Modellierung und Messung experimenteller Kompetenz

Analyse eines large-scale Experimentiertests

Inauguraldissertation

zur Erlangung der Würde eines Doktors der Philosophie
vorgelegt der Philosophisch-Naturwissenschaftlichen Fakultät
der Universität Basel

von

Christoph Gut-Glanzmann
aus Affoltern am Albis, Zürich

Basel, 2012

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät auf Antrag von

Prof. Dr. Christoph Bruder (Fakultätsverantwortlicher)
Prof. Dr. Peter Labudde (Referent)
Prof. Dr. Horst Schecker (Korreferent), Universität Bremen

Basel, den 22. Mai 2012

Der Dekan
Prof. Dr. Martin Spiess

Die Dissertation ist beim Logos Verlag Berlin (www.logos-verlag.de) in der Schriftenreihe *Studien zum Physik- und Chemielernen* als gedruckte Ausgabe erhältlich (ISBN 978-3-8325-3213-0).

Für Jutta, Alexander und Darius

Inhaltsverzeichnis

Kapitel 1

Einleitung

Die Veröffentlichung der ersten PISA-Ergebnisse zu Beginn dieses Jahrtausends markiert eine Zäsur im Schweizer Bildungsdiskurs wie auch in den Diskursen vieler anderer, vorwiegend europäischen Länder. Nicht nur wird in öffentlichen und politischen Schuldebatten seither regelmässig auf den internationalen Vergleichstest verwiesen, auch die fachdidaktische Forschung erhielt durch PISA wichtige Impulse. Nach PISA wurden in den deutschsprachigen Ländern erstmals Bildungsstandards formuliert, die sich auf Kompetenzdefinitionen abstützen. Es wurden Projekte in Angriff genommen, die Kompetenzdefinitionen in Kompetenzmodellen zusammenzufassen, und man ist daran, Testinstrumente zu entwickeln, um Standards zu überprüfen und Kompetenzmodelle zu validieren. In allen deutschsprachigen, aber z. B. auch in den skandinavischen Ländern sind Reformen eingeleitet worden, welche die Bildungssysteme langfristig zu mehr Output-Orientierung hinführen werden.

In der Schweiz wurde dieser Prozess von der Eidgenössischen Konferenz der kantonalen Erziehungsdirektoren (EDK) mit dem Entscheid ausgelöst, die kantonalen Bildungssysteme mit sprachregionalen Lehrplänen[1], verbindlichen Standards und einem nationalen Bildungsmonitoring zu harmonisieren. Die Reformen wurde vom Schweizer Stimmvolk durch die Annahme des revidierten Bildungsartikels 2006 in der Bundesverfassung verankert und mit Inkrafttreten des HarmoS-Schulkonkordats 2009 bekräftigt. 2004 erteilte die EDK dem Konsortium HarmoS Naturwissenschaften unter der Co-Leitung von Prof. Dr. Peter Labudde (PH FHNW, ehemals PHBern) und Prof. Dr. Marco Adamina (PHBern) den Auftrag, ein Kompetenzmodell für den naturwissenschaftlichen Unterricht in der obligatorischen Schule zu entwickeln und zu validieren sowie modellbasierte Vorschläge für

[1] Für die Deutschschweiz wird derzeit der *Lehrplan 21* ausgearbeitet. Er betrifft alle 21 Kantone, in welchen Deutsch eine offizielle Unterrichtssprache ist. Für die französischsprachige Schweiz gilt seit 2010 der *Plan d'études romand*. Das italienischsprachige Tessin behält seinen kantonalen Lehrplan.

Basisstandards am Ende des 2., 6. und 9. Schuljahres[2] auszuarbeiten. Der Auftrag wurde 2008 mit dem wissenschaftlichen Schlussbericht «HarmoS Naturwissenschaften: Kompetenzmodell und Vorschläge» zuhanden der EDK vom Konsortium abgeschlossen. Nach einer breit geführten Vernehmlassung wurden die Standards überarbeitet und als verbindliche Grundkompetenzen 2011 von der EDK in Kraft gesetzt (EDK, 2011). In der Deutschschweiz werden zudem die Grundkompetenzen mittelfristig durch den oben genannten Lehrplan 21 ergänzt.

Anhaltspunkte für die Standardvorschläge lieferten dem Konsortium HarmoS Naturwissenschaften Leistungstests, die vom Konsortium entwickelt und zwischen April 2007 und Mai 2008 schweizweit in jeweils mindestens zwei Sprachregionen durchgeführt wurden. Auf jeder der drei Zielstufen wurden mit einem Papier-und-Bleistifttest und mit einem Experimentiertest die naturwissenschaftlichen Kompetenzen der Schülerinnen und Schüler evaluiert. Die Tests wurden zudem dazu verwendet, das HarmoS-Kompetenzmodell zu validieren (Gut & Labudde, 2010; Ramseier, Labudde & Adamina, 2011).

Mit dem HarmoS-Projekt wurden in der Schweiz verschiedene fachdidaktische Diskurse initiiert, die in gewissen europäischen und in den angelsächsischen Ländern bereits seit Jahrzehnten geführt werden (Waddington, Nentwig & Schanze, 2007). Hierzu gehört der *Modelldiskurs*. Die Frage, welche Kompetenzen auf welche Weise modelliert werden sollen, stand bei HarmoS am Anfang der ganzen Entwicklungsarbeit (HarmoS, 2008). Mit Abschluss des Projekts und der vorläufigen Validierung des HarmoS-Kompetenzmodells findet dieser Diskurs jedoch vorerst ein Ende. Neue Impulse zum Modelldiskurs werden u. a. im Rahmen eines nationalen Bildungsmonitorings erwartet. Mehr Resonanz in der Öffentlichkeit und der Tagespolitik erreicht der *Standarddiskurs*. Mit den EDK-Grundkompetenzen sind die Standards zwar gegeben, ausstehend sind jedoch immer noch deren Implementation in der Praxis und deren Überprüfung durch ein nationales Bildungsmonitoring. Der Standarddiskurs wird daher auch weiterhin nicht nur in der Öffentlichkeit grosse Aufmerksamkeit erhalten.

Neben dem Modell- und Standarddiskurs hat sich in der Naturwissenschaftsdidaktik auch ein neuer *Assessmentdiskurs* etabliert. Besonders in Deutschland sind im Hinblick auf die erste nationale Evaluation der KMK-Regelstandards im Jahr 2012 grosse Anstrengun-

[2]Die hier verwandte Nummerierung der Schuljahre entspricht nicht der neuen Zählung der Konferenz der kantonalen Erziehungsdirektoren (EDK). Gemäss dem so genannten HarmoS-Schulkonkordat, das die meisten Deutschschweizer Kantone angenommen haben, werden die zwei obligatorischen Kindergartenjahre neu zur offiziellen Schulzeit gerechnet. Dies auch unter der Perspektive einer möglichen Einführung von Grund- oder Basisstufen in den Kantonen. Die Grundkompetenzen der EDK beziehen sich daher auf das "4.", "8." und "11." Schuljahr. Da die Reformen jedoch in nächster Zukunft kaum umgesetzt werden, zieht es der Autor zwecks besserer Vergleichbarkeit mit dem Ausland vor, konsequent die alte Nummerierung der Schuljahre zu verwenden.

gen bei der Messung von naturwissenschaftlichen Kompetenzen unternommen worden. In der Schweiz haben, von der zyklisch wiederkehrenden PISA-Berichterstattung abgesehen, vor allem die large-scale Assessments des Konsortiums HarmoS Naturwissenschaften neue Impulse in Forschung und Lehre gebracht. Dies gilt im Besonderen für die Experimentiertests. Bereits gibt es Pläne und erste Arbeiten für Vergleichstests im Bildungsraum Nordwestschweiz, die auf den Arbeiten von HarmoS aufbauen. Ebenfalls von der EDK initiiert und im HarmoS-Schulkonkordat gesetzlich verankert ist ein nationales Bildungsmonitoring. Der Assessmentdiskurs wird deshalb gegenüber dem derzeit stark geführten Standarddiskurs langfristig aufholen und deutlich an Wichtigkeit zulegen. Man wird sich vermehrt die Frage stellen müssen, mit welchen Tests die Grundkompetenzen und die Ziele des Lehrplans 21 sowohl im Unterricht als auch im Rahmen nationaler Leistungstests erfasst werden können. Man wird notwendigerweise mehr Entwicklungs- und Forschungsarbeit in die Konstruktion von Testaufgaben investieren müssen, will man die Wende zu einem Output-orientierten Bildungssystem aus wissenschaftlicher Sicht ernst nehmen. Im Rahmen dieser Arbeiten wird auch der Modelldiskurs wieder aufgenommen werden müssen.

Mit der vorliegenden Dissertation wird der Assessmentdiskurs weitergeführt. Thema der Arbeit ist der im Kapitel 5 vorgestellte HarmoS-Experimentiertest, der im April und Mai 2008 mit rund 1500 Schülerinnen und Schülern des 6. und 9. Schuljahres in der deutsch-, französisch- und italienischsprachigen Schweiz durchgeführt wurde.

Large-scale Experimentiertests sind ressourcenintensive und heikle Testinstrumente. Entsprechend selten wird diese Testart verwendet. Im Sinne von nationalen Assessments standen für die Testentwicklung und -auswertung bei HarmoS drei Beispiele zur Verfügung (Britton & Schneider, 2007): Die weltweit ersten nationalen Experimentiertests APU (Assessment of Performance Unit), welche in England, Wales und Nordirland in den Jahren 1980 bis 1984 durchgeführt wurden (Gott & Duggan, 1995), die nationalen Leistungserhebungen NAEP (National Assessment of Education Progress) der Vereinigten Staaten, in welchen seit 1986 regelmässig Experimentieraufgaben verwendet und ausführlich ausgewertet werden (Shavelson, Ruiz-Primo & Wiley, 1999), und der TIMSS-Experimentiertest (Third International Mathematics and Science Studies), mit welchem 1994/95 erstmals ein Ländervergleich – mit Beteiligung der Schweiz – gemacht wurde (TIMSS, 1997).

Im Rahmen des durch PISA ausgelösten Kompetenzdiskurses sind Experimentiertests nun erneut im Fokus fachdidaktischer Forschung. Dabei interessieren in erster Linie die Fragen, ob diese Testart als reliables und valides Messinstrument experimenteller Kompetenz taugt und inwiefern Hands-on-Tests durch andere weniger aufwändige Tests ersetzt werden können. In einem zweiten Schritt geht es darum, mit Hilfe solcher Tests Hinweise über die experimentellen Fähigkeiten von Schülerpopulationen zu erhalten und die neu

eingeführten Standards zu überprüfen.
Die vorliegende Studie nimmt diese Fragen mit einer breit gefächerten Analyse des HarmoS-Experimentiertests auf, die im Analyseteil ab S. 105) vorgestellt wird. Die Analyse baut auf der theoretischen Erörterung der Begriffe der experimentellen Kompetenz und der Itemschwierigkeit auf (Theorieteil ab S. 13). Eine Übersicht über die Arbeit folgt im nächsten Kapitel.

Kapitel 2

Übersicht und Forschungsfragen

2.1 Inhaltliche Übersicht

Übersicht. Die Dissertation ist in drei Teile gegliedert: in einen Teil zur Theorie der experimentellen Kompetenz, einen Analyseteil zum HarmoS-Experimentiertest und in die Diskussion.

Der erste Teil "*Zur Theorie der experimentellen Kompetenz*" ist der Erörterung der Begriffe experimentelle Kompetenz und Itemschwierigkeit gewidmet. Im Kapitel 3 wird die experimentelle Kompetenz anhand dreier wissenschaftlicher Diskurse zu Standards, Modellen und Assessments diskutiert. Aus der Analyse von aktuellen Bildungsstandards werden u. a. fünf verallgemeinerte Progressionsdimensionen für die experimentelle Kompetenz hergeleitet. Diese dem Standarddiskurs entstammenden Progressionsdimensionen, namentlich der *Aufgabenumfang* $[A]$, die *Problemkomplexität* $[P]$, die *Prozessqualität* $[Q]$, der *Transferumfang* $[T]$ und die *Eigenständigkeit* $[E]$, gilt es auf den Modell- und Assessmentdiskurs adäquat zu übersetzen. Im Falle der im Kapitel 7 durchgeführten Analyse der Itemschwierigkeit im HarmoS-Experimentiertest wird auf diese Progressionsdimensionen zurückgegriffen (cf. Abb. 2.1 auf S. 7).

Im Zentrum des Kapitels 4 steht die Modellierung der Itemschwierigkeit bei einem large-scale Experimentiertest. Um Faktoren für Itemschwierigkeit zu erfassen, wird der Aufgabenlösungs- und Bewertungsprozess beim Testen in *vier unabhängig voneinander gedachte Arbeitsschritte* zerlegt: ‹Aufgabe erfassen›, ‹Problem lösen›, ‹Antwort geben› und ‹Lösung kodieren›. Mit diesem Modell werden Schwierigkeiten, die sich auf die zu messende experimentelle Kompetenz beziehen und mit so genannten *kompetenzrelevanten Itemmerkmalen* verknüpft werden, von Schwierigkeiten abgegrenzt, die sich auf nicht kompetenzspezifische Fähigkeiten beziehen und mit so genannten *kompetenzirrelevanten Itemmerkmalen* verknüpft werden. Erstere werden mit den Arbeitsschritten ‹Problem lösen› und ‹Lösung kodieren›assoziiert, letztere mit den Schritten ‹Aufgabe erfassen› und

‹Antwort geben› (cf. Abb. 2.1).
Im empirischen zweiten Teil "*Analyse eines large-scale Experimentiertests*" werden, aufbauend auf den theoretischen Grundlagen des ersten Teils, vier Analysen des HarmoS-Experimentiertests vorgestellt. In den ersten zwei Analysen (Kap. 6 und 7) wird die Dimensionalität des Experimentiertests und die Itemschwiergkeit modelliert untersucht. In den zwei letzten Analysen werden die individuellen Fähigkeitsparameter mit weiteren Personenvariablen verknüpft, die mit einem separaten Fragebogen erhoben wurden. Die Analysen geben Aufschluss über die Verteilung der experimentellen Kompetenzen in verschiedenen Schülerstichproben. Untersucht werden Leistungsdifferenzen zwischen den Schulstufen, Geschlechtern, Sprachregionen und Anforderungsstufen.

Die Hauptarbeit der Dissertation besteht in der Post hoc-Analyse der Itemschwierigkeit des HarmoS-Experimentiertests (cf. Kap. 7). In der Abbildung 2.1 ist der hierfür entwickelte Merkmalkatalog schematisch dargestellt. Der Katalog nimmt die Grundstruktur der oben erwähnten vier Arbeitsschritte auf. Die kompetenzrelevanten Itemmerkmale werden aufgrund der fünf Progressionsdimensionen entwickelt. Die Festlegung von kompetenzirrelevanten Itemmerkmalen erfolgt mit Hilfe einer Analyse der Informationsformate der gedruckten Aufgabenstellung. Diese Analysemethode wird im Kapitel 7 vorgestellt.
Die abschliessende Diskussion der Resultate folgt in den Kapiteln 10, 11 und 12 mit einer Zusammenfassung der Ergebnisse, einem kritischen Rückblick auf den Experimentiertest und dessen Analyse und einem Ausblick auf anstehende Forschungsfragen.

Kennzeichnungen mit Klammern. Für die Analysen werden verschiedene Sätze von Merkmalen unterschieden. Bei der Analyse der Itemschwierigkeit sind dies Itemmerkmale, bei den anderen Analysen handelt es sich um Personenmerkmale. In beiden Fällen werden wir Merkmale durch geschweifte Klammern {...} kennzeichnen. Die für die Itemmerkmale wesentlichen Progressionsdimensionen werden wir mit Buchstaben in eckigen Klammern abkürzen: Als Beispiel sei der Aufgabenumfang $[A]$ genannt. Kompetenzbeschreibungen erfolgen anhand der Differenzierung von Prozessen oder Teilprozessen. Experimentelle Teilprozesse werden wir in diesem Zusammenhang zur besseren Lesbarkeit in Klammern des Typs «...» setzen, die vier Arbeitsschritte einer Aufgabenbearbeitung durch Klammern des Typs ‹...› (cf. Abb. 2.1).

CD-ROM. Zu den im Abschnitt 4.2.1 und in den Kapiteln 6, 7, 8 und 9 erwähnten statistischen Analysen wurde eine CD-ROM mit Daten und Auswertungen erstellt, die beim Autor nachgefragt werden kann. Ausgewählte Ergebnisse sind in den Anhängen C, D.2 und E abgedruckt.

