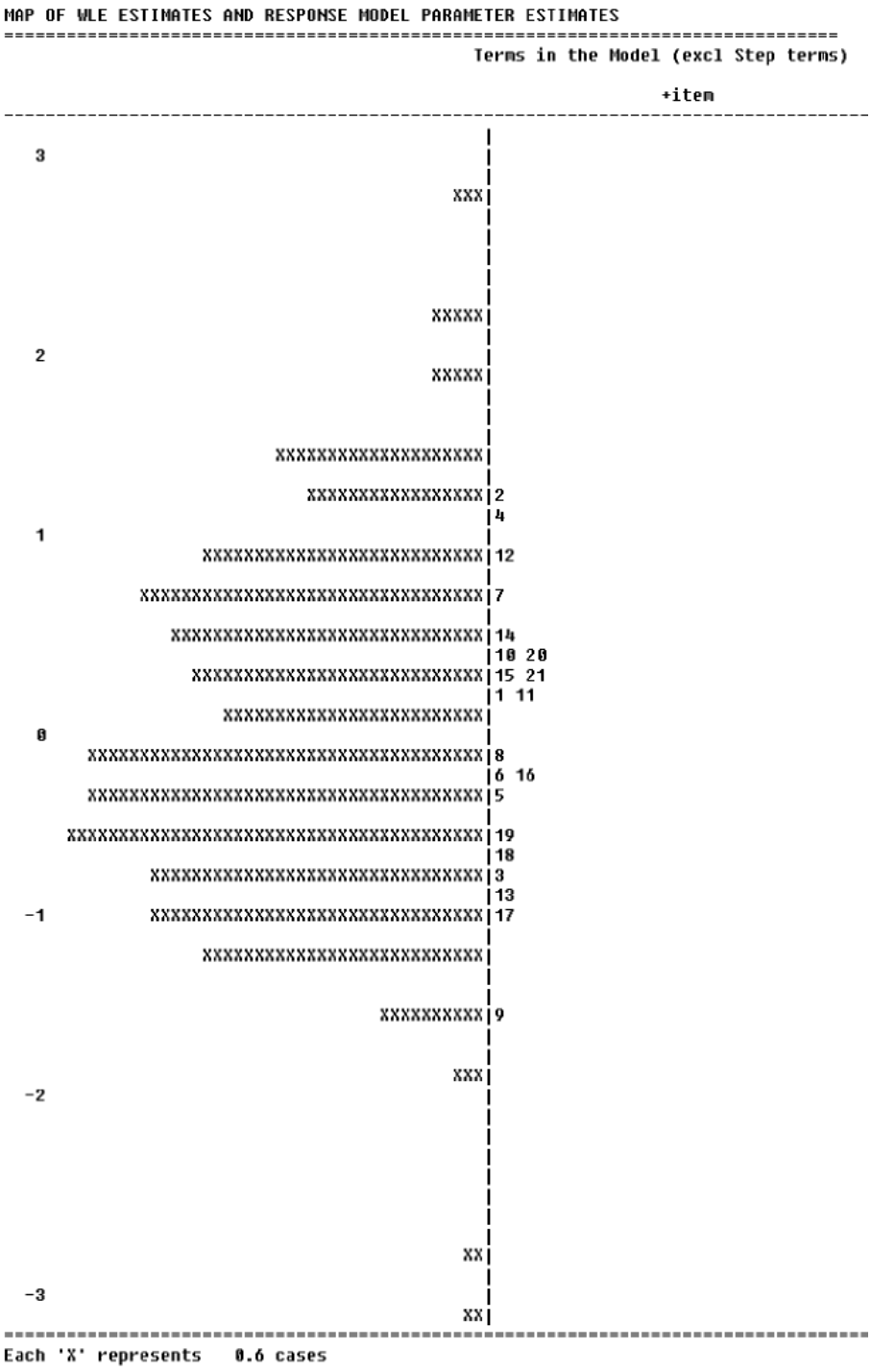


Abbildung D.10: Wright-Map Realschule

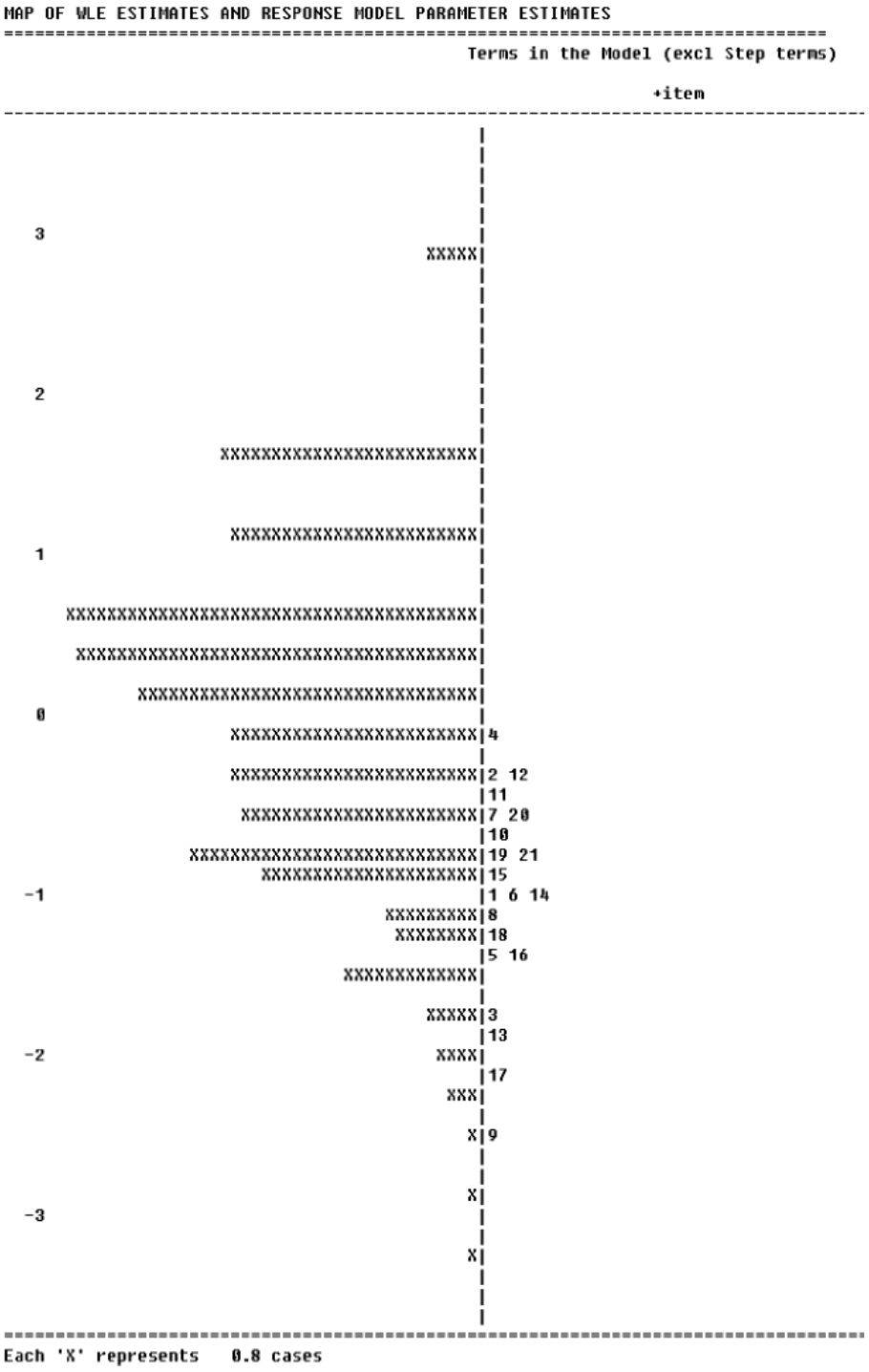


Abbildung D.11: Wright-Map Gymnasium

Abbildungsverzeichnis

Tabellenverzeichnis

Literaturverzeichnis

Adams, R. & Khoo, S. (1996). *Quest: The interactive test analysis system-version 2.1* (Bericht). Victoria: Australian Council for Educational Research.

Adams, R. J., Wilson, M. R. & Wang, W. C. (1997). The multidimensional random coefficients multinomial logit. *Applied Psychological Measurement*.

Aikenhead, G. S. (1988). An analysis of four ways of assessing student beliefs about STS topics. *Journal of Research in Science Teaching*, *25* (8), 607-629.

Amelang, M. & Zielinski, W. (2002). *Psychologische Diagnostik und Intervention*. Berlin: Springer-Verlag.

American Association of the Advancement of Science. (1989). *Science for all Americans*. New York: Oxford University Press.

American Association of the Advancement of Science. (2001). *Atlas of scientific literacy - Project 2061*. Washington, D.C.: American Association for the Advancement of Science (AAAS).

Amsel, E. & Brock, S. (1996). The development of evidence evaluation skills. *Cognitive Development*, *11*, 523-550.

Anderson, R. D. (2007). Inquiry as an organizing theme for science curricula. In S. K. Abell & N. G. Lederman (Hrsg.), *Handbook of research in science teaching* (S. 807-830). Mahwah: Lawrence Erlbaum.

Andrich, D. (1988). *Rasch models for measurement*. Newbury Park: Sage.

Artelt, C., Stanat, P., Schneider, W. & Schiefele, U. (2001). Lesekompetenz: Testkonzeption und Ergebnisse. In J. Baumert et al. (Hrsg.), *PISA 2000 - Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich.* Opladen: Leske + Budrich.