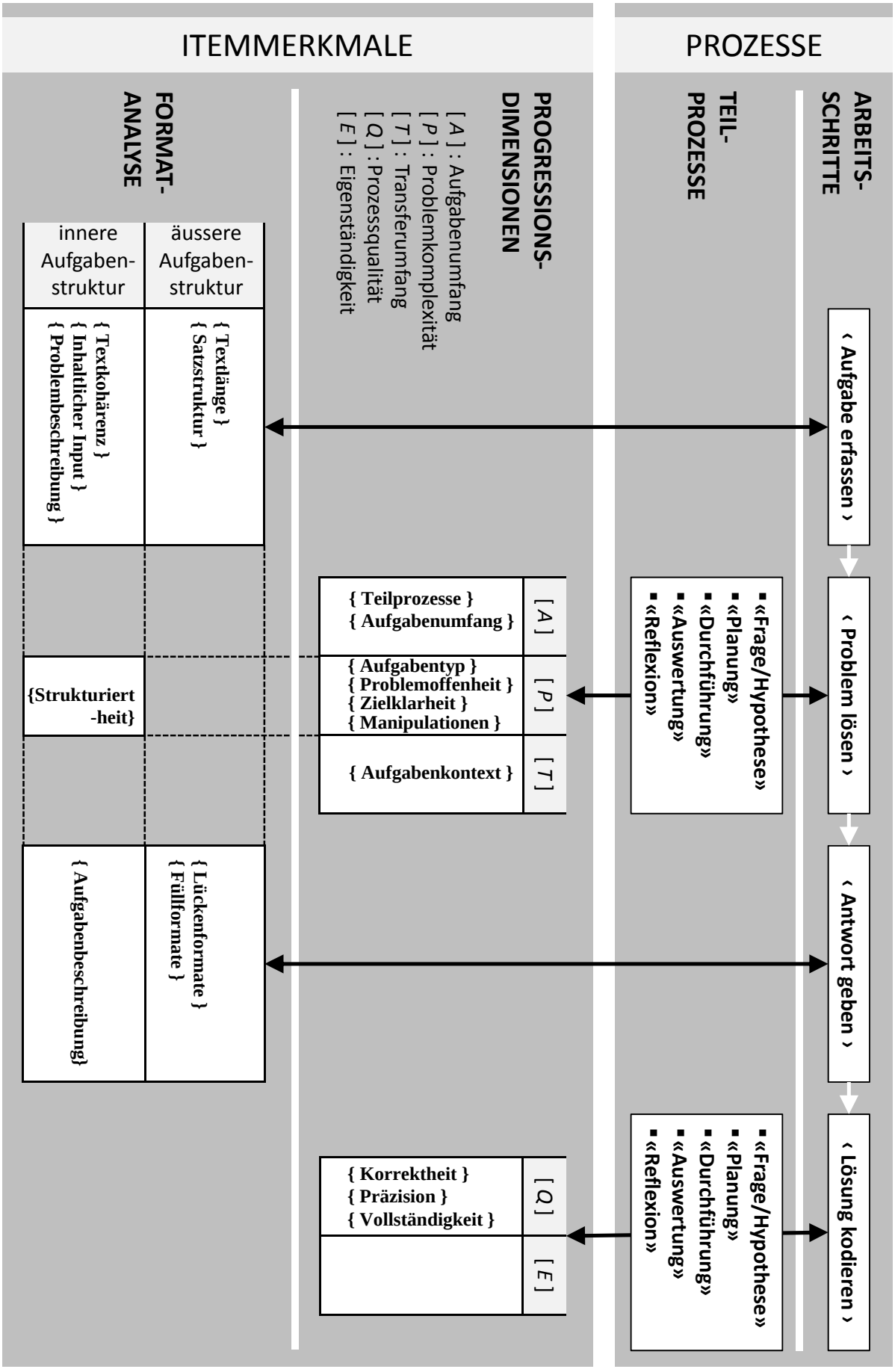


Abbildung 2.1 – Analyse der Itemschwierigkeit des HarmoS-Experimentiertests: Schematische Übersicht über die Struktur des verwandten Merkmalkatalogs.

2.2 Analyse des HarmoS-Experimentiertests: Forschungsfragen

Analyse-Kapitel. Mit den oben skizzierten vier Analysen des HarmoS-Experimentiertests sollen die anschliessend formulierten Fragen beantwortet werden. Die Fragen werden in den vier Analyse-Kapiteln "Test-Analyse" (Kap. 6), "Item-Test-Analysen" (Kap. 7), "Personen-Analysen" (Kap. 8) und "Personen-Test-Analysen" (Kap. 9) beantwortet. Alle Antworten werden zudem im Diskussionsteil zusammengefasst (Kap. 10).

Kennzeichnung von Itemstichproben und Personenstichproben. Nicht alle Auswertungen erfolgen anhand derselben Stichprobe. Aufgrund der besonderen Datenlagen oder der Forschungsfragen beziehen sich die einzelnen Analyse auf unterschiedliche Teiltests des HarmoS-Experimentiertests. Dies betrifft sowohl die getesteten Personen als auch die ausgewerteten Items. Um die unterschiedlichen Item- und Personenstichproben kenntlich zu machen, werden im Text symbolische Bezeichnungen verwendet, die sich – wenn nicht anders deklariert – grundsätzlich auf die Itemstichprobe beziehen. Mit der Verwendung der Klammern $\langle ... \rangle$ werden einzelne Items oder Menge von Items gekennzeichnet. Dies trifft auch auf ganze Tests zu. Folgende Kennzeichnungen von Teiltests werden häufig verwandt:

$\langle$E08|69df$\rangle$ Teiltest des HarmoS-Experimentiertests 2008 mit allen Experimentieritems, die im 6. oder 9. Schuljahr in der deutsch- oder französischsprachigen Schweiz eingesetzt wurden.
⇒ Kapitel "Personen-Test-Analysen"

$\langle$E08|69d$\rangle$ Teiltest des HarmoS-Experimentiertests 2008 mit allen deutschsprachigen Items, die im 6. oder 9. Schuljahr verwendet wurden.
⇒ Kapitel "Test-Analysen" und "Item-Test-Analysen"

$\langle$Q08|69dfi$\rangle$ Fragebogen mit allen Items, die im 6. oder 9. Schuljahr in der deutsch-, französisch- und italienischsprachigen Schweiz eingesetzt wurden.
⇒ Kapitel "Personen-Analysen"

$\langle$Q08|9df$\rangle$ Fragebogen mit allen Items, die im 9. Schuljahr in der deutsch- und französischsprachigen Schweiz eingesetzt wurden. Die Daten des Fragebogens werden im Kapitel 9 mit den Testergebnissen der entsprechenden Personen verknüpft.
⇒ Kapitel "Personen-Test-Analysen"

2.2.1 Test-Analysen: Dimensionalität des HarmoS-Experimentiertests

Für die Entwicklung des Experimentiertests wurden bei HarmoS verschiedene Teilaspekte experimentellen Handelns unterschieden, die unter dem übergeordneten Handlungsaspekt «Fragen und untersuchen» zusammengefasst wurden. Die Teilaspekte referieren sowohl auf unterschiedliche Teilprozesse beim Experimentieren (e. g. Planung, Durchführung und Reflexion eines Experimentes) als auch auf verschiedene Aufgabentypen (e. g. Observieren/Klassifizieren versus Untersuchen). Zwar wurde der Test nicht für den Zweck konstruiert, verschiedene Dimensionen zu unterscheiden, die Frage bezüglich der Dimensionalität stellt sich aber trotzdem.

1.1. Unterscheidet der HarmoS-Experimentiertest bezüglich der Teilprozesse zwischen verschiedenen Dimensionen?

1.2. Unterscheidet der HarmoS-Experimentiertest bezüglich der Aufgabentypen zwischen verschiedenen Dimensionen?

1.3. Unterscheidet der HarmoS-Experimentiertest bezüglich der Themenbereiche zwischen verschiedenen Dimensionen?

2.2.2 Item-Test-Analysen: Post hoc-Erklärung der Itemschwierigkeit

Beim HarmoS-Experimentiertest wurden keine A priori-Kompetenzstufen festgelegt. Mit dem Test wurde vorwiegend das Ziel verfolgt, das Niveau möglicher Basisstandards für die verschiedenen Teilaspekte zu erfassen. Die Erklärung von unterschiedlichen Testleistungen bedingt somit eine Post hoc-Analyse der Itemschwierigkeit. Dies eröffnet den Raum für folgende Forschungsfragen.

2.1. Lässt sich die Itemschwierigkeit des HarmoS-Experimentiertest anhand eines Systems von Itemmerkmalen modellieren?

2.2. Können schwierigkeitsinduzierende Anforderungen eruiert werden, die für die experimentelle Kompetenz relevant sind? Inwieweit lassen sich Einflüsse der Korrekturen erkennen, die bei der Aufgabenentwicklung verwendet wurden?

2.3. Können darüber hinaus schwierigkeitsinduzierende Anforderungen eruiert werden, die für die experimentelle Kompetenz nicht relevant sind? Inwieweit lassen sich Einflüsse der Korrekturen erkennen, die bei der Aufgabenentwicklung verwendet wurden?

2.4. Inwieweit eignet sich der HarmoS-Experimentiertest als Messinstrument experimenteller Kompetenz? Inwieweit misst er kompetenzrelevante und -irrelevante Fähigkeiten?

2.2.3 Personen-Analysen: Interessenverteilung in der Personenstichprobe

Die mit dem Fragebogen erhobenen persönlichen Daten der Testpersonen ermöglichen eine vertiefte Analyse der Personenstichprobe. Die Kombination der Daten des Fragebogens mit den Testresultaten des Experimentiertests erlauben letztlich weitere Zusammenhänge zu analysieren.

3.1. Welche Unterschiede bestehen in der Ausprägung des Interesses an Naturwissenschaften (Fach- und Sachinteressen) zwischen den Geschlechtern, Schulstufen, Sprachregionen und schulischen Anforderungsniveaus?

2.2.4 Personen-Test-Analysen: Kompetenzverteilung in der Personenstichprobe

Mit HarmoS wurde erstmals in der Schweiz ein large-scale Experimentiertest auf verschiedenen Schulstufen und in verschiedenen Sprachregionen durchgeführt. Von grossem Interesse sind daher Ergebnisse zur Leistungsfähigkeit verschiedener Schulsysteme, zum Genderaspekt und zum Leistungszuwachs zwischen den Stufen und Anforderungsniveaus. Aufgrund der Plausibilität solcher Ergebnisse könne zudem Rückschlüsse auf die Validität des Experimentiertests gemacht werden.

4.1. Welcher Kompetenzunterschied besteht zwischen den Geschlechtern?

4.2. Welche Kompetenzprogression besteht zwischen den Schulstufen?

4.3. Welcher Kompetenzunterschied besteht zwischen den Sprachregionen?

4.4. Welche Kompetenzprogression besteht zwischen den Anforderungsniveaus auf der Sekundarstufe I?

4.5. Welcher Zusammenhang besteht zwischen der Kompetenz und der Fremdsprachigkeit und dem familiären Bildungshintergrund?

4.6. Welcher Zusammenhang besteht zwischen der Kompetenz und den Fach- und Sachinteressen?

4.7. Wie schätzen Schülerinnen und Schüler ihre Testleistung ein?

Teil I

Zur Theorie der experimentellen Kompetenz

Kapitel 3

Der Kompetenzbegriff

3.1 Kompetenzdiskurse

Mit dem häufig verwendeten Begriff der Kompetenz wird in der Naturwissenschaftsdidaktik recht Unterschiedliches bezeichnet. Situativ wird der Kompetenzbegriff mit empirischen Sachverhalten, theoretischen Konstrukten und normativen Zielsetzungen in Verbindung gebracht (Weinert, 2001a; Klieme & Hartig, 2007). Hierfür gibt es zwei Gründe. Einerseits werden unter dem Kompetenzbegriff vielfältige Aspekte des menschlichen Handelns, Denkens und Empfindens subsummiert. Das entsprechend komplexe Konstrukt widerstrebt bislang einer scharfen Operationalisierung. Andererseits wird in der naturwissenschaftsdidaktischen Forschung versucht, sich dem Konzept der Kompetenz auf vielfältigen Wegen mit divergenten Ansätzen zu nähern. In den dazugehörigen wissenschaftlichen Kompetenzdiskursen erhalten empirische, theoretische und normative Aspekte des Kompetenzbegriffs unterschiedlich viel Gewicht und tragen ungleich viel Bedeutung. Die Vielfalt und Heterogenität der Kompetenzdiskurse macht die Übersetzung des einen Diskurses in einen anderen schwierig. Zuweilen ist sie sogar problematisch, weil man sich inkomensurablen Sprachsystemen bedient. Was nottut, ist ein einheitliches Begriffssystem, das die Übersetzbarkeit gewährleistet. Im Folgenden wird ein solches Sprachsystem für den Diskurs über experimentelle Kompetenz entwickelt und begründet. Dabei schlagen wir vor, die Unterscheidung zwischen Standard-, Modell- und Assessmentdiskurs zu machen. Jeder Diskurs wird beschrieben, charakterisiert und in Beziehung zu den anderen Diskursen gesetzt.

3.1.1 Der Standarddiskurs

In der Öffentlichkeit erhält der Kompetenzbegriff Aufmerksamkeit im Zusammenhang mit der Standarddiskussion. Ausgelöst durch die Resultate der TIMS- und PISA-Studien werden in etlichen OECD-Ländern erstmals oder erneut Bildungsstandards entwickelt und

implementiert (Weinert, 2001b; Waddington et al., 2007). Dabei geht es hauptsächlich um inhaltliche und leistungsbezogene Standards (Content Standards und Performance Standards), und nur am Rand um Fragen der schulischen Ressourcen und Rahmenbedingungen (Opportunity-to-learn-Standards) (Oelkers, 2007). Diese Standards werden im Auftrag der Regierungen in Fachgremien ausgehandelt. Dabei fliessen sowohl normative Vorstellungen über Bildungsideale als auch empirische Erfahrungen mit bestehenden Curricula, deren Umsetzung in der Praxis und theoretische Überlegungen aufgrund fachdidaktischer Erkenntnisse in die Standards ein. Es spielen jedoch immer auch politische Opportunitätsüberlegungen eine Rolle. Für Aussenstehende verwirrend ist, dass sich die aus solchen aufwändigen Prozessen resultierenden Leistungsstandards von gewissen Lernzielformulierungen in den bestehenden Lehrplänen nicht unterscheiden (cf. hierzu HarmoS, 2008; KMK, 2005a, 2005b, 2005c; Waddington et al., 2007). Vergessen geht zuweilen nicht nur in der öffentlichen Debatte, sondern auch im wissenschaftlichen Diskurs, dass sich Leistungsstandards von Lehrplanzielen vor allem darin unterscheiden, was nicht in den Standards geschrieben steht. Der Standarddiskurs unterliegt gewissen A priori-Annahmen, die bei Lehrplanzielen nicht oder weniger ausgeprägt sind. Im Wesentlichen unterscheiden sich Standards von Lehrplanzielen durch folgende Aspekte.[1]

Lehr- und Lernbarkeit von Standards. Leistungsstandards wie auch Lernziele implizieren, dass die beschriebenen Inhalte lehr- und lernbar sind. Bei Leistungsstandards wird darüber hinaus erwartet, dass unter der Bedingung, dass notwendige entwicklungsbedingte Lernvoraussetzungen bei den Schülerinnen und Schülern gegeben sind, ein gezieltes Unterrichtsarrangement das allgemeine Kompetenzniveau anhebt.[2]

Verbindlichkeit von Standards. Leistungsstandards sind verbindlich. Sie enthalten durchwegs Can-do-Formulierungen oder Formulierungen, die als solche zu interpretiert sind. Lehrplanziele enthalten demgegenüber auch Must-have-done-, Nice-to-have- oder Could-do-bestly-Formulierungen. Die Verbindlichkeit bei Lehrplanzielen ist dementsprechend gering (cf. Klieme et al., 2007, 27f). Zur Verbindlichkeit gehört zudem eine möglichst klare Festlegung, für wen und in welchem Masse die Standards verbindlich sind. Verschiedene Typen von Standards (Basis-, Regel- und Maximalstandards) sind in diesem Punkt unterschiedlich präzise (cf. Maag Merki, 2007, 23).

[1] Die Frage, inwiefern sich Standards von Lehrplanzielen unterscheiden, wird u. a. bei Labudde (2007, 279) aufgeworfen.

[2] Die Lehrbarkeit von Standards ist eng verknüpft mit der Messbarkeit von Kompetenzen. Beide Anforderungen sind hingegen problematisch (cf. Millar & Driver, 1987, 51-55).

Kompetenzbezug von Standards. Leistungsstandards heben sich von Lehrplanzielen durch ihren Kompetenzbezug ab (Nentwig & Waddington, 2007, 380ff; Klieme et al., 2007, 21ff). Ein Leistungsstandard ist die Beschreibung einer Kompetenzausprägung (Labudde, 2007; Herzog, 2007; Klieme, 2007; Schecker & Wiesner, 2007; Walpuski et al., 2008). Dementsprechend werden Leistungsstandards gelegentlich auch als Kompetenzstandards bezeichnet (Herzog, 2007; Klieme, 2007). Als solche haben Leistungsstandards denselben psychometrischen Anforderungen zu genügen wie Kompetenzkonstrukte. Dies betrifft die beiden nachfolgenden Aspekte.

Messbarkeit von Standards. Ein wissenschaftlicher Leistungsstandard bedingt eine minimale Messbarkeit. Mit Hilfe eines geeigneten Instruments sollte zumindest entschieden werden können, ob ein bestimmter Anteil einer Schülerstichprobe eine Kompetenzausprägung erreicht (cf. Rost, 2004b, 663). Steht ein solches Instrument zur Verfügung, ist grundsätzlich auch die individuelle Diagnose möglich.

Im Zusammenhang mit der geforderten Messbarkeit von Kompetenzausprägungen werden an Messinstrumente verschiedene Anforderungen gestellt, darunter auch die klassischen psychometrischen Bedingungen wie Objektivität, Reliabilität und Validität. In der Assessmentpraxis wird die Objektivität mittels einer hohen Inter-Rater-Reliabilität und die Konstruktreliabilität mittels eines hohen Werts des Cronbach α-Koeffizienten sichergestellt. Unter dem Titel der Konstruktvalidität wird u. a. die *statistische Abgrenzbarkeit* des Kompetenzkonstrukts von anderen Kompetenzkonstrukten eingefordert. Angestrebt wird daher meist ein übergeordnetes Kompetenzstrukturmodell, in das eine Kompetenz eingeordnet werden kann, und ein Messinstrument, das die Kompetenzen im Strukturmodell statistisch unterscheidet, i. e. die Kompetenzen nur schwach korrelieren. Wenn dies nicht gelingt, wird die Konstruktvalidität – zumindest aus psychometrischer Sicht – als defizitär betrachtet (cf. u. a. Ramseier et al., 2011). Entgegen dem gemeinen Diskurs gilt es aber zu betonen, dass die Messung eines Kompetenzkonstrukts auch dann valide sein kann, wenn es keine valide Messung eines übergeordneten Kompetenzstrukturmodells gibt. Es wird hier daher die Meinung vertreten, dass die statistische Abgrenzbarkeit gerade im Hinblick auf die Modellierung experimenteller Kompetenz keine notwendige Forderung für die Konstruktvalidität ist.

Skalierbarkeit von Standards. Sofern wissenschaftliche Leistungsstandards auch die Kompetenzentwicklung abbilden, ist ein skalierbares Messinstrument erforderlich. Das Messinstrument muss dabei die verschiedenen Möglichkeiten der Kompetenzentwicklung berücksichtigen. Diese betreffen zwei Ebenen der Kompetenzprogression (KP): der Aufgabenstellung einerseits und der Aufgabenlösung andererseits:

(KP1) Eine Kompetenzzunahme zeigt sich darin, dass umfangreichere und komplexere Aufgaben bewältigt werden.

(KP2) Eine Kompetenzzunahme zeigt sich darin, dass eine Aufgabe qualitativ besser bewältigt wird.

Die Messbarkeit eines Leistungsstandards bedingt eine präzise Beschreibung der Kompetenzausprägung. Diese wird erreicht, wenn die Aufgabenstellung (KP1), die geforderte Aufgabenlösung (KP2) und das Mass der zu gewährenden Assistenz hinreichend definiert sind (c.f. Rost, 2004b). Eine wissenschaftliche Standardbeschreibung sollte daher stets Aussagen zu folgenden Progressionsdimensionen[3] enthalten.

[*A*] *Aufgabenumfang* (KP1): Als Kompetenzausprägung beschreibt ein Leistungsstandard eine Aufgabe. Dazu muss festgelegt werden, welche Prozesse diese Aufgabe umfasst. Dies bedingt eine zweifache Abgrenzung. Erstens muss entschieden werden, welche Prozesse grundsätzlich zur Kompetenz gehören. Diese sogenannte *äussere Abgrenzung* ist teil der Kompetenzmodellierung und erfolgt häufig in Form von Kompetenzstrukturmodellen, d.h. die Kompetenzdefinition besteht in der Abgrenzung zu anderen Kompetenzen. Zweitens muss geklärt werden, auf welche in der äusseren Abgrenzung enthaltenen Prozesse sich ein bestimmter Leistungsstandard bezieht. Dadurch wird eine *innere Abgrenzung* vorgenommen, die Teilstrukturen oder Stufungen der Kompetenz wiedergibt.[4] Dabei muss auch festgelegt werden, inwieweit "komplexe" Prozesse in diskrete Teilprozesse zerlegt werden sollen (Messick, 1994, 19ff).

[*P*] *Problemkomplexität* (KP1): Das Anforderungsniveau einer Kompetenzausprägung wird wesentlich durch die Komplexität des zu bearbeitenden Problems mitbestimmt. Eine Standardaufgabe, die sich auf einen bestimmten Prozess bezieht, kann auf unterschiedlichem Anforderungsniveau gestellt werden.[5] Die Problemkomplexität er-

[3] Mit dem Begriff Dimensionen soll nicht angedeutet werden, dass es sich um vier statistisch unterscheidbare Dimensionen handelt. Sowohl zu den Dimensionen selbst und möglichen Subdimensionen als auch zum Zusammenspiel der verschiedenen Dimensionen in Bezug auf die Anforderungsprogression von Kompetenzen fehlen bislang empirische Ergebnisse. Wir vertreten die Meinung, dass die Progression nur mit einer Verschränkung der fünf Progressionsdimensionen erklärt werden kann.

[4] Übertragen auf das Beispiel der experimentellen Kompetenz bedeutet dies, dass festgelegt werden muss, inwieweit Teilprozesse wie «Hypothese aufstellen», «Untersuchung planen» oder «Daten auswerten» zur experimentellen Kompetenz gezählt werden und welche Teilprozesse dann auch zu einer bestimmten Kompetenzausprägung gehören. Nicht jede Kompetenzausprägung muss sich auf alle möglichen Teilprozesse beziehen.

[5] Zum Beispiel hängt die Schwierigkeit der Aufgabe, eine gegebene Hypothese empirisch zu überprüfen, massgeblich von der Art und Komplexität der Hypothese ab. Je nach Hypothese gestalten sich auch Manipulationen mit Experimentiermaterial anspruchsvoller (cf. das Beispiel im Abschnitt 4.2.1).

gibt sich grundsätzlich aus der Analyse der Lösungswege, die vom Gegebenen zum Gesuchten der Aufgabe führen, und beinhaltet Aspekte wie die Anzahl und Art der involvierten Variablen[6], der Problemoffenheit (Anzahl Lösungen und Lösungswege)[7], die Zielklarheit und Methodenbekanntheit. Als Teil der Problemkomplexität wird häufig gesondert die *Inhaltskomplexität* diskutiert. Die Ausübung einer Kompetenz erfolgt im Rahmen von Bildungsstandards stets an einem fachlichen Inhalt. Deshalb muss die Komplexität des fachlichen Inhalts als Faktor bei der Bestimmung der Anforderung einer Kompetenzausprägung mitberücksichtigt werden. Dazu gehören sowohl die Abstraktheit[8] als auch die Anzahl und Zusammenhänge der verwandten fachlichen Konzepte.[9]

[T] *Transferumfang* (KP1): Von einer kompetenten Person wird erwartet, dass sie ihre Kompetenz in verschiedenen Kontexten beweist. Eine wissenschaftliche Beschreibung einer Kompetenzausprägung bedingt daher die Definition des Kontextumfangs, innerhalb welchem ein Kompetenztransfer gefordert wird. Der Transferumfang kann dabei einerseits die *Reichweite der Konzeptbasis*[10] und andererseits den Grad der *Fremdheit des Kontextes.*[11] betreffen. Ein Zunahme des Transferumfangs kann daher auch als Erweiterung der Kompetenz interpretiert werden (Hodson, 1992, 126).[12]

[Q] *Prozessqualität* (KP2): Eine präzise Beschreibung einer Kompetenzausprägung be-

[6] im Sinne der "Procedural complexity" gemäss Gott und Duggan (1995, 53)

[7] cf. "Openness" gemäss Gott und Duggan (1995, 54f)

[8] im Sinne von "Conceptual demand" gemäss Gott und Duggan (1995, 53)

[9] Die Evaluation der deutschen KMK-Standards in Naturwissenschaften (KMK, 2005a, 2005b, 2005c) erfolgt anhand eines dreidimensionalen Kompetenzmodells, das neben den *Kompetenzbereichen* (Fachwissen, Erkenntnisgewinnung, Kommunikation, Bewertung) die Dimensionen *Kognitive Prozesse* und *Komplexität* unterscheidet. Die *Kognitiven Prozesse* sind hierarchisch geordnet mit den Stufen «Reproduzieren», «Selegieren», «Organisieren» und «Integrieren». Die *Komplexität* bezieht sich auf die inhaltlichen Zusammenhänge, die eine Aufgabenlösung erfordert. Differenziert wird zwischen «Ein Fakt», «Zwei Fakten», «Ein Zusammenhang», «Zwei Zusammenhänge» und «Übergeordnetes Konzept» (Walpuski et al., 2008; Kauertz et al., 2010). Die Dimension *Kognitive Prozesse* beschreibt eine Art "Prozesskomplexität", die Dimension *Komplexität* entspricht dem Konzept der Inhaltskomplexität. Beide Dimensionen gehören zur Problemkomplexität.

[10] Transferumfang verstanden im Sinne von "breadth of content coverage" (Linn et al., 1991, 20)

[11] Transferumfang verstanden als Graduierung der Kontextferne im Sinne vom Übergang von innerfachlichen Kontexten über Kontexte des persönlich-gesellschaftlichen Umfelds hin zu professionellen Anwendungen in Technik und Wissenschaft (Schecker & Parchmann, 2006, 58).

[12] Bekannte Kompetenzdefinitionen thematisieren den Transferumfang kaum (Bybee, 2002; Weinert, 2001b). Dies gilt auch für die diversen fachspezifischen Kompetenzbeschreibungen, welche innerhalb von Kompetenzstruktur- oder Kompetenzentwicklungsmodellen gemacht werden und die implizit davon ausgehen, dass der Transferumfang durch die lokal gültigen Fach-Curricula gegeben ist (Bernholt et al., 2009; Neumann et al., 2007). In wenigen Ausnahmen wird der Transferumfang explizit thematisiert (HarmoS, 2008).

dingt eine Aussage zur Qualität, mit der die beschriebenen Prozesse erfolgen sollen. Beim Experimentieren kann die Prozessqualität z. B. die Messgenauigkeit oder die Präzision und Korrektheit einer Beobachtung (Hodson, 1992, 126) beinhalten. An die Qualität von Prozessen werden meist gleichzeitig mehrere verschiedene *Massstäbe* angelegt, d.h. Prozesse werden anhand eines Sets unterschiedlicher Kriterien bewertet (cf. Abs. 4.3.2).

[*E*] *Eigenständigkeit* (KP2): Die Progressionsdimension der Eigenständigkeit betrifft Hilfestellungen, die Schülerinnen und Schülern individuell gegeben werden, um eine Barriere im individuellen Lösungsprozess zu überwinden. Standardformulierungen nehmen Aspekte der Eigenständigkeit mit Formulierungen wie "angeleitet" oder "eigenständig" auf (e. g. EDK, 2011, 40, 43). In large-scale Assessments kann der Aspekt der Eigenständigkeit berücksichtigt werden, indem bei der Handhabung von Geräten und Messinstrumenten bei Bedarf Unterstützung angeboten wird.[13] Die Eigenständigkeit ist ein Aspekt bei der Beurteilung einer Aufgabenlösung.

Im üblichen Standarddiskurs werden die Content standards und die Performance standards unabhängig voneinander behandelt und festgeschrieben. Da mit den Content standards aber der *Transferumfang* für die entsprechenden Performance standards bereits gegeben ist, wird dieser Progressionsdimension im Diskurs über Leistungsstandards meist keine Beachtung geschenkt. Ähnliches gilt für die *Eigenständigkeit*. Die Eigenständigkeit lässt sich in einem large-scale Assessment nur bedingt modellieren. Als Progressionsaspekt wird sie daher sowohl im Modell- als auch im Assessmentdiskurs systematisch vernachlässigt. Beim letztlich geführten Diskurs geht es darum, das Anforderungsniveau von Standards mit dem reduzierten Modell der drei Progressionsdimensionen *Aufgabenumfang*, *Problemkomplexität* und *Prozessqualität* zu modellieren. Ein idealisierter wissenschaftlicher Standard entspricht daher einem Punkt im in der Abbildung 3.1 dargestellten dreidimensionalen Koordinatensystem.

3.1.2 Der Modelldiskurs

In der Fachdidaktik erhält der Kompetenzbegriff vor allem im Zusammenhang mit der Diskussion um Kompetenzmodelle Aufmerksamkeit. Der Modelldiskurs bildet dabei das Bindeglied zwischen dem Standard- und dem Assessmentdiskurs. Einerseits fliessen im Modelldiskurs die Ergebnisse des Assessmentdiskurses ein und andererseits liefert er die

[13] Diese Art der Unterstützung wurde z. B. bei den TIMSS-Experimentiertests geboten, wobei die Inanspruchnahme von Hilfe kodiert wurde (e. g. für den Gebrauch des Thermometers bei der Tabletten-Aufgabe von TIMSS, 1994, 27).

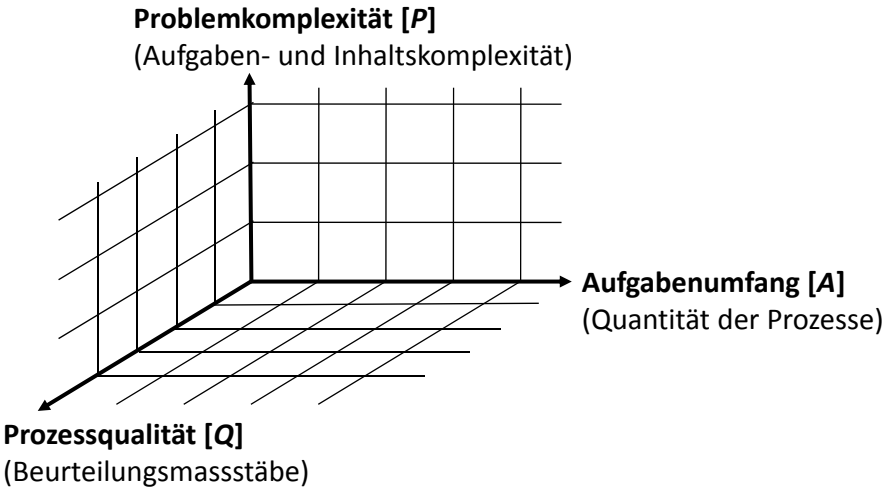


Abbildung 3.1 – Reduziertes Modell der Progressionsdimensionen für Standards

wissenschaftliche Begründung für die bislang normativ gesetzten Standards. Der Modelldiskurs ist entsprechend vielfältig und komplex und allzu oft auch undurchsichtig, weil nicht deutlich gemacht wird, wo und wann der Diskurs auf Beschreibungen, Werturteilen, empirischen Behauptungen oder Postulaten bzw. A priori-Annahmen basiert.[14] Im Zentrum des Modelldiskurses stehen daher die vier Funktionen, die Kompetenzmodelle – seine dies nun Struktur- oder Progressionsmodelle – wahlweise erfüllen.[15]

Deskriptive Funktion von Kompetenzmodellen. Kompetenzmodelle beschreiben und interpretieren empirische Ergebnisse zur Struktur und Entwicklung von Kompetenzen. Beispiele von deskriptiven Modellen experimenteller Kompetenz sind HarmoS (2008); C. von Aufschnaiter und Rogge (2010).

Normative Funktion von Kompetenzmodellen. Kompetenzmodelle bilden den Raster für die Formulierung von Standards. In dieser Funktion sind Kompetenzmodelle bislang vor allem pragmatisch begründet und stark normativ geprägt. Sie widerspiegeln daher kulturell bedingte Wert- und Leistungsvorstellungen: Sie werden praxisnah und praxistauglich formuliert und sollten auf die Entwicklung des Unterrichts innovativ wirken (Labudde, 2007; Bernholt et al., 2009, 223ff).

[14] Die Konfundierung von Beschreibungen, Werturteilen, Behauptungen und Postulaten lässt sich sehr gut an der Aufsatzsammlung zu naturwissenschaftlichen Standards im internationalen Kontext von Waddington et al. (2007) beobachten. Der internationale Diskurs ist stark defizitär, da die Länderdiskurse auf unterschiedlichen Ebenen und teils mit inkomensurablen Sprachsystemen erfolgen.

[15] Schecker und Parchmann (2006) differenzieren nur zwischen der deskriptiven und normativen Funktion von Kompetenzmodellen, wobei die normative Funktion Aspekte der spekulativen und postulativen Funktion enthält. Im Gegensatz zu Schecker und Parchmann (2006) heben wir hervor, dass Kompetenzmodelle mehrere Funktionen gleichzeitig erfüllen.

Spekulative Funktion von Kompetenzmodellen. Kompetenzmodelle stellen Hypothesen dar, die der empirischen Überprüfung – im Modelldiskurs "Validierung" genannt – unterliegen. Da der empirische Gehalt von Modellvorschlägen meist nicht explizit ausgewiesen wird, bleibt unklar, was unter der Validierung eines Kompetenzmodells zu verstehen ist. Konsens besteht darüber, dass die in einem Kompetenzmodell definierten Konstrukte objektiv, reliabel und valide gemessen werden müssen (cf. Bemerkung zur Messbarkeit von Standards auf S. 15). Weniger Klarheit herrscht darüber, inwieweit die implizite *Disjunktivitätsannahme von Kompetenzmodellen* statistisch bestätigt werden muss bzw. wie hoch akzeptable Korrelationen zwischen den Konstrukten sein dürfen. Die Frage stellt sich, da die Disjunktivität bereits auf der Modellebene nicht immer und strikt gegeben ist.[16]

Postulative Funktion von Kompetenzmodellen. Kompetenzmodelle bilden den Ausgangspunkt für large-scale Assessments. Aus den Kompetenzmodellen werden reduzierte Testmodelle mit dazugehöriger Testkonstruktion abgeleitet. Die Ergebnisse eines Assessments werden nun auf zweierlei Arten verwendet: Einerseits werden die Ergebnisse herangezogen, um das Testmodell bzw. das übergeordnete Kompetenzmodell zu "validieren". Das Kompetenzmodell übernimmt dann die Funktion einer Hypothese. Andererseits werden aufgrund der Ergebnisse Aussagen zu Kompetenzausprägungen in der Stichprobe gemacht. In dieser Verwendung bekommt das empirisch nicht überprüfte Kompetenzmodell den Status eines Postulats. Der Modelldiskurs wird heute vor allem im deutschen Sprachraum geführt. Sämtliche fachdidaktisch wesentlichen Aspekte lassen sich auf die allgemeine und grundlegende Frage zurückführen, inwiefern Kompetenzmodelle gleichzeitig verschiedene Funktionen übernehmen oder übernehmen sollten bzw. nicht übernehmen oder nicht übernehmen sollten (cf. u. a. Bernholt et al., 2009; Herzog, 2007; Kauertz et al., 2008; Klieme, 2007; Klieme & Hartig, 2007; Labudde, 2007; Labudde, Duit et al., 2009; Schecker & Parchmann, 2006; Schecker & Wiesner, 2007; C. von Aufschnaiter & Rogge, 2010; Weinert, 2001b).

3.1.3 Der Assessmentdiskurs

Im Zentrum des Assessmentdiskurses steht der Bezug des Kompetenzbegriffs zu konkreten Aufgaben (cf. Klieme et al., 2007, 23ff). Die Abhängigkeit besteht dabei auf dreierlei Weise.