Ayala, C. C., Shavelson, R. J., Yin, Y. & Schultz, S. E. (2002). Reasoning dimensions underlying science achievement: The case of perfomance assessment. *Educational Assessment*, *8* (2), 101-121.

Baker, F. B. & Kim, S.-H. (2004). *Item response theory: Paramter estimation techniques*. New York, NY: Dekker.

Ballstaedt, S.-P. (1997). *Wissensvermittlung. Die Gestaltung von Lernmaterial*. Weinheim: Beltz PVU.

Baumert, J. et al. (Hrsg.). (2001). *PISA 2000: Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich*. Opladen: Leske + Budrich.

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

Baxter, G. P., Shavelson, R. J., Goldmann, S. R. & Pine, J. (1992). Evaluation of procedure-based scoring for hands-on science assessment. *Journal of Educational Measurement*, *29* (1), 1-17.

Beaton, A. & Allen, N. (1992). Interpreting scales through scale anchoring. *Journal of Educational Statistics*, *17* (2), 191-204.

Biological Sciences Curriculum Study. (1965). *Process of science test: Form A*. New York: The Psychological Corporation.

Bortz, J. (1999). *Statistik für Sozialwissenschaftler* (5. Aufl.). Heidelberg: Springer.

Branch, J. L. (2000). Investigating the information-seeking processes of adolescents: the value of using think alouds and think afters. *Library & Information Science Research*, *22* (4), 371-392.

Bresler, S., Kuck, C., Lichtenberger, J. & Pollmann, M. (2006). Physik Interaktiv: Naturwissenschaftliches Arbeiten. Berlin: Cornelsen.

Burnham, K. P. & Anderson, D. R. (2002). *Model selection and multimodal inference*. New York: Springer.

Bybee, R. (2002). Scientific Literacy - Mythos oder Realität. In W. Gräber, P. Nentwig, T. Koballa & R. Evans (Hrsg.), *Scientific literacy* (S. 21-43). Opladen: Leske + Budrich.

Bybee, R. W. (1997). *Achieving scientific literacy: from purposes to practices*. Oxford: Heinemann Educational Books.

Camilli, G. (1993). The case against item bias detection techniques based on internal criteria: Do item bias procedures obscure test fairness issues? In P. W. Holland & H. Wainer (Hrsg.), *Differential Item Functioning* (S. 397-413). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Carey, S., Evans, R., Honda, M., Jay, E. & Unger, C. (1989). ´An experiment is when you try it and see if it works‘: a study of grade 7 students‘ understanding of the construction of scientific knowledge. *International Journal of Science Education*, *11*, 514-529.

Carey, S. & Smith, C. (1993). On understanding the nature of scientific knowledge. *Educational Psychologist*, *28* (3), 235-251.

Chen, Z. & Klahr, D. (1999). All other things being equal: acquisition and transfer of the control of variables strategy. *Child Development*, *70* (5), 1098-1120.

Chinn, C. A. & Malhotra, B. A. (2002b). Epistemologically authentic inquiry in schools: A theoretical framework for evaluating inquiry tasks. *Science Education*, *86*, 175-218.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge: MIT Press.

Cooley, W. W. & L., K. (1961). *Test on understanding science: Form W*. Princeton, NJ: Educational Testing Service.

Crehan, K. D., Haladyna, T. M. & Brewer, B. W. (1993). Use of an inclusive option and the optimal number of options for multiple-choice items. *Educational and Psychological Measurement*, *53*, 241-247.

Daum, A. (2002). *Wissenschaftspopularisierung im 19. Jahrhundert: Bürgerliche Kultur, naturwissenschaftliche Bildung und die deutsche Öffentlichkeit 1848-1914*. München: Oldenbourg Wissenschaftsverlag.

DeBoer, G. E. (1997). Historical perspectives on scientific literacy. In W. Gräber & C. Bolte (Hrsg.), *Scientific literacy.* Kiel: IPN.

DeGarmo, C. (1895/2007). *Herbart and the Herbartarians*. Whitefish, MT: Kessinger Pub. & Co.

Deloache, J. S., Miller, K. F. & Pierroutsakos, S. L. (1998). Reasoning and problem solving. In W. Damon, D. Kuhn & R. S. Siegler (Hrsg.), *Handbook of Child Psychology: Cognition, Perception and Language* (Bd. 2, S. 801-849). Chichester: John Wiley & Sons, Inc.

Department of Education and Science. (1984). *Science in schools age 11: Report No.3.* London.

Dewey, J. (1916/1993). *Demokratie und Erziehung: eine Einleitung in die philosophische Pädagogik* (J. Oelkers, Hrsg.). Weinheim und Basel: Beltz.

Dillashaw, F. G. & Okey, J. R. (1980). Test of integrated science process skills for secondary science students. *Science Education*, *64* (5), 601-608.

Drechsel, B. & Artelt, C. (2007). Lesekompetenz. In M. Prenzel et al. (Hrsg.), *PISA ´06 - Die Ergebnisse der dritten internationalen Vergleichsstudie* (S. 225-248). Münster: Waxmann.

Du, Y. (1995). When to adjust for differential item functioning. *Rasch Measurement Transactions*, *9* (1).

Duit, R., Häußler, P. & Prenzel, M. (2001). Schulleistungen im Bereich der naturwissenschaftlichen Bildung. In F. Weinert (Hrsg.), *Leistungsmessung in Schulen* (S. 169-186). Weinheim: Beltz.

Duncker, K. (1945). On problem-solving. In J. Dashiell (Hrsg.), *Psychological monographs* (S. 1-114). Washington, D.C.: American Psychological Association.

Duschl, R. A., Hamilton, R. J. & Grandy, R. E. (1992). Psychology and epistemology: Match or mismatch when applied to science education? In R. A. Duschl & R. J. Hamilton (Hrsg.), *Philosophy of science, cognitive psychology, and educational theory and practice* (S. 19-47). Albany, NY: State University of New York Press.

Edmondson, K. M. & Novak, J. D. (1993). The interplay of scientific epistemological views, learning strategies, and attitudes of college students. *Journal of Research in Science Teaching*, *30* (6), 547-559.

Education, N. S. for the Study of. (1932). *A program for teaching science: Thirty-first yearbook of the NSEE*. Chicago: University of Chicago Press.

Engeln, K. & Rost, J. (2006). Increasing students´interest: Informal learning in authentic science labs. In L. Xingfeng & W. Boone (Hrsg.), *Applications of Rasch Measurement in Science Education.* Maple Grove, MN: Jam Press.

Ericsson, K. A. & Simon, H. A. (1980). Verbal reports as data. *Psychological Review*, *87* (3).

Ericsson, K. A. & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data*. Cambridge, MA: MIT Press.

Field, H. & Powell, P. (2001). Public understanding of science versus public understanding of research. *Public Understanding of Science*, *10*, 421-426.

Flanagan, J. G. (1954). The critical incident technique. *Psycholgical Bulletin*, *51*, 327-358.

Flick, L. B. (1993). The meanings of hands-on science. *Journal of Science Teacher Education*, *4* (1), 1-8.

Fraser, B. J. (1979). Development and validation of a test of enquiry skills. *Journal of Research in Science Teaching*, *17* (1), 7-16.

Fraser, B. J., Walberg, H. J., Welch, W. W. & Hattie, J. (1987). Syntheses of educational productivity research. *International Journal of Educational Research*, 145-252.

Frey, A. (2006). Strukturierung und Methoden zur Erfassung von Kompetenz. *Bildung und Erziehung*, *59* (2), 125-166.

Gagné, R. M. (1963). The learning requirements for enquiry. *Journal of Research in Science Teaching*, *1* (2), 144-153.

Gerber, B. L., Cavallo, A. M. L. & Marek, E. A. (2001). Relationship among informal learning enviroments, teaching procedures and scientific reasoning ability. *International Journal of Science Education*, *23* (5), 535-549.