[16] Die Frage nach dem fachdidaktischen Nutzen disjunktiver Kompetenzmodelle erhält im Modelldiskurs keine Beachtung. Da Unterrichtshandlungen jedoch de facto nicht disjunktiv sind, ergibt sich durch ein disjunktives Modell auch kein nennenswerter Vorteil bei der Beschreibung, Evaluation und Steuerung von Unterricht. Die Disjunktivitätsannahme hat eher den Stellenwert eines psychometrischen Nice-to-have.

Definitorische Aufgabenabhängigkeit von Kompetenzen. Die im Kontext von large-scale Assessments häufig zitierte Weinertsche Kompetenzdefinition stellt die Aufgabenabhängigkeit paradigmatisch her: Dergemäss "versteht man unter Kompetenzen die bei Individuen verfügbaren oder durch sie erlernbaren kognitiven Fähigkeiten und Fertigkeiten, um bestimmte Probleme zu lösen, sowie die damit verbundenen motivationalen, volitionalen und sozialen Bereitschaften und Fähigkeiten, um die Problemlösungen in variablen Situationen erfolgreich und verantwortungsvoll nutzen zu können" (Weinert, 2001b, 27f). Eine "Kompetenz ist nach diesem Verständnis eine Disposition, die Personen befähigt, bestimmte Arten von Problemen erfolgreich zu lösen, also konkrete Anforderungssituationen eines bestimmten Typs zu bewältigen", wobei die "individuelle Ausprägung einer Kompetenz [...] von verschiedenen Facetten bestimmt wird [wie] Fähigkeit, Wissen, Verstehen, Können, Handeln, Erfahrung, Motivation" oder Wille (Klieme et al., 2007, 27f). Im Assessmentdiskurs zerfällt diese Definition in zwei Teile, in einen "Weinert vor und nach dem Komma", wobei das Komma vor dem *sowie* gemeint ist (cf. A. Wellensiek in Labudde, Duit et al., 2009, 363ff). Während die motivationalen und volitionalen Kompetenzaspekte, die im "Weinert nach dem Komma" enthalten sind, im allgemeinen fachdidaktischen Diskurs als wertvoll beschrieben werden, spielen diese Aspekte im Assessmentdiskurs nur eine marginale Rolle. Es haben sich damit zwei Verständnisse von Kompetenz etabliert: eine *Kompetenzkonzeption im engeren Sinne*, die nur die messbaren (kognitiven) Aspekte betrifft, und eine *Konzeption im weiteren Sinn*, die auch nicht messbare Aspekte umfasst.

Konzentriert man sich vorerst auf den messbaren Kompetenzaspekt, werden Kompetenzen gemäss Weinert nicht als *bestimmte Fähigkeiten und Fertigkeiten* verstanden, um Probleme zu lösen, sondern als "die [...] Fähigkeiten und Fertigkeiten, um *bestimmte Probleme* [Hervorhebung durch den Autor] zu lösen" . Die subtile Verschiebung des Prädikats *bestimmte* macht deutlich, dass es sich bei den diskutierten Kompetenzen – abgesehen von den motivationalen und volitionalen Aspekten – nicht um allgemeine (kognitiven) Fähigkeiten wie logisches Denken, räumliches Vorstellungsvermögen oder quantitatives Begriffsverständnis geht, sondern um situationsspezifische aufgabengebundene Fähigkeiten.[17] In ähnlicher Weise wird die Aufgabenabhängigkeit der Kompetenzdefinition im

[17] Die Unterscheidung zwischen allgemeinen und problemspezifischen Fähigkeiten zieht sich als Konstante durch alle Kompetenzdiskurse hindurch. *Specialized cognitive competencies* sind für Weinert (2001a, 46f) an eine Domäne oder einen Aufgabenbereich gebunden und heben sich von den inhalts- und kontextunabhängigen *General cognitive competencies* ab. Eine ähnliche Unterscheidung wird bei der Analyse der TIMSS-Erhebungen gemacht (TIMSS, 2000b, 2000a): Um die Aufgabenschwierigkeit zu beschreiben, wird einer Aufgabe einerseits *eine Kompetenzstufe* und andererseits *verschiedene Anforderungsmerkmale* zugeordnet. Dieselbe Unterscheidung wird von H. E. Fischer und Draxler (2007, 648ff) bei der Beschreibung von Physikaufgaben übernommen. Während sich ihre Kompetenzstufe auf eine Beschreibung von problem- und situationsspezifischen Fähigkeiten bezieht, umfassen ihre Anforderungsmerkmale vor allem allgemeine aufgabenunabhängige Fähigkeiten.

Assessmentdiskurs des DFG-Schwerpunktprogramms «Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen» deutlich. Kompetenzen werden dort als "kontextspezifische Leistungsdispositionen, die sich funktional auf Situationen und Anforderungen in bestimmten Domänen beziehen" gehandelt (Klieme & Leutner, 2006; Klieme & Hartig, 2007, 14). Der Aufgabenbezug bleibt auch bestehen, wenn Kompetenz statt im engeren Sinne im weiteren Sinne verstanden wird: Integraler Bestandteil jeder Kompetenzdefinition ist eine Beschreibung einer Aufgabe, die im Falle einer Kompetenzkonzeption im erweiterten Sinne durch eine Beschreibung der Situation ergänzt wird, in welcher die Anforderungen der gestellten Aufgabe zu meistern sind.

Die Aufgabenabhängigkeit von Kompetenzdefinitionen tritt letztlich im Modelldiskurs deutlich zutage: Sowohl die Abgrenzung von Kompetenzen in Strukturmodellen als auch die Abgrenzung von Stufen in Progressionsmodellen erfolgt ausschliesslich über die Differenzierung von Aufgaben- und Bearbeitungsmerkmalen (KP1 und KP2). Die beschriebenen Merkmale erhalten somit definitorischen Charakter.

Explikatorische Aufgabenabhängigkeit von Kompetenzen. Obwohl die im Assessmentdiskurs gehandelten Kompetenzformulierungen sich auf aufgabenabhängige, spezifische Fähigkeiten beziehen, bleiben viele dieser Beschreibungen schwammig. Gerade bei Standardbeschreibungen bleibt oft unklar, was für Aufgaben mit dem Standard verknüpft werden sollen. Und die Resultate von grossen Tests lassen sich erst dann richtig einordnen, wenn Testaufgaben bekannt sind (cf. Streitgespräch von H. Schecker und H. Fischer in Labudde, Duit et al., 2009, 346ff). Den Aufgaben kommt daher im Kompetenzdiskurs eine explikative Rolle zu. Kompetenzbeschreibungen werden erst durch Aufgabenbeispiele und Kodierschemen expliziert (cf. Resultate des Kapitels 7). Dabei wird der Kompetenz eine Menge von Testaufgaben zugeordnet, die sich durch ein Set gemeinsamer Problemmerkmale auszeichnen. Diese Merkmale können sich auf oberflächliche Aufgabeneigenschaften (e. g. Antwortformate, Textlängen) wie auch auf theoretische Konstrukte (e. g. Anforderungsmerkmale im Sinne von H. E. Fischer und Draxler (2007)) beziehen, die beide mit der zu messenden Kompetenz in Verbindung gebracht werden. Die Aufgabenmerkmale übernehmen daher eine explikative Funktion bei der Kompetenzbeschreibung, die mit ihrer definitorischen Rolle verschmilzt.

Im Rahmen von large-scale Assessments wurden verschiedene Versuche unternommen, die Schwierigkeit spezifischer Aufgabenkompetenzen auf allgemeine kognitive Fähigkeiten zurückzuführen, sei dies unter dem Titel von *Anforderungsmerkmalen* bei TIMSS (TIMSS, 2000b, 2000a), bei PISA in Form von *kognitiven Teilkompetenzen* (PISA, 2004, 2007; Senkbeil et al., 2005) oder im angloamerikanischen Assessmentdiskurs mit sogenannten *Reasoning dimensions* (Hamilton et al., 1997; Shavelson et al., 2002) oder *Cognitive activities* bzw. *Cognitive components of competence* (Baxter et al., 1995; Baxter & Glaser, 1998, 38).

Abhängigkeit von Testaufgaben. Mit der schulischen Kompetenzvermittlung wird das Ziel verfolgt, Schülerinnen und Schüler zu befähigen, reale Anforderungen in authentischen Situationen ausserhalb der Schule zu meistern. Die Kompetenzausübung in authentischen Situationen ist per se ad hoc und erfolgt auch spontan. In beiden Eigenschaften unterscheidet sich die authentische Kompetenzausübung in realen Situationen von der erzwungenen experimentellen Kompetenzmessung im Rahmen einer large-scale Erhebung. Diese im Schulbereich gegebene, systematisch grosse Kluft zwischen angestrebtem Ziel und gemessener Praxis tut sich auch im Assessmentdiskurs auf und führt dazu, dass der Anspruch fallengelassen wird, die Kompetenzausübung in authentischen ausserschulischen Situationen zu modellieren. Tatsächlich begnügt man sich mit der Modellierung von möglichst authentischen Unterrichtssituationen (Messick, 1994, 17f).[18] Dabei beschränkt man sich experimentell auf Kompetenzaspekte, die mit vertretbarem Aufwand und psychometrisch erfolgsversprechend gemessen werden können. Der Kompetenzdiskurs gerät daher in starke Abhängigkeit von den messtechnischen Möglichkeiten.

3.1.4 Probleme beim Zusammenspiel der Kompetenzdiskurse

Konsistenz von Diskursen. Wichtige Merkmale eines Kompetenzdiskurses sind seine Sprache, seine Ziele und die Voraussetzungen. Die *Diskurssprache* ist durch die Begrifflichkeit, d.h. durch die im Diskurs verwandten Begriffe und Begriffsbeziehungen gegeben. Die Voraussetzungen sind das, was an Begrifflichkeit vorgegeben ist. Die *Diskursziele* wiederum betreffen die Anforderungen, denen ein optimierter Diskurs genügen sollte. Sie dienen gleichzeitig als Orientierungspunkt und Bewertungskriterium eines sich dynamisch entwickelnden Diskurses. Im Rahmen der zulässigen Dynamik eines Diskurses sind Diskurssprache, Voraussetzungen und Diskursziele unterschiedlich starr.

An einem Diskurs sind stets viele und verschieden motivierte Partner beteiligt, die unterschiedliche Sprachen führen und unterschiedliche Ziele verfolgen. Nebst den diskursspezifischen Zielen (e. g. die erfolgreiche Implementierung von Standards in der Schulpraxis oder das erfolgreiche Assessment nationaler Standards) wird daher mit einem Diskurs immer auch das übergeordnete Ziel verfolgt, die Sprache, Voraussetzungen und Ziele zu vereinheitlichen. Angestrebt wird sowohl in Bezug auf Standards, Modelle oder Assessments *Diskurskonsistenz.*

[18] Die im Assessmentdiskurs gehandelten naturwissenschaftlichen Kompetenzmodelle sind allesamt *stufenspezifische Modelle des naturwissenschaftlichen Unterrichts.* Die Kompetenzmodelle sind auf typische Aufgabensituationen in einem Unterricht ausgerichtet: Weder beschreiben sie ausserschulische Tätigkeiten wie diejenigen von Wissenschaftlerinnen und Ingenieuren (e. g. das HarmoS-Kompetenzmodell Naturwissenschaften (Harmonisierung obligatorischer Schule), cf. HarmoS, 2008) noch können sie auf alle Schulstufen angewandt werden (e. g. das ESNaS-Kompetenzmodell (Evaluation der Standards in den Naturwissenschaften für die Sekundarstufe I), cf. Kauertz et al., 2010; Walpuski, 2010).

Kongruenz von Diskursen. In der Regel wird ein Standarddiskurs durch einen Modell- und Assessmentdiskurs begleitet und vice versa. Fasst man alle drei Diskurse zu einem Gesamtdiskurs zusammen, folgt als Anwendung der Konsistenzforderung die Vereinheitlichung der drei Diskurssprachen. Die Optimierung der diversen diskursspezifischen Ziele im Rahmen der Gesamtdiskursdynamik bedingt zudem, dass in den Teildiskursen die Voraussetzungen und Ziele der anderen Diskurse berücksichtigt werden. Begriffe, Ziele und Voraussetzungen müssen in allen drei Diskursen kongruent gehandhabt werden. Übergeordnetes Ziel des Gesamtdiskurses ist somit die *Kongruenz der Teildiskurse.*

Ungleichzeitigkeit von Diskursen. Um die Kongruenz zwischen Teildiskursen herzustellen, ist es von Vorteil, wenn sich Sprache, Ziele und Voraussetzungen der Diskurse gleichzeitig weiterentwickeln. Ist nur ein Diskurs dynamisch, während andere Diskurse entweder nicht vorhanden (*Diskursabsenz*) oder starr und daher meist unterentwickelt (*Diskursstagnation*) sind, gelingt das Abgleichen der Sprachen und Ziele auf die gegenseitigen Voraussetzungen nicht optimal. Die Optimierung des Gesamtdiskurses wird also durch die Ungleichzeitigkeit der Teildiskurse, die sich als (oft zeitlich beschränkte) Absenz oder Stagnation eines Teildiskurses manifestiert, erschwert.[19]

3.2 Diskurse zur experimentellen Kompetenz

So wie in den verschiedenen Diskursen unterschiedliche Seiten des Kompetenzbegriffs im Allgemeinen beleuchtet werden, so erhält auch die experimentelle Kompetenz im Speziellen je nach Diskurs eine unterschiedliche Tönung. Diesen Farbnuancen ist der folgende Abschnitt gewidmet.

3.2.1 Der experimentelle Standarddiskurs

In den OECD-Ländern werden experimentelle Leistungsstandards im Sinne von kompetenzbezogenen, messbaren und verbindlichen Lernzielen unter unterschiedlichen Bezeichnungen wie goals, aims, standards diskutiert. Die experimentellen Standards beziehen sich je nach landesspezifischem Kompetenzdiskurs auf *scientific literacy*, *experimental abilities*, *competencies* oder *skills* und werden unter unterschiedlichen Oberbegriffen wie *scienti-*

[19] In Deutschland wurde der Standarddiskurs abgeschlossen, bevor der Modell- und Assessmentdiskurs überhaupt aufgenommen wurde. In der Schweiz wurde der Gesamtdiskurs zwar gleichzeitig gestartet (HarmoS, 2008, Kap. 1). Während sich aber der Standarddiskurs nach Vorlegen der Vorschläge zu Bildungsstandars durch das Konsortium HarmoS Naturwissenschaften in verschiedenen Gremien weiterentwickelte, stagnierten der Modell- und Assessmentdiskurs. Der Assessmentdirskurs wird erst nach Vorliegen der Standards im Hinblick auf ein nationales Bildungsmonitoring wieder aufgenommen.

fic activities, *scientific inquiry*, *Erkenntnisgewinnung* oder *Fragen und untersuchen* zu Kompetenzen zusammengefasst (cf. Waddington et al., 2007). Im Folgenden werden die wesentlichen Merkmale experimenteller Standards besprochen.

Äussere Abgrenzung: Problemspezifische und nicht problemspezifische Fähigkeiten. In einigen Ländern mit Standarddiskurs wird der Bezug zu einer experimentellen Kompetenz nicht explizit gemacht. Dies weil entweder der dazugehörgige Modelldiskurs nicht geführt (e. g. Dolin, 2007) oder eine speziell experimentelle Kompetenz nicht von einer allgemeinen Naturwissenschaftskompetenz unterschieden wird (e. g. Hafner, 2007). Bei diesen Standards kann daher nicht entschieden werden, inwieweit sie einer experimentellen Kompetenz zugeordnet werden sollen. Auch in Ländern mit explizitem Modelldiskurs bleibt unklar, welche Teilprozesse zur experimentellen Kompetenz gezählt werden, wenn in den entsprechenden Modelldiskursen Teilprozesse nur vage umschrieben sind oder in Varianten verschiedenen Kompetenzen zugeordnet werden. Dies trifft zum Beispiel auf das *Ordnen, Strukturieren und Modellieren von Beobachtungen* zu. Es stellt sich die Frage, inwiefern dieser rein kognitive Prozess integraler Bestandteil einer Beobachtung ist und daher zur Kompetenz der Erkenntnisgewinnung gezählt wird. Oder ob das Ordnen, Strukturieren und Modellieren vor allem auf Vorwissen beruht und daher eine Anwendung von Fachwissen ist. Je nach Fachgebiet kann dieser Teilprozess unterschiedlich eingeordnet werden (e. g. KMK, 2005a, 14; KMK, 2005b, 11) oder bei fächerübergreifenden Standardformulierungen wird der Teilprozess in verschiedenen Kompetenzen untergebracht (e. g. EDK, 2011).

In experimentellen Standards erscheinen zuweilen Formulierungen, die sich auf allgemeine, nicht problemspezifische Fähigkeiten wie kommunikative oder mathematische Fähigkeiten oder die Fähigkeit, eigenständig und sicher zu arbeiten, beziehen. Wie bereits auf Seite 21 diskutiert, werden für die kompetente Erfüllung einer Arbeit immer auch nicht aufgabenspezifische Fähigkeiten benötigt. Trotzdem stellen diese Fähigkeiten gemäss der hier vertretenen Interpretation keine eigenständige Kompetenz bzw. Teilkompetenz dar. Beim Assessment experimenteller Kompetenz werden Aspekte dieser allgemeinen Fähigkeiten – wie Aspekte der Kommunikationsqualität, der mathematischen Auswertungsqualität oder der Eigenständigkeit und Sicherheit – als Teil der Prozessqualität interpretiert (vgl. Diskussion im Abschnitt 4.3.2).

Beschränkung auf Erkenntnisgewinnungsprozesse. "Die Wissenschaft dient in erster Linie der Verbesserung der Interpretation der Welt und erst in zweiter Linie der Verbesserung der Welt."[20] Was für die Wissenschaft allgemein gilt, trifft auch auf die standardisierten

[20] mündliches Zitat von Prof. Dr. Roland Reichenbach

Ziele des schulischen Experimentierens zu: Die nationalen Standards zur experimentellen Kompetenz beziehen sich in erster Linie auf Aktivitäten der Erkenntnisgewinnung. Aktivitäten der Naturgestaltung sind in den Standards implizit enthalten, sofern sie der Erkenntnisgewinnung dienen (cf. Waddington et al., 2007). Nimmt man die deutschen Regelstandards für den Mittleren Schulabschluss der Kultusministerkonferenz (2005a, 14; 2005b, 12; 2005c, 11) zum Beispiel, so werden den Schülerinnen und Schüler in der Beziehung zur Natur zwei Rollen zugestanden: Einerseits nehmen sie die Natur *passiv* wahr, indem sie die Phänomene der sich spontan offenbarenden Natur "beobachten", "beschreiben", "darstellen", "vergleichen", "analysieren", "bestimmen" und "erklären". Andererseits greifen sie in die Natur *aktiv* ein mit dem Ziel, die Welt besser zu verstehen bzw. interpretieren z können. Dabei provozieren sie künstlich Phänomene, indem sie "Untersuchungen" und "Experimente" "planen", "durchführen", "Daten erheben", "auswerten" und in ihnen "Trends, Strukturen oder Beziehungen finden" sowie die "Ergebnisse dokumentieren". In beiden Rollen geht es darum, die Natur im Sinne einer Analyse zu zerlegen. Nicht vorgesehen ist hingegen, dass Schülerinnen und Schüler auch eine konstruierende Rolle übernehmen und die Natur im Sinne einer Synthese zusammensetzen. In den Standards der Kultusministerkonferenz fehlen Begriffe wie «erfinden», «entwickeln», «herstellen», «fertigen», «konstruieren», «zusammenbauen», «zum Laufen bringen», «testen», «optimieren» oder «reparieren». Dieser Befund trifft auf die meisten OECD-Länder zu.[21] Technisch-praktische Aspekte der Naturwissenschaft finden dort nur als Mittel zum Zweck Eingang in die Standards, um erkenntnisbringende Experimente zum Funktionieren zu bringen. Sie bilden jedoch keine eigenständigen Ziele.[22]

Idealisierung der Experimentierprozesse. Mit der Ausrichtung auf Prozesse der Erkenntnisgewinnung ist eine weitere Einengung der Konzeption experimenteller Kompetenz festzustellen. Standardformulierungen in diesem Bereich orientieren sich an der Struktur ei-

[21] In den lateinischen Ländern ohne starke Berufsbildung wird den technisch-praktischen Aspekten der Naturwissenschaft kulturphilosophisch wenig Wert beigemessen. Deren Bildungsstandards bilden diese kulturell bedingte Wertvorstellung ab. In den deutschsprachigen Ländern, wo sich ab der zweiten Sekundarstufe anerkannte berufsspezifische Bildungswege vom akademischen Bildungsweg ablösen, gibt es starke Initiativen u. a. seitens der technischen Industrie, die Technik im Kompetenzdiskurs zu verankern (cf. die Vorschläge für Technikstandards vom Verband Deutscher Ingenieure, 2007).

[22] Ausnahmen bilden u. a. die Niederlande, die Schweiz und die USA. Während die niederländischen Standards für die Sekundarstufe I und II separate und detaillierte *technical-instrumental skills* und *design skills* unterscheiden (Driessen, 2007, 229f), bleiben die Formulierungen zur Technik in den Schweizer Standards für die obligatorische Schule vage. Der im Teilaspekt "Geeignete Werkzeuge, Instrumente und Materialien auswählen und verwenden" genannte Zweck "für [...] technische Konstruktionen" wird nicht näher erläutert (EDK, 2011, 7). Die amerikanischen National Standards (1996) wiederum unterscheiden neben dem Kompetenzbereich *Using scientific inquiry* auch den Bereich *Using technological design*.

nes *idealisierten wissenschaftlichen Experimentierprozesses*, wie er exemplarisch in der Abbildung 3.2 dargestellt wird (Modell von Murphy und Gott (1984) entnommen aus B. Fairbrother (1991, 159)). Mit der Idealisierung des Experimentierprozesses ist jedoch eine Standardisierung der Prozesse, Strategien und Resultate des Experimentierens verbunden. In der idealisierten Konzeption verläuft der Experimentierprozess linear, durch-

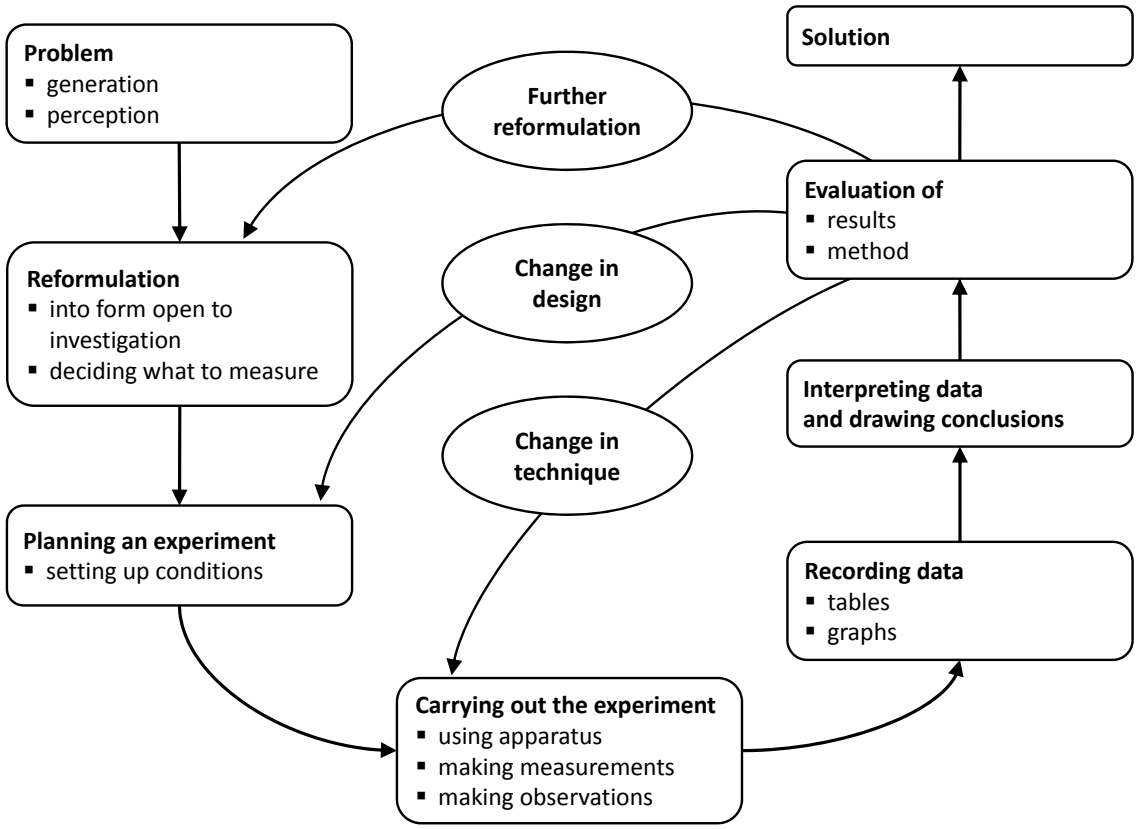


Abbildung 3.2 – Idealisierung des wissenschaftlichen Experimentierprozesses gemäss Murphy & Gott

strukturiert und jederzeit kontrolliert. Jedem Teilprozess geht ein bestimmter anderer Teilprozess voraus und folgt in eindeutig gegebener Weise ein dritter Teilprozess. Experimentelle Handlungen bauen stets auf einer wohldefinierten Fragestellung auf, die in eine überprüfbare Hypothese umformuliert werden kann. Zwar ist im Modell von Murphy und Gott (1984) eine zyklische Vorgehensweise beim Experimentieren angelegt, jedoch nur innerhalb der linearen Abfolge gegebener Teilprozesse. Spontane Abbrüche, Umwege, unvorhergesehene Abweichungen vom Plan oder Neuanfänge sind ebenso nicht vorgesehen wie planloses Ausprobieren verschiedener Varianten, Experimentieren "auf gut Glück [...], um zu sehen, was geschieht" (T. S. Kuhn, 1997 [1962], 100) oder Problemlösen nach der Methode «Versuch und Irrtum». Gemäss dieser Konzeption sind die Auswertungsmethoden bereits in der Planungsphase bekannt und alle möglichen Ausgänge der Untersuchung

können vorgängig benannt werden. Für überraschende Entdeckungen gibt es in diesem Modell keinen Platz.

Die Ziele, die mit dieser Art des standardisierten Experimentierens erreicht werden, entsprechen den Zielen, die T. S. Kuhn mit der "normalen Wissenschaften" verbindet. "Normale Wissenschaft strebt nicht nach neuen Tatsachen oder Theorien und findet auch keine, wenn sie erfolgreich ist." Ihr Zweck ist die stetige "Ausweitung des Umfangs und der Exaktheit wissenschaftlicher Kenntnisse" (T. S. Kuhn, 1997 [1962], 65). Diese Art der Wissenschaft findet unter unverändertem Paradigma (sensu T. S. Kuhn, 1997 [1962]) statt. Ein Paradigmenwechsel wird weder erwartet, noch wird er angestrebt. In ähnlicher Weise verhält es sich bei den Leistungsstandards zur experimentellen Kompetenz: Fragestellung, Planung, Durchführung und Auswertung eines standardisierten Experimentierens erfolgen stets in derselben Theorie und derselben Sprache. In den Standards kommen keine Anomalien vor, und alternative Hypothesen, neue Theorien oder Konzeptwechsel sind als Resultate einer Untersuchung nicht vorgesehen.

Das Modell von Murphy und Gott (1984) bildet sehr gut ab, wie Forschungsprozesse heutzutage in wissenschaftlichen Projektanträgen dargestellt werden: Überraschungen oder Paradigmenwechsel kommen nicht vor, weil sie nicht planbar sind. Forschung verläuft aber nicht immer nach Plan, und Erkenntnisgewinnung erfolgt zuweilen spontan. Das Modell beschreibt daher nur bedingt, wie in Labors gearbeitet wird. Noch weniger entspricht das Modell dem Vorgehen von Laien. Das tatsächliche Experimentier- und Problemlöseverhalten von Schülerinnen und Schülern muss im Vergleich zum dargestellten Ideal als stark lücken- und sprunghaft angenommen werden. Das Vorgehen in Experimentiersituationen ist auf jeden Fall kein linearer, sondern ein suchender, herantastender Prozess, wie die Analysen von Theyssen et al. (2001); S. von Aufschnaiter und Welzel (1997); S. von Aufschnaiter et al. (2000); C. von Aufschnaiter und Rogge (2010) zur Kompetenzprogression in experimentellen Lernsituationen deutlich aufzeigen.

Mit der Idealisierung der Experimentierprozesse wird schliesslich die empiristische Auffassung transportiert, Erkenntnisse über die Welt würden nach rationalen, kontextunabhängigen und allgemeingültigen Methoden gewonnen. Die Standards widerspiegeln die gängige Unterscheidung der deduktiven und induktiven Methoden, wobei auf der einen Seite der hypothetisch-deduktive Weg als erkenntnisgenerierendes Ideal angestrebt wird und auf der anderen Seite Sinneswahrnehmungen als offensichtlich behandelt werden. Schülerinnen und Schüler bekommen somit klare Rollen zugeteilt: Inhaltlich nehmen sie entweder eine rein aktive Rolle (als bürokratisch hypothesengestützte Experimentatoren) oder eine rein passive Rolle (als vorurteilslose, objektive Beobachter) ein (cf. Seite 25). Eine Rolle dazwischen gibt es nicht. Mit der Idealisierung und Standardisierung des Experimentierens läuft man Gefahr, dass ein falsches Bild der Naturwissenschaft ver-

mittelt wird, das sowohl wissenschaftsphilosophischen Erkenntnissen (Carrier, 2006) als auch kognitionspsychologischen Ergebnissen zum problemlösenden Denken von Kindern (Koslowski, 1996) widerspricht.

Innere Abgrenzung: Kompetenzzerlegung in Teilaufgaben. Charakteristisch für experimentelle Standards ist die Zerlegung der Kompetenz in Teilaufgaben, die unterschiedliche Teilprozesse umfassen und unterschiedliche Vorgaben enthalten. Der per se zyklische Experimentierprozess, wie er in der Abbildung 3.2 dargestellt ist, wird dabei in sinnvolle Teilaufgaben mit festem Ausgangs- und Endpunkt zerlegt. So wird beispielsweise das Formulieren und das Überprüfen von Hypothesen zuweilen als separate Aufgaben beschrieben, wobei die Hypothese einmal gesucht und einmal gegeben ist. Die zwei Teilaufgaben können aber auch zu einer umfangreicheren Aufgabe zusammengefasst werden mit dem Effekt, dass das Anforderungsniveau anhoben wird. Teilprozesse verändern zudem ihren Charakter, wenn sie in andere Teilaufgaben eingebettet werden. Der Prozess des Beobachtens im Sinne von "Bewusst wahrnehmen"[23] ist ein anderer als das Beobachten, das bei der Durchführung von quantitativen Messungen stattfindet. Beim "Bewusst wahrnehmen" wird die Natur stark subjektiv interpretiert. Das Sinnesorgan selbst ist das Messinstrument und die Wahrnehmungen unterliegen dem Einfluss von Präkonzepten und Sinnestäuschungen. Beim Beobachten eines Messinstruments ist der Spielraum für subjektive Interpretation wegen der dem Messinstrument zugrundeliegenden einheitlichen Theorie geringer.

Die Abgrenzung von Teilaufgaben experimenteller Kompetenz dient somit folgenden drei Zwecken: Erstens werden unterschiedliche Aspekte von Teilprozessen hervorgehoben. Zweites wird das Anforderungsniveau der experimentellen Kompetenzausprägungen den situationsbedingten Bedürfnissen angepasst. Drittens dient die Zerlegung der Kompetenz in Teilaufgaben auch der Bildung von Teilkompetenzen. Der entsprechende Diskurs ist Teil des Modelldiskurses und wird dort vertieft behandelt.

Prämisse der Inhaltsunabhängigkeit von Prozessstrategien. Von einer kompetenten Experimentatorin wird erwartet, dass sie ihr Können in unterschiedlichen Kontexten beweist und somit eine gewisse Transferleistung erbringt. Wie bereits im Abschnitt 3.1.1 erwähnt, gehört zu einer vollständigen Beschreibung eines Standards die Festlegung des Transferumfangs. In den allermeisten Ländern erfolgt dieser via Content standards unabhängig von der Festlegung der Performance standards. Somit unterliegt der Diskurs über experimentelle Leistungsstandards der *Prämisse, dass transferfähige Prozessstrategien unabhängig von den Inhalten existieren, beschrieben und gemessen werden können.* Viele experimen-

[23] cf. (HarmoS, 2008, 46ff)

telle Standards enthalten dementsprechend keinen Bezug zu Inhalten. Ausdruck dieser Prämisse ist der im Rahmen des Modelldiskurses zur Entwicklung der Experimentierfähigkeit unternommene Versuch, theorieunabhängige "korrekte" Experimentierstrategien zu beschreiben (Hammann, 2004; Hammann et al., 2008; Mayer et al., 2008) sowie kontextunabhängige Strategiefehler, sogenannte Bias, zu orten und hierarchisch zu ordnen (e. g. Ehmer & Hammann, 2007; Hammann et al., 2006; D. Kuhn & Phelps, 1982; Schauble, 1990; Schauble, Klopfer & Raghavan, 1991; Tschirgi, 1980). Die Probleme, die mit diesem Ansatz verbunden sind, werden im Abschnitt 4.3.3 diskutiert.

Scientific inquiry versus nature of science. Das wissenschaftliche Experimentieren ist eng mit wissenschafts- und erkenntnistheoretischen Fragen verflochten. Dazu wird in den NAEP (2008, 73) festgehalten: "Scientific inquiry is more complex than simply making, summarizing, and explaining observations, and it is more flexible than the rigid set of steps often referred to as the 'scientific method'. The *National Standards* make it clear that inquiry goes beyond 'science as a process' to include an understanding of the nature of science [...] . [...] when students use scientific inquiry they are drawing on their understanding about the nature of science [...].". Die Grenze zwischen wissenschaftlichem Handeln (*scientific inquiry*) und metawissenschaftlichen Überzeugungen (*nature of science*) sind unscharf. Standards zum Experimentieren enthalten daher zuweilen auch explizit Inhalte zu Natur der Naturwissenschaft (e. g. Mayer, 2007; Mayer et al., 2008). Da die von Schülerinnen und Schülern gewählten Experimentierstrategien auch von deren epistemischen Überzeugungen abhängen (Carey et al., 1989; Höttecke, 2001; Leach, 2002; Lunetta, 1998; D. Kuhn, 1989; Millar, 1989), sind Aspekte zur Natur der Naturwissenschaften zumindest implizit immer in experimentellen Standards enthalten.

3.2.2 Der experimentelle Modelldiskurs

Nicht zu jedem Standarddiskurs findet auch explizit ein Modelldiskurs statt. Ein Modelldiskurs zeichnet sich durch den Gebrauch von Begriffen wie Dimensionen oder Stufen aus, die der Differenzierung der Struktur und Anforderungsprogression der experimentellen Kompetenz dienen, und bildet die Basis für ein entsprechendes Assessment. Zu jedem Modelldiskurs gibt es daher den entsprechenden Assessmentdiskurs.