Germann, P. J. (1989). The processes of biological investigations test. *Journal of Research in Science Teaching*, *26* (7), 609-625.

Germann, P. J., Aram, R. & Burke, G. (1998). Identifying patterns and relationships among the responses of seventh-grade students to the science process skill of designing experiments. *Journal of Research in Science Teaching*, *33*, 79-99.

Gräber, G., Hampl, U., Otteni, M., Pälchen, U., Pondorf, P., Ruppert, W. et al. (2006). Biologie Interaktiv: Naturwissenschaftliches Arbeiten. Berlin: Cornelsen.

Gräber, W. & Bolte, C. (1997). *Scientific Literacy - An international symposium*. Kiel: IPN.

Haladyna, T. M. (1994). *Developing and validating multiple-choice test items*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Haladyna, T. M., Downing, S. M. & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurment in Education*, *15* (3), 309-334.

Halldis, A., Blume, R., Eilks, I., Kienast, S., Knobloch, A., Kuck, C. et al. (2005). Chemie Interaktiv: Naturwissenschaftliches Arbeiten. In R. Blume & I. Eilks (Hrsg.), (1. Aufl.). Berlin: Cornelsen.

Hart, C., Mulhall, P., Berry, A., Loughran, J. & Gunstone, R. (2000). What is the Purpose of this Experiment? Or can students learn something from doing experiments? *Journal of Research in Science Teaching*, *37* (7), 655-675.

Hartig, J., Frey, A. & Jude, N. (2007). Validität. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion.* Heidelberg: Springer Medizin Verlag.

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

Häcker, T. (2006). Ein Medium des Wandels in der Lernkultur. In I. Brunner, T. Häcker & F. Winter (Hrsg.), *Das Handbuch Portfolioarbeit - Konzepte, Anregungen, Erfahrungen aus Schule und Lehrerbildung* (S. 15-18). Seelze-Velber: Erhard Friedrich Verlag GmbH.

Heckhausen, J. & Heckhausen, H. (Hrsg.). (2006). *Motivation und Handeln*. Heidelberg: Springer.

Heller, K. A. & Perlet, C. (2000). *Kognitiver Fähigkeitstest für 4.-12. Klassen, Revision (KFT4-12+R).* Göttingen: Hogrefe.

Henke, C. (2006). *Experimentell-naturwissenschaftliche Arbeitsweisen in der Oberstufe. Untersuchung am Beispiel des HIGHSEA-Projekts in Bremerhaven. Studien zum Physik- und Chemielernen*. Berlin: Logos Verlag.

Herbart, J. F. (1806/1965). Allgemeine Pädagogik, aus dem Zweck der Erziehung abgeleitet. In J. F. Herbart (Hrsg.), *Pädagogische Schriften* (Bd. 2, S. 9-155). Düsseldorf: Küpper.

Hoffmann, L., H. & P., M., Lehrke. (1998). *Die IPN-Interessenstudie Physik*. Kiel: Institut für die Pädagogik der Naturwissenschaften.

Hofstein, A. & Lunetta, V. (2004). The Laboratory in science education: Foundations for the twenty-first century. *International Journal of Science Education*, *88* (1), 28-54.

Hofstein, A. & Lunetta, V. N. (1982). The Role of the Laboratory in Science Teaching: Neglected Aspects of Research. *Review of Educational Research*, *52* (1), 201-217.

Hungerford, H. & Walding, H. (1974). *The modification of elementary methods: Students´ concepts concerning science and scientists.* Paper presented at the Annual Meeting of the National Science Teachers Association.

Hurd, P. D. (1958). Science literacy: Its meaning for American schools. *Educational Leadership*, 13-16.

Johnstone, C. J., Bottsford-Miller, N. A. & Thompson, S. J. (2006). *Using the thinking aloud method (cognitive labs) to evaluate test design for students with disabilities and english language learners* (Bericht Nr. 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Jungwirth, E. E. (1970). An evaluation of the attained development of the intellectual skills needed for ´Understanding of the Nature of Scientific Inquiry´ by BSCS pupils in Israel. *Journal of Research in Science Teaching*, *7*, 141-151.

Kail, R. V. & Bisanz, J. (1992). The information-processing perspective on cognitive development in childhood and adolescence. In R. J. Sternberg & C. A. Berg (Hrsg.), *Intellectual Development* (S. 229-260). New York: Cambridge University Press.

Kelava, A. & Moosbrugger, H. (2007). Deskriptivstatistische Evaluation von Items (Itemanalyse) und Testwertverteilungen. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion.* Heidelberg: Springer Medizin Verlag.

Kersting, M. (2003). Augenscheinvalidität (Face Validity). In K. D. Kubinger & R. S. Jäger (Hrsg.), *Schlüsselbegriffe der Psychologischen Diagnostik* (S. 54-55). Weinheim: Beltz, PVU.

Kieren, C. (2004). *Naturwissenschaftliches Arbeiten im Anfangsunterricht - Entwicklung eines Testverfahrens.* Schriftliche Hausarbeit im Rahmen der ersten Staatsprüfung für das Lehramt für die Sekundarstufe I/II. Universität Duisburg-Essen.

Klahr, D. (2000). *Exploring science*. Cambridge, Massachusetts: MIT Press.

Klahr, D. & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Science*, *12*, 1-55.

Klahr, D., Fay, A. L. & Dunbar, K. (1993). Heuristics for scientific experimentation: A developmental study. *Cognitive Psychology*, *25*, 111-146.

Klieme, E., Funke, J., Leutner, D., Reimann, P. & Wirth, J. (2001). Problemlösen als fächerübergreifende Kompetenz. *Zeitschrift für Pädagogik*, *Jg. 2001* (2).

Klieme, E. & Hartig, J. (2008). Kompetenzkonzepte in den Sozialwissenschaften und im Erziehungswissenschaftlichen Diskurs. In P. M., I. Gogolin & H. Krüger (Hrsg.), *Evaluation psychologischer Interventionsmaßnahmen Standards und Kriterien: Ein Handbuch* (Bd. 8). Wiesbaden: VS Verlag für Sozialwissenschaften, GWV Fachverlag GmbH.

Köller, O., Baumert, J. & Neubrand, J. (2000). Epistemologische Überzeugungen und Fachverständnis im Mathematik- und Physikunterricht. In J. Baumert, W. Bos & R. Lehmann (Hrsg.), *TIMSS/III: Dritte Internationale Mathematik- und Naturwissenschaftsstudie - Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn, Mathematische und physikalische Kompetenzen am Ende der gymnasialen Oberstufe* (Bd. 2, S. 229-270). Opladen: Leske + Budrich.

Klos, S., Henke, C., Kieren, C., Walpuski, M. & Sumfleth, E. (2008). Naturwissenschaftliches Experimentieren und chemisches Fachwissen - zwei verschiedene Kompetenzen. *Zeitschrift für Pädagogik*, *54* (3).

KMK. (2005a). *Beschlüsse der Kultusministerkonferenz - Bildungsstandards im Fach Biologie für den mittleren Bildungsabschluss (Beschluss vom 16. Dezember 2004)*. München: Wolters Kluwer.

KMK. (2005b). *Beschlüsse der Kultusministerkonferenz - Bildungsstandards im Fach Chemie für den mittleren Bildungsabschluss (Beschluss vom 16. Dezember 2004)*. München: Wolters Kluwer.

KMK. (2005c). *Beschlüsse der Kultusministerkonferenz - Bildungsstandards im Fach Physik für den mittleren Bildungsabschluss (Beschluss vom 16. Dezember 2004)*. München: Wolters Kluwer.

Krapp, A. (1992). Konzept und Forschungsansätze zur Analyse des Zusammenhangs von Interesse, Lernen und Leistung. In A. Krapp & M. Prenzel (Hrsg.), *Interesse, Lernen, Leistung.* Münster: Aschendorff.

Krapp, A. (1996). Die Bedeutung von Interesse und intrinsischer Motivation für den Erfolg und die Steuerung schulischen Lernens. Theorie und Praxis der Unterrichtsforschung. In G. W. Schnaitmann (Hrsg.), *Theorie und praxis der unterrichtsforschung* (S. 88-111). Donauwörth: Auer.

Krapp, A. (2000). Interest and human development during adolescence: An educational–psychological approach. In J. Heckhausen (Hrsg.), *Motivational psychology of human development* (S. 109–128). London: Elsevier.

Kuhn, D., Amsel, E. & O´Loughlin, M. (1988). *The development of scientific thinking skills*. New York: Academic Press.

Langer, I., Thun, F. Schulz v. & Tausch, R. (2006). *Sich verständlich ausdrücken*. München: Reinhardt.

Laugksch, R. C. (2000). Scientific Literacy: A conseptual overview. *Science Education*, *84*, 71-94.

Lüdtke, O., Robitzsch, A., Trautwein, U. & Köller, O. (2007). Umgang mit fehlenden Werten in der psychologischen Forschung. *Psychologische Rundschau*, *58* (2), 103-117.

Lederman, N. G. (2002). Views of nature of science questionnaire: Toward valid and meaningful assessment of learners´ conceptions of nature of science. *Journal of Research in Science Teaching*, *39* (6), 497-521.

Lederman, N. G. (2007). Nature of science: past, present and future. In S. K. Abell & N. G. Lederman (Hrsg.), *Handbook of research in science teaching* (S. 831-879). Mahwah: Lawrence Erlbaum.

Lederman, N. G., Wade, P. & Bell, L. (1998). Assessing understanding of the nature of science: A historical perspective. In M. W. F. (Hrsg.), *The Nature of Science in Science Education: Rationales and Strategies* (S. 331-350). Dordrecht: Kluwer Academic Publishers.

Lienert, G. A. & Raatz, U. (1998). *Testaufbau und Testanalyse*. Weinheim: Psychologie Verlags Union.

Lord, F. M. (1980). *Applications of item response theory to practical test problems*. Hillsdale, NJ: Lawrence Erlbaum.

Lowrie, T. & Diezmann, C. M. (2007). Solving graphics problems: student performance in junior grades. *Journal of Educational Research*, *100* (6).

Mackworth, N. H. & Bruner, J. S. (1970). How adults and children search and recognize pictures. *Human Development*, *13*, 149-177.

Martin, M. O., Mullis, I. V. S., Gonzales, E. J. & Chrostowski, S. J. (Hrsg.). (2004). *TIMSS 2003 International Science Report*. Boston: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149-174.

McComas, W. F. (Hrsg.). (1998). *The nature of science in science education: Rationales and strategies* (Bd. 5). Dordrecht: Kluwer Academic Publishers.

McComas, W. F. (2005). The principal elements of the nature of science: dispelling the myths. *California Journal of Science Education*, *5* (2), 53-70.

Molenaar, I. W. (1995). Some background for item response theory and the Rasch model. In G. H. Fischer & I. W. Molenaar (Hrsg.), *Rasch Models: Foundations, recent developments, and applications* (S. 3-14). New York: Springer-Verlag.

Moosbrugger, H. (2007). Item-Response-Theorie. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 217-259). Heidelberg: Springer Medizin Verlag.

Mullis, I. V. S., Martin, M. O., Ruddock, G. J., O´Sullivan, C. Y., Arora, A. & Erberber, E. (Hrsg.). (2005). *TIMSS 2007 Assessment Frameworks*. Boston: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Newell, A. & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, New Jersey: Prentice-Hall.

Nielsen, J. (1993). Estimating the number of subjects needed for a thinking aloud test. *International Journal of Human-Computer Studies*, *41*, 385-397.