Kompetenzstrukturmodelle: Äussere Abgrenzung. Eine Kompetenzstruktur wird einerseits negativ durch die Abgrenzung zu anderen Kompetenzen und andererseits positiv durch die Präzisierung des Kompetenzinhalts definiert. Zur Inhaltspräzisierung gehört die innere Abgrenzung von Teilkompetenzen und Teilaufgaben. Mit der Formulierung von Teilaufgaben werden unterschiedliche Aspekte einer Kompetenzausprägung beschrie-

ben, von denen vermutet wird, dass sie eine intern konsistente Skala für die Kompetenz bilden (*implizite Konsistenzannahme*). Teilaufgaben werden häufig zu Teilkompetenzen zusammengefasst, womit implizit die Behauptung verbunden ist, dass die Teilkompetenzen konsistente Skalen bilden und sich statistisch getrennt messen lassen (cf. *implizite Disjunktivitätsannahme* auf Seite 20).

Die Abgrenzung der experimentellen Kompetenz von nicht-experimentellen (naturwissenschaftlichen) Kompetenzen geschieht aufgrund von vier Differenzierungsprinzipien (DP), die je nach Modelldiskurs unterschiedlich gewichtet werden.

DP1 *Methodenwissen statt Inhaltswissen*: Experimentelle Kompetenz beruht auf dem Wissen über naturwissenschaftliche Methoden. Es geht dabei nicht um das "knowing that" oder das "knowing why", sondern um das "knowing how" in der naturwissenschaftlichen Praxis (cf. NAEP, 2008, 65).

DP2 *Prozessaufgaben statt Konzeptaufgaben*: Beim Experimentieren, verstanden als wissenschaftliches Problemlösen (cf. Mayer, 2007), stehen die Prozesse im Vordergrund. Konzepte spielen nur soweit eine Rolle, wie sie für die Problembeschreibung und -lösung benötigt werden. Z. B. will man mit einer Testaufgabe, bei der Waldvögel beobachtet werden, nicht das Fachwissen über Waldvögel evaluieren, sondern die Art und Weise wie beobachtet wird.

DP3 *Handarbeit statt Kopfarbeit*: Gemäss Hodson (1990, 1992) dient die schulische Laborarbeit drei Zwecken: dem "Learning science", dem "Learning about science" und dem "Learning to do science". Mit der experimentellen Kompetenz wird der zweite und dritte Aspekt eingefangen. Beim "Doing science" wird das "Knowing about science" wichtig. Dazu wird auch Kopfarbeit benötigt. Vor allem aber geht es um praktische Handarbeit.

DP4 *Erkenntnisgewinnung statt Erkenntnisanwendung*: Experimentelle Kompetenz wird immer auch als experimentelle Erkenntnisgewinnung, selten jedoch als praktisch-technische Erkenntnisanwendung interpretiert (cf. Seite 25f). Die technische Anwendung bleibt im Modelldiskurs zweitrangig.

Kompetenzstrukturmodelle unterscheiden sich im Hinblick auf die äussere Abgrenzung der experimentellen Kompetenz durch die unterschiedliche Graduierung, mit der die obigen Prinzipien zur Anwendung kommen. Am Beispiel des "Ordnens und Strukturierens" von Beobachtungen soll dies erläutert werden.

Diese wissenschaftliche Tätigkeit wird in den Kompetenzmodellen der Schweiz, der USA und von Deutschland unterschiedlich eingeordnet (cf. Tab. 3.1). Im Schweizer Modell von HarmoS (2008) stellt das «Ordnen, strukturieren und modellieren» eine eigene,

Kompetenz	CH: HarmoS (2008)	USA: NAEP (2008)	D: KMK (2005a)	TIMSS (1997)
experimentell	Fragen und untersuchen	Using science inquiry	Erkenntnisgewinnung	Scientific investigation
		Using technological design		Using scientific procedures
nicht-experimentell	Ordnen, strukturieren, modellieren	Identifying science principles		
	Informationen erschliessen	Using science principles	Anwendung von Fachwissen	Scientific problem solving and applying concept knowledge
	Einschätzen und beurteilen		Bewertung	
	Mitteilen und austauschen		Kommunikation	

Tabelle 3.1 – Vergleich von Kompetenzstrukturmodellen

nicht-experimentelle Kompetenz dar, die sich von der experimentellen Kompetenz «Fragen und untersuchen» abgrenzt. In den National Standards (1996) wird das Beschreiben, Messen und Klassifizieren von Beobachtungen dem nicht experimentellen Kompetenzbereich «Identifying science principles» zugeordnet, bei dem es auch um die Erkennung und Anwendung von wissenschaftlichen Prinzipien und Gesetzen geht. Im deutschen Modell der KMK (2005a) ist das Ordnen und Modellieren im Kompetenzbereich Erkenntnisgewinnung integriert. Die unterschiedliche Einordnung der Tätigkeit des Ordnens und Strukturierens von Beobachtungen geht einher mit einer unterschiedlichen Interpretation dieser Tätigkeit, die zu einem gewissen Grad subjektiv ist. Es stellen sich die Fragen, ob eine Beobachtung vor allem Methode oder Inhalt ist. Ob der Prozess des Beobachtens oder die für das Ordnen notwendigen Konzepte im Vordergrund stehen. Ob das Ordnen von Beobachtungen als reine Kopfarbeit oder auch als Handarbeit gedeutet wird. Oder ob das Ordnen von Beobachtungen als Anwendung von Wissen oder als Generierung von Wissen interpretiert wird.

Kompetenzstrukturmodelle: Innere Abgrenzung. Die innere Struktur experimenteller Kompetenz wird durch die Differenzierung von Teilkompetenzen und deren Präzisierung durch Teilaufgaben bestimmt.[24] Die hierbei verwandten Differenzierungsprinzipien ermöglichen die Bestimmung von Modelltypen. Wir schlagen die Unterscheidung von vier Differenzie-

[24] Die Definition erfolgt mit der Zielsetzung, den Modelldiskurs mit dem begleitenden Standard- und Assessmentdiskurs abzugleichen, indem sowohl fachdidaktische Überlegungen zu Standards als auch assessmentspezifische Desiderata berücksichtigt werden. Unter der Annahme, grösstmögliche Kongruenz zwischen Modell- und Assessmentdiskurs herzustellen, müssen die mit der Formulierung einer Modellstruktur implizit gemachten *psychometrischen Annahmen* ernst genommen werden: Diese sind, 1.) dass die Teilaufgaben eine konsistente Skala für die jeweilige Teilkompetenz ergeben (*Konsistenzannahme*), 2.) dass sich die Teilkompetenzen statistisch unterscheiden (*Differenzierungsannahme*).

rungsprinzipien vor.

PrA *Differenzierung von Prozessalternativen*: Prozessalternativen sind mehr oder weniger ausschliessende experimentelle Handlungen.

PrZ *Differenzierung einer Prozessabfolge*: Eine Prozessabfolge besteht meist aus disjunkten Teilprozessen, die nacheinander ausgeführt werden.

PrE *Differenzierung einer Prozesserweiterung*: In einer Prozesserweiterung ist der eine Prozess vollständig im nächsthöheren Prozess enthalten.

WiA *Differenzierung von Wissensarten*: Unterschiedliche Handlungen erfordern unterschiedliche Wissensarten.

Auf der Ebene der Teilkompetenzen erfolgt die Differenzierung von Teilaufgaben bei allen bekannten Modellen ausschliesslich auf der Prozessebene, wobei verschiedene Differenzierungsprinzipien in einer Teilkompetenz gemischt vorkommen können (cf. Modell von TIMSS (1997), Tab. 3.6). Erwähnenswert ist das Modell der Gruppe Schecker (Nawrath et al., 2011; Schreiber et al., 2009), dessen innere Abgrenzung auf nur einem Differenzierungsprinzip basiert (cf. Tab. 3.7).

	Teilkompetenzen	**Teilaufgaben**	
WiA	Wissenschaftsverständnis	Grundzüge und Grenzen der Wissenschaft	**PrA**
		Beurteilen der Aussagekraft von Modellen	
		Naturwissenschaft und Gesellschaft	
	Wissenschaftliches Denken	Naturwissenschaftliche Fragen und Hypothesen formulieren	**PrZ**
		Untersuchungen planen und durchführen	
		Daten auswerten: Mathematisierung	
		Empirische Daten interpretieren	
		Beobachten, Untersuchen, Beschreiben, Vergleichen, Bestimmen, Experimentieren	**PrA**
		Modelle zur Erkenntnisgewinnung nutzen	
	Manuelle Fertigkeiten	Mikroskopieren	**PrA**
		Chemische Nachweise / physikalische Messungen	
		Sicherheitsaspekte im Labor	

Tabelle 3.2 – Innere Struktur der experimentellen Kompetenz basierend auf den KMK-Biologiestandards gemäss Mayer(2007, 178)

	Teilkompetenzen	Teilaufgaben	
WiA	Using scientific inquiry	Design or critique aspects of scientific investigations	**PrZ**
		Conduct scientific investigations using appropriate tools and techniques	
		Identify patterns in data and/or relate patterns in data to theoretical models	
		Use empirical evidence to validate or criticize conclusions about explanations and predictions	
	Using technological Design	Propose and criticize solutions to problems, given criteria and scientific constraints	**PrA**
		Identify scientific tradeoffs in design decisions and choose among alternative solutions	
		Apply science principles or data to anticipate effects of technological design decisions	

Tabelle 3.3 – Innere Struktur der experimentellen Kompetenz bei NAEP(2008, 72ff)

	Teilkompetenzen	Teilaufgaben	
PrA	Bewusst wahrnehmen	Phänomene (Lebewesen, Gegenstände, Situationen, Prozesse) aufmerksam betrachten, genauer erkunden, beobachten, beschreiben und vergleichen.	
	Geeignete Werkzeuge, Instrumente und Materialien auswählen und verwenden ...	... für Erkundungen, Untersuchungen, Experimente und technische Konstruktionen	
PrZ	Fragen, Probleme und Hypothesen aufwerfen, ...	... um Beobachtungen, Entdeckungen und technische Konstruktionen zu ermöglichen und zu steuern	
	Erkundungen, Untersuchungen und Experimente durchführen	Fragen und Probleme aufgrund von Beobachtungen und Vorkenntnissen aufwerfen	**PrZ**
		Erkundungen, Untersuchungen oder Experimente planen und durchführen	
		Daten sammeln und auswerten, Hypothesen überprüfen bzw. Sachverhalte und Regelhaftigkeiten erkennen und festhalten	
	Über Ergebnisse und Untersuchungsmethoden nachdenken	Ergebnisse und Schlussfolgerungen aus Untersuchungen, Erkundungen und Experimenten beurteilen und bewerten	**PrZ**
		Frage- und Problemstellungen, Versuchsanlagen, Untersuchungs- und Messmethoden sowie technische Konstruktionen reflektieren, hinterfragen und dazu Verbesserungen vorschlagen.	

Tabelle 3.4 – Innere Struktur der experimentellen Kompetenz bei HarmoS(2008, 46)

	Teilkompetenzen	Teilaufgaben	
PrA	Use of apparatus and measuring instruments	Using measuring instruments	**PrA**
		Estimating physical quantities	
	Observation	Making and interpreting observations	
	Interpretation and application	Interpreting presented information (judging the applicability of a given generalisation)	**PrA**
		Applying science concepts to make sense of new information (generating new hypothesis)	
PrE	Planning of investigations	Planning parts of investigations	**PrE**
		Planning entire investigations	
	Performance of investigations	Performance of entire investigations	

Tabelle 3.5 – Innere Struktur der experimentellen Kompetenz bei APU(1988a, 2)

	Teilkompetenzen	Teilaufgaben	
WiA	Scientific investigation	Desiging and conducting investigations	**PrZ**
		Interpreting investigational data	
		Formulating conclusions from investigational data	
	Using scientific procedures	Using apparatus or equipment	**PrA**
		Conducting routine experimental operations	**PrZ**
		Gathering data	
		Organizing, representing, and interpreting data	

Tabelle 3.6 – Innere Struktur der experimentellen Kompetenz bei TIMSS(1997, 114ff)

	Teilkompetenzen	Teilaufgaben	
PrZ	Planung	Vorgegebene Fragestellung klären	**PrZ**
		Fragestellung entwickeln	
		Erwartungen formulieren	
		Hypothese bilden	
	Durchführung	Geräte zusammenstellen	**PrZ**
		Versuchsanordnung aufbauen	
		Messungen durchführen	
		Messungen dokumentieren	
	Auswertung	Messdaten verarbeiten	**PrZ**
		Messdaten interpretieren	
		Ergebnisse auf Fragestellung / Erwartung / Hypothese beziehen	

Tabelle 3.7 – Innere Struktur der experimentellen Kompetenz bei Schreiber et al.(2009, 93)

Aufgrund der Zerlegung von Teilkompetenzen können zwei Grundmodelle unterschieden werden. Beim *wissensbasierten Modelltyp* werden die Teilkompetenzen anhand spezifischer Wissensarten unterschieden, die für die erfolgreiche Kompetenzausübung benötigt werden. Unterschieden wird u. a. das Wissen über die Natur der Naturwissenschaften, das naturwissenschaftliche Methodenwissen und praktisches Fertigkeitswissen (cf. Modell von Mayer (2007), Tab. 3.2) sowie technisches Verständnis (cf. Modell von NAEP (2008), Tab. 3.3). Beim *prozessbasierten Modelltyp* erfolgt mit den Teilkompetenzen eine Differenzierung von experimentellen Prozessen. Hierbei kommen drei Differenzierungsprinzipien zur Anwendung: Die Differenzierung von Prozessalternativen, die Differenzierung einer Prozessabfolge (cf. das Modelle von HarmoS (2008), Tab. 3.4) sowie die Differenzierung einer Prozesserweiterung (cf. Modell von APU (1988a), Tab. 3.5).

Kompetenzprogressionsmodelle. Unter den Progressionsmodellen werden zwei Arten unterschieden: *Stufenmodelle* beschreiben die Anforderungsprogression von Kompetenzausprägungen, die in einer repräsentativen Jahrgangsstichprobe besteht (cf. Rost, 2004b, 664ff). *Entwicklungsmodelle* beschreiben demgegenüber die Abfolge von Kompetenzausprägungen beim individuellen Lernprozess oder in der individuellen Entwicklung. Der Zusammenhang zwischen den beiden Modellarten wird von Schecker und Parchmann (2006, 56f) wie folgt beschrieben: "Selbst wenn man davon ausgeht, dass Stufen naturwissenschaftlicher Kompetenz sinnvoll hierarchisch zu beschreiben sind [...], bleibt es eine Hypothese, dass damit gleichzeitig eine Abfolge einhergeht. Empirisch bisher gar nicht geklärt ist, in welcher Weise und in welcher Verknüpfung sich die Ausprägungen naturwissenschaftlicher Kompetenz beim Individuum zeitlich entwickeln [...]." Es stellt sich also u. a. die Frage, inwieweit die "natürliche" Kompetenzentwicklung infolge der Progression in der Aufgabenstellung oder infolge der Verbesserung der Aufgabenlösungen fortschreitet. Die Untersuchung dieser Frage wird strukturell durch den Umstand erschwert, dass sich mit zunehmender Erfahrung die Anforderungen der gestellten Aufgaben und Problemstellungen von selbst verändern. Dieses Problem zeigt sich exemplarisch an dem von Hammann (2004) aus unterschiedlichen Forschungsquellen zusammengetragenen Entwicklungsmodell zum Umgang mit Hypothesen, Variablen und Daten beim Experimentieren. Eine Analyse der Progressionen mit Hilfe der Progressionsdimensionen Aufgabenumfang $[A]$, Eigenständigkeit $[E]$, Problemkomplexität $[P]$, Prozessqualität $[Q]$ und Transferumfang $[T]$ legt prima facie die Vermutung nahe, dass die Entwicklung in den unteren Primarklassen ausschliesslich in Form eines Zuwachs der Qualität, mit der experimentiert wird, stattfindet und erst in den oberen Klassen Aufgaben mit grösserem Umfang, Transfer oder höherer Komplexität bewältigt werden (cf. Tab. 3.8, 3.9 und 3.10: Eine Änderung der Graustufe deutet einen Progressionsschritt in der entsprechenden Dimension an). Die-

Stufe	Kompetenzausprägung	[A]	[E]	[P]	[Q]	[T]
Primarstufe	Keine Hypothese beim Experimentieren					
Klasse 5	Unsystematische Suche nach Hypothesen					
Klasse 5	Systematische Suche nach Hypothesen					
Klasse 7	Systematische Suche nach Hypothesen und erfolgreiche Hypothesenrevision					

Tabelle 3.8 – Deskriptiv-spekulatives Entwicklungsmodell für das Formulieren von Hypothesen im Sinne von der "Suche im Hypothesensuchraum" gemäss Hammann(2004, 201)

Stufe	Kompetenzausprägung	[A]	[E]	[P]	[Q]	[T]
Primarstufe	Unsystematischer Umgang mit Variablen					
Klasse 5	Teilweise systematischer Umgang mit Variablen					
Klasse 5	Systematischer Umgang mit Variablen in bekannten Domänen					
Klasse 6	Systematischer Umgang mit Variablen in unbekannten Domänen					

Tabelle 3.9 – deskriptiv-spekulatives Entwicklungsmodell für das Experimentieren im Sinne von "Suche im Experimentiersuchraum" gemäss Hammann(2004, 200)

ser Schluss ist jedoch nicht gerechtfertigt – und wurde vom Autor in dieser Auslegung wahrscheinlich auch nicht intendiert –, da die dem Entwicklungsmodell zugrundeliegenden Daten mit unterschiedlichen Testaufgaben ermittelt wurden, die sich in Bezug auf Inhalt, Umfang und Komplexität schlecht vergleichen lassen. Die Modelle von Hammann (2004) sind in Bezug auf die Progression eines spezifischen Kriteriums der Prozessqualität (im folgenden werden wir ein solches Kriterium als Massstab bezeichnen) deskriptiv, bezüglich der Verknüpfung von Progressionsdimensionen rein spekulativ. Die Frage bleibt bisher unbeantwortet, inwiefern die dargestellten Verknüpfungen der Progressionsdimensionen eine "wahre" Kompetenzentwicklung wiedergeben.