Nott, M. & Wellington, J. (1998). Eliciting, interpreting and developing teachers´ understanding of the nature of science. *Science and Education*, *7*, 579-594.

NRC. (1996). *National science education standards*. Washington, D.C.: National Academy Press.

OECD (Hrsg.). (1999). *Measuring student knowledge and skills: A new framework for assessment*. Paris: OECD.

OECD (Hrsg.). (2005). *PISA 2003 technical report*. Paris: OECD.

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

OECD. (2007). *PISA 2006. Schulleistungen im internationalen Vergleich. Naturwissenschaftliche Kompetenzen für die Welt von Morgen*. Bielefeld: W. Bertelsmann Verlag.

OECD. (2009). *Pisa 2006 technical report* (Bericht). OECD Publishing.

Oelkers, J. (1997). How to define and justify scientific literacy for everyone. In W. Gräber & C. Bolte (Hrsg.), *Scientific literacy* (S. 87-101). Kiel: IPN.

Orlando, M. & Marshall, G. N. (2002). Differential item functioning in a spanish translation of the PTSD checklist: Detection and evaluation of impact. *Psychological Assessment*, *14* (1), 50-59.

Paulsen, C. A., Best, C., Levine, R., Milne, A. & Ferrara, S. (1999, April). *Lessons learned: Results from try-outs of items in cognitive labs.* Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal.

Paulson, L. & Paulson, P. (1991). What makes a portfolio a portfolio? *Educational Leadership*, *48* (5), 60-63.

Peeck, J. (1994). Wissenserwerb mit darstellenden Bildern. In B. Weidenmann (Hrsg.), *Wissenserwerb mit Bildern* (S. 59-94). Bern: Verlag Hans Huber.

Penner, D. E. & Klahr, D. (1996a). The interaction of domain-specific knowledge and domain-general discovery strategies: A study with sinking objects. *Child Development*, *67*, 2709-2727.

Phan, T. (2007). *Testing levels of competencies in biological experimentation*. Dissertation, Christian-Albrechts-Universität zu Kiel, Kiel.

Piaget, J. (1974). *Der Aufbau der Wirklichkeit beim Kinde*. Linson: Klett-Cotta.

Prenzel, M. (1988). *Die Wirkungsweise von Interesse. Ein Erklärungsversuch aus pädagogischer Sicht*. Opladen: Westdeutscher Verlag.

Prenzel, M., Artelt, C. et al. (Hrsg.). (2007a). *PISA ´06 - Die Ergebnisse der dritten internationalen Vergleichsstudie*. Münster: Waxmann.

Prenzel, M., Artelt, C. et al. (Hrsg.). (2007b). *PISA 2006: Die Ergebnisse der dritten internationalen Vergleichsstudie*. Münster: Waxmann.

Prenzel, M. et al. (Hrsg.). (2008). *PISA 2006 in Deutschland. Die Kompetenzen der Jugendlichen im dritten Ländervergleich*. Münster: Waxmann Verlag.

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

Prenzel, M. & Ringelband, U. (2001). Lernort Labor - neue Initiativen. In U. Ringelband, M. Prenzel & M. Euler (Hrsg.), *Lernort Labor - Initiativen zur Naturwissenschaft. Bildung zwischen Schule, Forschung und Wirtschaft* (S. 7-12). Kiel: Leibniz-Institut für die Pädagogik der Naturwissenschaften (IPN).

Prenzel, M., Rost, J., Senkbeil, H. M., M. & Klopp, A. (2001). Naturwissenschaftliche Grundbildung: Testkonzeption und Ergebnisse. In J. Baumert et al. (Hrsg.), *Pisa 2003. Der Bildungsstand der Jugendlichen in Deutschland - Ergebnisse des zweiten internationalen Vergleichs* (S. 192-248). Opladen: Leske + Budrich.

Prenzel, M., Schütte, K. & Walter, O. (2007). Interesse an den Naturwissenschaften. Münster: Waxmann Verlag.

Rennie, L. J. & Williams, G. F. (2002). Science centers and scientific literacy: Promoting a relationship with science. *Science Education*, *86* (5), 706-726.

Rönnebeck, S., Schöps, K., Prenzel, M. & Hammann, M. (2008). Naturwissenschaftliche Kompetenz im Ländervergleich. In M. Prenzel et al. (Hrsg.), *PISA 2006 in Deutschland. Die Kompetenzen der Jugendlichen im dritten Ländervergleich* (S. 67-94). Münster: Waxmann Verlag.

Rogers, W. T. & Harley, D. (1999). An empirical comparison of three- and four-choice items and tests: Susceptibility to testwiseness and internal consistency reliability. *Educational and Psychological Measurement*, *59*, 234-247.

Rosenstiel, L. v. (2003). *Grundlagen der Organisationspsychologie* (5. Aufl.). Stuttgart: Schäffer-Poeschel Verlag.

Rost, J. (2004a). *Lehrbuch Testtheorie - Testkonstruktion*. Bern: Verlag Hans Huber.

Rost, J. (2004b). Psychometrische Modelle zur Überprüfung von Bildungsstandards anhand von Kompetenzmodellen. *Zeitschrift für Pädagogik*, *50* (5), 662-678.

Roth, W.-M. & Roychoudhury, A. (1993). The development of science process skills in authentic contexts. *Journal of Research in Science Teaching*, *30* (2), 127-152.

Ruffman, T., Perner, J., Olson, D. R. & Doherty, M. (1993). Reflecting on scientific thinking: Children´s understanding of the hypothesis-evidence-relation. *Child Development*, *64*, 1617-1636.

Ruiz-Primo, M. A., Li., M. & Shavelson, R. J. (2001). *Looking into students' science notebooks: What do teachers do with them?* (Bericht).

Ruiz-Primo, M. A. & Shavelson, R. J. (1996). Rhetoric and reality in science performance assessments: An update. *Journal of Research in Science Teaching*, *33* (10), 1045-1063.

Rutherford, F. J. (1993). Hands-on: A means to an end. *2061 Today*, *3* (1), 5.

Sadler, P. M. (2000). The relevance of multiple-choice tests in assessing science. In J. J. Mintzes, H. Wandersee & J. D. Novak (Hrsg.), *Assessing science understanding* (S. 249-278). San Diego, CA: Academic Press.

Schauble, L. (1990). Belief revision in children: The role of prior knowledge and strategies for generating evidence. *Journal of Experimental Child Psychology*, *49*, 31-57.

Schermelleh-Engel, K. & Werner, C. (2007). Methoden der Reliabilitätsbestimmung. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion.* Heidelberg: Springer Medizin Verlag.

Schiefele, U. & Schreyer, I. (1994). Intrinische Lernmotivation und Lernen. Ein Überblick zu Ergebnissen der Forschung. *Zeitschrift für Pädagogische Psychologie*, *8*, 1-13.

Schmitt, A. P., Holland, P. W. & Dorans, N. J. (1993). Evaluating hypotheses about differential item functioning. In P. W. Holland & H. Wainer (Hrsg.), *Differential item functioning* (S. 281-319). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Schnotz, W. (1994). Wissenserwerb mit logischen Bildern. In B. Weidenmann (Hrsg.), *Wissenserwerb mit Bildern* (S. 95-148). Bern: Verlag Hans Huber.

Schnotz, W. (1997). Wissenserwerb mit Diagrammen und Texten. In L. J. Issing & P. Klimsa (Hrsg.), *Information und Lernen mit Multimedia* (S. 85-106). Weinheim: Beltz PVU.

Schrader, F.-W. & Helmke, A. (2001). Alltägliche Leistungsbeurteilung durch Lehrer. In F. E. Weinert (Hrsg.), *Leistungsmessung in Schulen* (S. 45-58). Weinheim: Beltz.

Schütte, K., Frenzel, A. C., Asseburg, R. & Pekrun, R. (2007). Schülermerkmale, naturwissenschaftliche Kompetenz und Berufserwartung. In M. Prenzel et al. (Hrsg.), *PISA ´06 - Die Ergebnisse der dritten internationalen Vergleichsstudie* (S. 125-146). Münster: Waxmann.

Schuler, H. (2000). *Psychologische Personalauswahl*. Göttingen: Verlag für Angewandte Psychologie.

Schuler, H. (2006). Noten als Prädiktoren von Studien- und Berufserfolg. In D. H. Rost (Hrsg.), *Handwörterbuch Pädagogische Psychologie* (S. 535-541). Weinheim: Psychologie Verlags Union.

Schwartz, R. S., Lederman, N. G. & Crawford, B. A. (2004). Developing views of nature of science in an authentic context: An explicit approach to bridging the gap between nature of science and scientific inquiry. *Science Education*, *88*, 610-645.

Schweizer, K. (2006). *Leistung und Leistungsdiagnostik*. Heidelberg: Springer Medizin Verlag.

Scientific Literacy Center. (1967). *Wisconsin inventory of science processes* (Bericht). The Regeants of the University of Wisconsin.

Seidel, T., Prenzel, M., Wittwer, J. & Schwindt, K. (2007). Unterricht in den Naturwissenschaften. In M. Prenzel et al. (Hrsg.), *PISA ´06 - Die Ergebnisse der dritten internationalen Vergleichsstudie* (S. 147-180). Münster: Waxmann.

Shamos, M. H. (2002). Durch Prozesse ein Bewusstsein für die Naturwissenschaften entwickeln. In W. Gräber, P. Nentwig, T. Koballa & R. Evans (Hrsg.), *Scientific literacy* (S. 45-68). Opladen: Leske + Budrich.

Shavelson, R. J. & Ruiz-Primo, M. A. (1999). On the psychometrics of assessing science understanding. In J. Mintzes, J. H. Wandersee & J. D. Novak (Hrsg.), *Assessing science understanding* (S. 304-341). San Diego: Academic Press.

Sidick, J. T., Barrett, G. V. & Doverspike, D. (1994). Three-alternative multiple choice tests: An attractive option. *Personnel Psychology*, *47*, 829-835.

Siegler, R. S. & Alibali, M. W. (2005). *Children´s thinking*. Upper Saddle River, NJ: Prentice Hall.

Smith, E. R. & Mackie, D. M. (2007). *Social psychology*. New York, NY: Psychology Press.

Smith, R. M. (1995, April). *Using item mean squares to evaluate fit to the rasch model.* Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA.

Smith, S. (1979). *Ideas of the great educators*. New York: Barnes and Noble.

Snijkers, G. J. (2002). *Cognitive laboratory experiences : On pre-testing computerised questionnaires and data quality*. Unveröffentlichte Dissertation, Utrecht University.

Sodian, B., Thoermer, C. & Kircher, E. (2002). Vermittlung von Wissenschaftsverständnis in der Grundschule. In *Zeitschrift der pädagogik* (Bd. 45, S. 192-206). Weinheim, Basel: BELTZ.

Solomon, J. (1991). Teaching about the nature of science in the british national curriculum. *Science Education*, *75* (1), 95-103.

Songer, N. B. & Linn, M. C. (1991). How do students´ views of science influence knowledge integration? *Journal of Research in Science Teaching*, *28* (9), 761-784.

Sternberg, R. J. (2003). *Cognitive Psychology*. Belmont, CA: Wadsworth/Thomson Learning.

Tamir, P., Nussinovitz, R. & Fiedler, Y. (1982). The design and use of a practical tests assessment inventory. *Journal of Biological Education*, *16* (1), 42-50.