Die Überprüfung von Kompetenzstandards stützt auf Stufenmodellen ab. Dabei gilt es zwischen Modellen zu unterscheiden, die die Stufungen der experimentellen Kompetenz als Ganzes (e. g. das Modell der Scientific literacy von Bybee, 1997) oder einzelner Teilkompetenzen (e. g. die Modelle von Hammann, 2004; HarmoS, 2010; Mayer et al., 2008) beschreiben. Da die unter dem Konstrukt der experimentellen Kompetenz zusammengefassten Prozesse sehr heterogen sind, kann eine allgemeine Progression nur mit Hilfe allgemeiner, problemunspezifischer Fähigkeiten, die keinen Bezug zu konkreten Testaufgaben zulassen, erfasst werden. Demgegenüber beziehen sich Modelle für Teilkompetenzen meist auf problemspezifische Fähigkeiten. Sofern dadurch ein konkreter Aufgabenbezug gege-

Stufe	Kompetenzausprägung	[A]	[E]	[P]	[Q]	[T]
Primarstufe	Daten werden nicht auf Hypothesen bezogen					
Klasse 5	Unlogische Analyse der Daten					
Klasse 6	Weitgehend logische Analyse der Daten, jedoch Probleme bei der Bewertung von Daten, die den eigenen Erwartungen widersprechen					
ab Klasse 7	Daten werden in adäquater Weise zur Überprüfung von Hypothesen herangezogen					

Tabelle 3.10 – Deskriptiv-spekulatives Entwicklungsmodell für die Teilkompetenz «Auswerten von Daten» im Sinne von der Überprüfung einer Hypothese gemäss Hammann(2004, 202)

Stufe	Kompetenzausprägung	[A]	[E]	[P]	[Q]	[T]
1	Einfache naturwissenschaftliche Fragen auf Phänomenebene stellen					
2	Naturwissenschaftliche Fragen nach Zusammenhang zweier Variablen formulieren					
3	Naturwissenschaftliche Fragen zu einem quantitativen Zusammenhang von Variablen formulieren					
4	Naturwissenschaftliche Fragen nach einem verallgemeinerten Zusammenhang formulieren					
5	Eigene naturwissenschaftliche Fragen zur Problemlösung formulieren					

Tabelle 3.11 – Spekulatives Stufenmodell für die Teilkompetenz «Fragestellung formulieren» gemäss Mayer(2008, 67)

Stufe	Kompetenzausprägung	[A]	[E]	[P]	[Q]	[T]
1	Einfache, testbare Hypothese generieren					
2	Hypothesen mit Analogie begründen					
3	Hypothesen auf der Basis von Kompetenzverständnis begründen					
4	Generalisierende Hypothese formulieren					
5	Alternative Hypothesen berücksichtigen					

Tabelle 3.12 – spekulatives Stufenmodell für die Teilkompetenz «Hypothesen generieren» gemäss Mayer(2008, 67)

ben ist, können Progressionsmodelle mit Hilfe der Progressionsdimensionen $[A,E,P,Q,T]$ analysiert werden. Hier unterscheiden wir eindimensionale Stufenmodelle, bei denen die Progression nur in einer Dimension fortschreitet (cf. Tab. 3.11, 3.12 und 3.13), und mehrdi-

Stufe	Kompetenzausprägung	[A]	[E]	[P]	[Q]	[T]
1	Eine Variable identifizieren					
2	Veränderte und zu messende Variable in Beziehung setzen					
3	Kontrollvariable berücksichtigen					
4	Stichprobe, Messwiederholung und Versuchsdauer berücksichtigen					
5	Untersuchungsmethoden, Genauigkeit, Fehler abwägen					

Tabelle 3.13 – Spekulatives Stufenmodell für die Teilkompetenz «Planung einer Untersuchung» gemäss Mayer(2008, 67)

Stufe	Kompetenzausprägung	[A]	[E]	[P]	[Q]	[T]
1	Sch. können zu vertrauten Lebewesen alltäglichen Gegenständen, Situationen und Prozessen einfache Fragen aufwerfen.					
2	Sch. können zu vertrauten Lebewesen, alltägliche Gegenstände, Situationen und Prozessen Fragen aufwerfen.					
3	Sch. können zu vertrauten Lebewesen, alltäglichen Gegenständen, Situationen und Prozessen verschiedenartige Fragen und einfache Probleme aufwerfen.					
4	Sch. können zu Lebewesen, Gegenständen, Situationen und Prozessen aus ihrer Umgebung verschiedenartige Fragen und einfache Probleme aufwerfen.					

Tabelle 3.14 – Normatives Stufenmodell für die Teilkompetenz «Fragen, Probleme und Hypothesen aufwerfen» im 2. Schuljahr gemäss HarmoS(2008, 79)

mensionale Modelle, welche auch Aussagen zur Verknüpfung der Progressionsdimensionen machen (cf. Tab. 3.14 und 3.15).

3.2.3 Der experimentelle Assessmentdiskurs

Die Forschung zur Messung experimenteller Kompetenz kann als «Problemlöseprozess» im Sinne von David Klahrs und Kevin Dunbars "Scientific discovery as dual search"[25] interpretiert werden, bei dem das Ziel verfolgt wird, Kompetenzstrukturen und Kompetenzprogressionen möglichst valide und möglichst effizient zu unterscheiden und zu messen. Die gemäss Klahr und Dunbar beim Lösungsprozess parallel laufenden und sich gegenseitig bedingenden "Suche im Hypothesensuchraum" und "Suche im Experimentiersuchraum"[26] entsprechen grosso modo dem Modell- und dem Assessmentdiskurs. Während im Modelldiskurs die Hypothesen entstehen, auf deren Basis der Assessmentdiksurs stattfindet, generiert der Assessmentdiskurs Daten, anhand derer die Modelle überprüft werden. Auf

[25] cf. Klahr und Dunbar (1988); Klahr (2000)

[26] Übersetzung von Hammann (2004), vgl. hierzu Hammann (2007); Hammann et al. (2007)

Stufe	Kompetenzausprägung	[A]	[E]	[P]	[Q]	[T]
1	Sch. können zu vorgegebenen Fragen mit vorgegebenem Material angeleitet einfache Erkundungen und Untersuchungen durchführen.					
2	Sch. können zu vorgegebenen einfachen Fragen und Hypothesen mit vorgegebenem Material angeleitet einfache Erkundungen und Untersuchungen durchführen sowie zu den Fragen und Hypothesen subjektiv Stellung nehmen.					
3	Sch. können zu vorgegebenen einfachen Fragen und Hypothesen mit teils vorgegebenem Material angeleitet einfache Erkundungen und Untersuchungen durchführen, Daten sammeln und auswerten (mögliche Gesetzmässigkeiten ansatzweise erkennen), sowie zu den Fragen und Hypothesen sachbezogen Stellung nehmen.					
4	Sch. können zu vorgegebenen einfachen Fragen und Hypothesen mit teils vorgegebenem Material angeleitet einfache Erkundungen und Untersuchungen planen und durchführen, Daten sammeln und auswerten (mögliche Gesetzmässigkeiten ansatzweise erkennen) sowie damit zu den Fragen und Hypothesen sinnvoll Stellung nehmen.					

Tabelle 3.15 – normatives Stufenmodell für die Teilkompetenz «Erkundungen, Untersuchungen oder Experimente durchführen» für Ende 2. Schuljahr gemäss HarmoS (2008, 80)

der dualen Suche nach adäquaten Modellen und validen Assessments stellen sich u. a. folgende Fragen:

1. Welche Kompetenzstrukturen und Kompetenzprogressionen lassen sich valide und effizient messen?

2. Mit welchen Messinstrumenten können Kompetenzstrukturen und Kompetenzprogressionen valide und effizient gemessen werden?

3. Bei welchen Populationen können Kompetenzstrukturen und Kompetenzprogressionen gemessen werden?

Die drei Fragen lassen sich nicht unabhängig voneinander beantworten, vielmehr verschmelzen sie zu einer einzigen Frage der Form: Welche Kompetenzstrukturen und Kompetenzprogressionen lassen sich mit welchen Messinstrumenten bei welchen Populationen valide und effizient messen? Im Folgenden soll trotzdem der Versuch unternommen werden, empirische Ergebnisse zur Kompetenzmessung anhand der drei Teilfragen zu strukturieren. Dabei werden Resultate aus der aktuellen Kompetenzforschung aus dem deutschspra-

chigen Raum[27], Resultate aus der Tradition der large-scale Experimentiertests von APU[28], NAEP[29], TIMSS[30] und HarmoS[31] sowie ausgewählte Resultate der Kognitionsforschung zum wissenschaftlichen Denken und Problemlösen[32] berücksichtigt.

3.2.3.1 Empirische Ergebnisse zu Kompetenzstrukturen und Kompetenzentwicklungen

Die genannte Literaturliste enthält Beiträge zu drei Themenkreisen: zur Modellierung und Messung einer einzelnen Kompetenz, zur Differenzierung von mehreren Kompetenzen (innere oder äussere Abgrenzung) und zur Modellierung und Messung einer Kompetenzprogression. Die Tabelle 3.16 ordnet die verschiedenen Beiträge nach deren Zweck, wobei unterschieden wird, ob es sich bei den behandelten Konstrukten um eine umfassende Kompetenz oder um Teilkompetenzen handelt. Die Literaturliste teilt sich zudem auf in Beiträge qualitativer Kompetenzanalysen und in Beiträge "quantitativer" Kompetenzmessungen (siehe kursive Einträge in Tab. 3.16). Erstere liefern Erkenntnisse über das qualitative Lösungsverhalten bestimmter Populationen bei bestimmten kompetenzorientierten Aufgaben. Letztere enthalten Resultate, die im direkten Zusammenhang mit konkreten Messinstrumenten und mit Bezug auf Analysen von Itemschwierigkeiten entstanden sind. Im Folgenden wird ein Überblick über die Literatur gegeben. Die Zusammenfassung erfolgt gemäss der Tabelle 3.16 geordnet nach der Art und dem Inhalt der Resultate (qualitativ oder quantitativ bzw. Konstrukt, Struktur oder Progression).

Qualitative Ergebnisse

Die Kognitionsforschung hält eine Fülle an qualitativen Erkenntnissen über das wissenschaftliche Argumentieren und Handeln von Kindern und Erwachsenen sowie von Laien und Expertinnen bereit, die für die Kompetenzmodellierung und Kompetenzmessung

[27] Ehmer und Hammann (2007); Grube et al. (2007); Hammann (2004); Hammann et al. (2006, 2007, 2008); Klos et al. (2008); Mannel et al. (2010); Mayer et al. (2008); Möller et al. (2007); Nawrath et al. (2011); Schreiber et al. (2009); C. von Aufschnaiter und Rogge (2010); Walpuski (2010); Wellnitz und Mayer (2008); Wellnitz et al. (2010)

[28] APU (1988a, 1988b, 1989); Toh und Woolnough (1990); Woolnough und Toh (1990)

[29] Ayala et al. (2001, 2002); Bass et al. (2002); Baxter und Glaser (1998); Baxter und Shavelson (1994); Baxter et al. (1995, 1992); Rosenquist et al. (2000); Ruiz-Primo et al. (1993); Ruiz-Primo und Shavelson (1996); Shavelson et al. (1993, 2002, 1999); Solano-Flores et al. (1999); Zimmerman und Glaser (2001)

[30] TIMSS (1997)

[31] Gut und Labudde (2010); HarmoS (2008); Labudde, Metzger und Gut (2009); Ramseier et al. (2011)

[32] Chinn und Brewer (1998); Chinn und Malhotra (2002); Dunbar (1993); Dunbar und Klahr (1989); Klahr (2000); Klahr und Dunbar (1988); Klayman und Ha (1987, 1989); Koslowski (1996); D. Kuhn (1989); D. Kuhn und Phelps (1982); D. Kuhn et al. (1988); Samarapungavan (1992); Schauble (1990); Schauble, Glaser et al. (1991); Schauble, Klopfer und Raghavan (1991); Schauble et al. (1992); Siegler und Liebert (1975); Tschirgi (1980); Wason (1960)

	experimentelle Kompetenz	**experimentelle Teilkompetenzen**
Konstrukt	*[Kl.2,6,9] HarmoS (2008)* [gr.4,8] Bass et al. (2002) *[gr.5-6] Baxter und Shavelson (1994); Shavelson et al. (1993)* *[gr.5-6,13] Rosenquist et al. (2000)* *[gr.13] Toh und Woolnough (1990); Woolnough und Toh (1990)* *[st.] Ruiz-Primo et al. (1993)*	[gr.1,3,5] Samarapungavan (1992) [gr.2,4,6/ust.] Tschirgi (1980) [gr.4-5] D. Kuhn und Phelps (1982) [gr.4-6] Chinn und Malhotra (2002) [gr.5-6] Schauble (1990) [gr.5-6] Schauble, Klopfer und Raghavan (1991) [gr.6] Zimmerman und Glaser (2001) [gr.6,9/ust.] Koslowski (1996)* [gr.6,9/gst./ad.] D. Kuhn et al. (1988); D. Kuhn (1989) [gr.8] Siegler und Liebert (1975) [ust.] Wason (1960) [ust.] Schauble, Glaser et al. (1991) [ust.] Chinn und Brewer (1998) [st.] Klayman und Ha (1989) [ad.] Klahr und Dunbar (1988); Dunbar und Klahr (1989) [ad.] Dunbar (1993)
Abgrenzung	*[gr.5] Baxter et al. (1992)* *[Kl.6,9] Labudde, Metzger und Gut (2009); Gut und Labudde (2010)* *[Kl.7,12] Klos et al. (2008)*	*[gr.4,8] TIMSS (1997)* *[gr.5] Solano-Flores et al. (1999)* *[Kl./gr.5-6] Hammann et al. (2007, 2008)* *[Kl.5-10] Grube et al. (2007); Mayer et al. (2008)* *[J.11,13,15] APU (1988a, 1988b, 1989)** *[gr.7-9] Tannenbaum (1971)* *[gr.7-10] Fraser (1980)* *[Kl.9] Wellnitz und Mayer (2008); Wellnitz et al. (2010)* *[Kl.10] Schreiber et al. (2009)* *[ad.] Ayala et al. (2001, 2002)*
Progression	[Sek.I,II] C. von Aufschnaiter und Rogge (2010)** [ust.] Schauble et al. (1992)	*[gr.4-5] Baxter et al. (1995)* [indef.] Hammann et al. (2006) [bis Kl.7] Hammann (2004) *[Kl.5] Mannel et al. (2010)* *[Kl.5-10] Mayer et al. (2008); Möller et al. (2007)* [Kl.8] Ehmer und Hammann (2007) *[Sek.I] Walpuski (2010)*

Tabelle 3.16 – Überblick über ausgewählte Literatur zur Messung experimenteller Kompetenz geordnet nach Konstrukten (individuelle Messung oder * = auch Gruppenassessment, ** = nur Gruppenassessment; Population: gr. = grade, Kl. = Klassenstufe, ust. = undergraduate students, gst. = graduate students, st. = students, ad. = adults, J. = Alter in Jahren). Kursiv: quantitative Beiträge, nicht kursiv: qualitative Beiträge.

wichtige Anhaltspunkte liefern. Diese Informationen werden durch Resultate aus quantitativen Messungen ergänzt. Um die Qualität von Prozessen sinnvoll einschätzen und eine Progression festlegen zu können, muss bekannt sein, welche Denk- und Handlungsvarianten beim Lösen bestimmter Probleme vorkommen und welche strukturellen Fehlverhalten – sogenannte *Bias* – auftreten. Qualitative Kompetenzanalysen allein reichen jedoch nicht

aus, um Kompetenzstrukturen und Kompetenzprogressionen zu definieren. Hierfür werden zusätzlich Ergebnisse aus quantitativen Analysen benötigt.

Die in der zitierten Forschung verwandten Modelle zum experimentellen Problemlösen

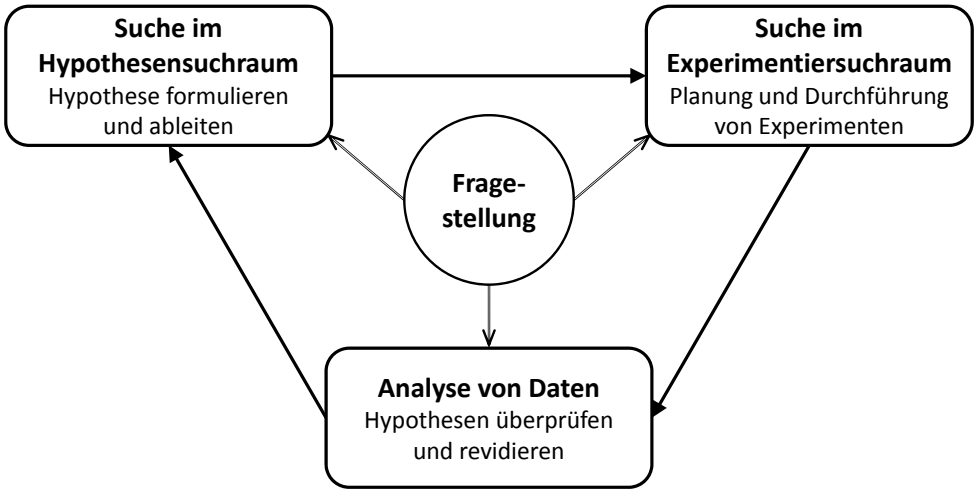


Abbildung 3.3 – Vierteiliges Modell zur Analyse experimenteller Problemlösekompetenz

basieren allesamt auf der grundsätzlichen Unterscheidung der drei Teilprozesse «Suche im Hypothesensuchraum», «Suche im Experimentiersuchraum» und «Datenanalyse» und der Annahme einer logischen zirkulären Abfolge der drei Teilprozessen (siehe Abb. 3.3). Die Trias der Teilprozesse geht auf das Modell der *Scientific discovery as dual search* von Klahr und Dunbar (1988) zurück, das den Prozess des experimentellen Problemlösens als eine gleichzeitige oder zirkuläre Suche von Hypothesen und Experimenten und der damit verbundenen Datenanalyse versteht. Die Trias ist zudem eine Reduktion des linearen Modells des idealisierten Experimentierens (cf. Abb. 3.2). Um den verschiedenen Modellansätzen der psychologischen und fachdidaktischen Kompetenzforschung gerecht zu werden, wird das Basismodell um das Element «Fragestellung» erweitert, das sowohl als eigenständiger Teilprozess als auch als Vorgabe der Problemziele interpretiert werden kann. Grundsätzlich kann eine Problemlöseaufgabe bei entsprechender Vorgabe der Fragestellung mit jedem beliebigen Teilprozess starten und nur einzelne Elemente und ihre Verbindungen ansprechen. Zudem können in der zirkulären Abfolge der Teilprozesse einzelne Elemente übersprungen werden. Teilprozesse können auch zusammengefasst und gemeinsam evaluiert werden. Es ergeben sich somit vier Aspekte, in welchen sich die verwandten Forschungsansätze zum Problemlösen wesentlich unterscheiden.