Tannenbaum, R. S. (1969). *The development of the ´Test of science processes´.* Paper presented at the 42nd meeting of the National Association for Research in Science Teaching, Pasadena, California.

Thomas, G. & Durant, J. (1987). Why should we promote the public understanding of science? In M. Shortland (Hrsg.), *Scientific literacy papers.* Oxford, UK: Department for External Studies, University of Oxford.

Tobin, K. G. & Capie, W. (1982). Development and validation of a group test of integrated science processes. *Journal of Research in Science Teaching*, *19* (2), 133-141.

Tsai, C. C. (1998). An analysis of scientific epistemological beliefs and learning orientations of Taiwanese eighth graders. *Science Education*, *82*, 473-489.

Tsai, C. C. (1999). "Laboratory exercises help me memorize the scientific truths": A study of eighth graders' scientific epistemological views and learning in laboratory activities. *Science Education*, *83* (6), 654-674.

University of York Science Education Group. (2006). *Twenty first century science. GCSE additional applied science module 3 textbook: Scientific detection*. Oxford: Oxford University Press.

Vroom, V. (1964). *Work and motivation*. New York: Wiley.

Wainer, H. (1980). A test of graphicacy in children. *Applied Psychological Measurement*, *4* (3), 331-340.

Wainer, H. (1992). Understanding graphs and tables. *Educational Researcher*, *21* (4), 14-23.

Walpuski, M. (2006). *Optimierung von experimenteller Kleingruppenarbeit durch Strukturierungshilfen und Feedback. Eine empirische Studie. Studien zum Physik- und Chemielernen*. Berlin: Logos Verlag.

Walter, O. (2005). *Kompetenzmessung in den PISA-Studien*. Lengerich: Pabst Science Publishers.

Wang, W. C. (1995). *Implementation and application of the multidimensional random coefficients multinomial logit*. Unveröffentlichte Dissertation, University of Berkeley, California.

Waterman, M. A. (1982). *College biology students beliefs about scientific knowledge: Foundation for study of epistemological commitments in conceptual change*. Unveröffentlichte Dissertation, Cornell University, Ithaca, NY.

Watson, F. & Cohen, I. B. (1952). *General education in science*. Cambridge, MA: Harvard University Press.

Weidenmann, B. (2001). Lernen mit Medien. In A. Krapp & B. Weidenmann (Hrsg.), *Pädagogische Psychologie* (4., vollständig überarbeitete Auflage Aufl.). Weinheim: Beltz PVU.

Weiner, B. (1986). *An attributional theory of motivation and emotion*. New York: Springer.

Weinert, F. E. (1999). *Konzepte der Kompetenz. Unveröffentlichtes Gutachten zum OECD-Projekt „Definition and Selection of Competencies:Theoretical and Conceptual Foundations (DeSeCo)“* (Bericht). München: Max-Planck-Institut für psychologische Forschung.

Weinert, F. E. (2001). Concept of Competence: A Conceptual Clarification. In D. S. Rychen & L. H. Salganik (Hrsg.), *Defining and selecting key competencies* (S. 46-65). Seattle: Hogrefe & Huber Publishers.

Welch, W. W. (1969). *Welch science process inventory: Form D* (Bericht). Minneapolis, MN.

Welch, W. W. & Pella, M. O. (1967). The development of an instrument for inventory knowledge of the process of science. *Journal of Research in Science Teaching*, *5* (1).

Wertheimer, M. (1938). *Laws of organization in perceptual forms in a source book for Gestalt Psychology*. London: Routledge.

Wild, K.-P., Krapp, A. & Winteler, A. (1992). Die Bedeutung von Lernstrategien zur Erklärung des Einflusses von Studieninteresse auf Lernleistungen. In A. Krapp & M. Prenzel (Hrsg.), *Interesse, Lernen, Leistung*. Münster: Aschendorff.

Wilson, M. (2003). On choosing a model for measuring. *Methods of Psychological Research Online*, *8* (3), 1-22. (Internet: http://www.mpr-online.de)

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Wright, B. D., Mead, R. J. & Bell, S. R. (o. J.). BICAL: Calibrating items with the Rasch model [Software-Handbuch].

Wu, A. R. J. . W. M., M. L. (1997). *ConQuest - Generalised item response modelling software, draft release 2*. Camberwell: Australian Council for Educational Research.

Zimmermann, C. (2007). The development of scientific thinking skills in elementary and middle school. *Developmental Review*, *27*, 172-223.

Lebenslauf

Persönliche Daten

Inga Glug
geboren am 03.11.1976
in Flensburg
unverheiratet

Ausbildung

1983-1987:	Grundschule Oeversee
1987-1996:	Auguste-Viktoria-Schule Flensburg, Abschluss: allgemeine Hochschulreife
1996 – 1999:	Ausbildung zur Industriekauffrau bei der Stora Spezialpapiere GmbH in Flensburg
10 1999 – 09 2004:	Psychologie-Studium an der Christian-Albrechts-Universität zu Kiel mit Nebenfach Pädagogik, Abschluss Diplom
10 2001 – 01 2002:	Forschungspraktikum im Bereich Arbeits-, Organisations- und Marktpsychologie zum Thema „Motivation und Leistung“
07 2002 – 10 2002:	Praktikum in der Unternehmensberatung Moldzio & Partner - Institut für Personalauswahl
seit 12 2005:	Dissertation zum Thema „Entwicklung und Validierung eines Multiple-Choice-Tests zur Erfassung prozessbezogener naturwissenschaftlicher Grundbildung“

Berufliche Tätigkeit

10 2002 – 09 2004:	projektbezogene Mitarbeit bei Moldzio & Partner – Institut für Personalauswahl
08 2004 – 11 2008:	wissenschaftliche Mitarbeiterin in der Abteilung für Physikdidaktik des IPN
12 2008 – heute:	wissenschaftliche Mitarbeiterin in der Abteilung für Erziehungswissenschaft des IPN