1. *Fragestellung*: Resultate zum experimentellen Problemlösen sind bei unterschiedlicher Fragestellung (e. g. einen kausalen Zusammenhang zu entdecken, eine technische Funktion zu analysieren oder einen Effekt zu optimieren) nur bedingt vergleichbar.

2. *Fokus*: Nicht bei allen Forschungsansätzen stehen alle drei Teilprozesse im Fokus der Beobachtung. Teilprozesse werden zum Teil gar nicht oder nur zusammen mit anderen Teilprozessen gemessen (innere Abgrenzung).

3. *Umfang*: Die gestellten Aufgaben umfassen nicht immer alle Teilprozesse. Einzelne Teilprozesse werden sowohl integriert als auch isoliert modelliert.

4. *Zirkularität*: Das experimentelle Problemlösen kann als linearer, nicht zirkulärer Vorgang modelliert werden, bei dem eine Hypothese am Anfang und deren Beurteilung am Schluss steht. Demgegenüber kann das experimentelle Problemlösen auch als zirkulärer Prozess verstanden werden, der u. a. auch die Revision von Hypothesen beinhaltet.

Für die nachfolgende Zusammenstellung relevanter Forschungsresultate zu Kompetenzkonstrukten werden wir, wo dies erforderlich ist, zwischen dem *vierteiligen Problemlösemodell* und dem um den Teilprozess «Fragestellung» *reduzierten dreiteiligen Problemlösemodell* unterscheiden.

A. Überprüfung von (gegebenen) Hypothesen

i. Existenz problemspezifischer Strategien und Progressionen: Es gibt erkennbare Präferenzen für bestimmte problemspezifische Strategien, die vom Alter und der Expertise der Testpersonen abhängen. Beispielsweise zeigen Klayman und Ha (1989) und Wason (1960), dass Hypothesen, die gleichzeitig Aussagen darüber machen, was der Fall ist und was nicht der Fall ist, tendenziell häufiger an positiven Beispielen getestet werden, die gemäss der Hypothese der Fall sind, als an negativen Beispielen, die gemäss Hypothese nicht der Fall sind. Diese so genannte *Positive test strategy* führt zu unzureichend verifizierten Hypothesen, weil zutreffende negative Beispiel nicht evaluiert werden.[33]

ii. Einfluss des Aufgabenkontextes auf Strategie: Die gewählten Strategien hängen nicht nur vom logischen Inhalt der Hypothese, sondern auch von der sprachlichen Form ab: Tschirgi (1980) und (Zimmerman & Glaser, 2001) weisen z. B. nach, dass logisch äquivalente Behauptungen sowohl bei Primarschülerinnen und -schülern als auch bei Erwachsenen unterschiedliches Experimentierverhalten bewirken, wenn einmal ein positiver Effekt und einmal ein negativer Effekt behauptet wird. Die Strategien

[33]Die Vorliebe einer Teststrategie muss nicht unbedingt ein Fehlverhalten anzeigen. Überlegungen zur Wahrscheinlichkeit, mit welcher eine bestimmte Strategie ein Gegenbeispiel erwartet lässt, kann die Präferenz einer Teststrategie rational begründen. Je nach Problemstellung sind z. B. die *Positive test strategy* und die *Negative test strategy* nicht gleich informativ (Klayman & Ha, 1987).

werden mehrheitlich so gewählt, dass positive Effekte reproduziert und negative Effekte vermieden werden.

B. Experimetelle Problemlösestrategien beim Untersuchen kausaler Zusammenhänge

iii. *Strategie-Variablität*: Kinder wie auch Erwachsene zeigen zu einer gegebenen Problemlöseaufgabe eine breite Palette an Lösestrategien (D. Kuhn & Phelps, 1982,39; Schauble et al., 1992, 342).

iv. *Strategie-Rekurrenz*: Die Entwicklung adäquater Experimentierstrategien verläuft nicht monoton. Vielmehr fallen Schülerinnen und Schüler bei wiederholten Experimentiersequenzen auf alte ungültige Verhaltensmuster zurück (D. Kuhn & Phelps, 1982, 31f).[34]

v. *Erfolgreiche Strategien*: Empirisch kann bei Testpersonen ein positiver Zusammenhang zwischen der Adäquatheit kausaler Denkmodelle und der Entwicklung der Experimentierstrategien hergestellt werden. Sowohl Kinder als auch Erwachsene, die zum Inhalt einer Problemlöseaufgabe über höher entwickelte Theorien verfügen, zeigen erfolgreichere Experimentierstrategien (Schauble, 1990, 37f; Schauble et al., 1992, 342). Im Gegenzug macht, wer über höher entwickelte Experimentierstrategien verfügt, mehr Lernfortschritte. Insbesondere können auf das spezifische Problem bezogene Merkmale der Lösungsprozesse genannte werden, die gute und schlechte Lerner unterscheiden: "These included activities in the class of evidence generation (controlling extraneous variation), evidence interpretation (generating and evaluating alternative hypotheses, inferring regularities in the data, producing sufficient evidence to support a hypothesis), dta management /sytsematic data recording), and planning (developing plans that are goal oriented rather than procedure orientied)" (Schauble et al., 1992), 342). Gute Problemlöser scheinen sich generell durch den effektiven Gebrauch alternativer Hypothesen auszuzeichnen (Klayman & Ha, 1989, 603).

vi. *Fehlerhafte Strategien*: Die kognitionspsychologische und fachdidaktische Literatur beschreibt unterschiedliche Fehlverhalten beim experimentellen Problemlösen (für eine Zusammenfassung siehe Hammann, 2004). Die Fehlerhaftigkeit der beschriebenen Strategien ergibt sich allein aus einer logischen Analyse der Aufgabenstellung.

[34]Die Rekurrenz im Experimentierverhalten zeigt sich auch bei den untersuchten Lernvorgängen während des Experimentierens. Das Erfassen von Konzepten mit Hilfe von manipulativer Exploration mit Geräten und Materialien bedingt auch auf konzeptionell hohem Niveau immer wieder den Rückschritt auf experimentell weniger komplexe Konzeptebenen (C. von Aufschnaiter & Rogge, 2010, 106).

Diese Fehlverhalten gelten daher auch nur für bestimmte Aufgabentypen und lassen sich nicht ohne Weiteres verallgemeinern. Einige Fehlverhalten treten bei empirischen Untersuchungen gehäuft auf und werden daher als allgemeine *Bias* gehandelt. Viele dieser als fehlerhaft taxierten Strategien erweisen sich jedoch bei genauerer Analyse nicht zwingend als falsch. Betrachtet man nicht nur die Experimentierstrategien, sondern bezieht in die Analysen auch die doxastischen Voraussetzungen der Testpersonen mit ein, lassen sich fehlerhafte Strategien meist rational begründen (Koslowski, 1996).[35] Tatsächlich kann jede beliebige Experimentierstrategie einer Testperson mit Hilfe von geeigneten Überzeugungen zu kausalen Zusammenhängen und Wahrscheinlichkeiten von Ereignissen rational erklärt werden. Eine Experimentierstrategie erweist sich daher erst dann als fehlerhaft, wenn ihre rationale Erklärung auf falschen Überzeugungen beruht (die sich daraus ergebenden Konsequenzen für das Assessment von Experimentierfähigkeiten werden im Abschnitt 4.3.3 diskutiert). Auch erweisen sich vermeintlich adäquatere Strategien nicht immer als die erfolgreicheren. Schülerinnen und Schüler der 5. und 6. Klasse zeigen beispielsweise beim Untersuchen kausaler Zusammenhängen zwei verschiedene Experimentierstrategien: Im Wissenschaftlermodus wird der Einfluss einzelner Variablen auf den Effekt zielgerichtet durch Variation der Variablen untersucht. Im Ingenieurmodus werden die Variablen so manipuliert, dass der gewünschte Effekt eintritt oder optimiert wird. Schauble, Klopfer und Raghavan (1991) zeigen, dass die erfolgreichsten Strategien diejenigen sind, die im Ingenieurmodus beginnen und im Wissenschaftlermodus enden.

C. Datenanalyse: Umgang mit anomalen Daten.

vii. Reaktionsmodi auf anomale Daten: Chinn und Brewer (1993, 1998) differenzieren sieben Modi, wie auf anomale Daten – Daten, die den eigenen Überzeugungen, i. e. der akzeptierten Theorie, widersprechen – reagiert wird.[36] Die Modi unterscheiden

[35] Ein sehr schönes Beispiel einer "fehlerhaften", aber rational begründbaren Strategie beim Schluss von einem Phänomen auf kausale Faktoren ist die von Ehmer und Hammann (2007) beschriebene so genannte Mit-Strategie (cf. Abs. 4.3.3).

[36] Folgende Modi werden unterschieden: *1. Ignoranz*: Die anomalen Daten werden ohne Begründung als nicht gültig betrachtet. *2. Ablehnung*: Die Daten werden mit Hilfe einer Begründung als ungültig erklärt. *3. Ausschluss*: Der Anwendungsbereich der akzeptierten Theorie wird so angepasst, dass die Theorie zu den Daten keine Aussage machen. *4. Schwebe*: Die Daten und die Theorie werden akzeptiert im Glauben, dass es für die Anomalität der Daten später eine Erklärung gibt. *5. Reinterpretation*: Die Daten und die Theorie werden zwar akzeptiert, jedoch ist man nicht sicher, ob es für die Anomalität der Daten später eine Erklärung gibt. *6. Periphere Theorierevision*: Die Daten werden so reinterpretiert, dass keine Theorierevision notwendig ist. Die Theorie wird peripher revidiert. *7. Theorierevision*: Die Theorie wird grundsätzlich revidiert.

sich im Grad und der Art der Interpretation der Gültigkeit der anomalen Daten, der Interpretation ihrer Aussagekraft für akzeptierte Theorie sowie in der Reinterpretation der akzeptierten Theorie, die eine Theorierevision beinhalten kann.

viii. Einfluss individueller doxastischer Voraussetzungen auf Datenanalyse: Empirische Untersuchungen zeigen, dass der Umgang mit anomalen Daten durch die Verankerung der bestehenden Überzeugung und das Vorhandensein wissenschaftlichen Hintergrundwissens – Wissen das als gültig vermutet wird, aber nicht Teil der evaluierten Theorie ist – beeinflusst wird (e. g. Chinn & Brewer, 1992, 1993). Kinder wie Erwachsene reagieren zudem in theoriebewahrender Weise, solange keine alternative Theorie zur Verfügung steht, welche die anomalen Daten plausibel erklärt (D. Kuhn, 1989). Bei der Wahl alternativer Theorien orientieren sich bereits Schülerinnen und Schüler der ersten Primarstufen daran, dass mit einer Theorie möglichst viele Daten möglichst präzise und möglichst konsistent erklärt werden (e. g. Burbules & Linn, 1988; Johsua & Dupin, 1987; Samarapungavan, 1992). Wenig Beachtung erhält von Unterstufenschülerinnen und -schülern hingegen die Idee, dass eine Theorie die Welt auf möglichst einfache Weise und ohne Gebrauch von ad hoc-Hypothesen erklären sollte (Samarapungavan, 1992).

ix. Fehlerhaftigkeit von Datenanalysen: Der von einem Subjekt gewählte Modus, auf anomale Daten zu reagieren, deutet im Allgemeinen kein Fehlverhalten an. Wissenschaftsphilosophische Studien zeigen, dass auch in der "hohen" Forschung viele Reaktionsmodi auftreten. Die Vielfalt der Reaktionsmodi, interpretiert als wissenschaftliches "anarchisches Verfahren" im Sinne von Feyerabend (1999 [1976], 238ff), kann auch als Bedingung für die Möglichkeit wissenschaftlichen Fortschritts verstanden werden. Im konkreten Beispiel einer bestimmten Laborsituation können hingegen Fehlverhalten eruiert und beschrieben werden. Die Ergebnisse der Kognitionsforschung zu diesem Thema sind nicht derart eindeutig, wie sie zuweilen dargestellt werden, und nicht jedes behauptete Fehlverhalten entpuppt sich bei genauerer Betrachtung als falsch (Koslowski, 1996). Trotzdem können Fehlverhalten und deren Entwicklung erkannt werden, wenn sie auf einem fehlerhaften Verständnis über die Natur der Naturwissenschaften basieren. So stellen D. Kuhn et al. (1988) fest, dass viele "Fehlverhalten" von Schulkindern auf der ungenügenden Unterscheidung von Theorie und Evidenz beruhen, i. e. es wird nicht unterschieden, ob eine Aussage mit einer Theorie oder einer Evidenz begründet wird. Dieses unsichere "Schwanken zwischen Theorie und Evidenz" nimmt mit zunehmendem Alter und Expertise ab (D. Kuhn, 1989, 676f).

Quantitative Ergebnisse

Quantitative, auf der Analyse von Itemschwierigkeiten basierende Untersuchungen belegen, dass mit Papier-und-Bleistift-Tests gewisse Teilkompetenzen, die beim Experimentieren erforderlich sind, reliabel gemessen und statistisch unterschieden werden können. Die Teilkompetenzen lassen sich auch zu einer übergeordneten Kompetenz zusammensetzen, die sich bei Vergleichen mit Fachwissenstests statistisch absetzt. Mit Hilfe von Modellen a priori können zudem Anforderungsprogressionen sowohl für diese übergeordnete Kompetenz als auch für die Teilkompetenzen modelliert werden. Dies ist bisher mit Hilfe von Experimentiertests nicht der Fall.

A. Abgrenzung von Kompetenzstrukturen

i. *Äussere Abgrenzung mit Hilfe von Experimentiertests*: Leistungsvergleiche von Experimentiertests und Fachwissenstests ergeben allgemein stark ausgeprägte Unterschiede. Die Korrelationen zwischen den Testarten für verschiedene Messarten pendeln zwischen den Werten .40 und .50 (Baxter et al., 1992; Baxter & Shavelson, 1994). Man vergleiche hierzu die Validierung des Schweizer Kompetenzmodells HarmoS Naturwissenschaften im Kapitel 5.1 auf S. 105).

ii. *Äussere Abgrenzung mit Hilfe von Papier-und-Bleistifttests*: Messungen von experimentellen Problemlösekompetenzen mit Papier-und-Bleistift-Tests lassen sich statistisch deutlich von Leistungsmessungen mit Fachwissenstests unterscheiden. Die Korrelationen zwischen den Leistungsmessungen betragen zwischen .07 und .38. Dies gilt sowohl für die Messung der experimentellen Kompetenz als Ganzes (Klos et al., 2008) als auch für die Messung einzelner Teilkompetenzen (Hammann et al., 2008).

iii. *Innere Abgrenzung: Differenzierung von Wissensarten*: Obwohl Experimentieraufgaben immer mehrere Wissensarten (e. g. deklaratives, prozedurales und schematisches Wissen) ansprechen, können Experimentieraufgaben entwickelt werden, die hauptsächlich auf eine Wissensart ausgerichtet sind. Ayala et al. (2002); Shavelson et al. (2002) legen dies für die Wissensdimensionen "Basic knowledge and reasoning", "Spatial-mechanical reasoning" und "Quantitativ science" am Beispiel dreier Experimentieraufgaben empirisch nahe.

iv. *Innere Abgrenzung: Differenzierung von Teilkompetenzen mit Hilfe von Experimentiertests*: In der TIMS-Studie wurden die drei experimentellen Teilkompetenzen «Scientific problem solving and Applying concept knowledge» («SPS»), «Using scientific procedures» («USP») und «Scientific investigating» («SI») (cf. das entsprechende Strukturmodell auf S. 35) bei 4. und 8. Klässlern mit einem internatio-

nalen large-scale Experimentiertest evaluiert. In beiden Jahrgangsgruppen wurden die Aufgaben der Teilkompetenz «SPS» signifikant schlechter gelöst als die Aufgaben der Teilkompetenzen «USP» und «SI». Während bei den Schülerinnen und Schüler der 4. Jahrgangsstufe die Teilkompetenz «USP» signifikant besser ausgeprägt ist als die Teilkompetenz «SI», gleichen sich die Differenzen in der 8. Jahrgangsstufe aus (TIMSS, 1997, 116).

v. *Innere Abgrenzung: Differenzierung von Teilkompetenzen mit Hilfe von Papier-und-Bleistift-Tests*: In Deutschland werden 2012 erstmals die Regelstandards evaluiert, wobei u. a. der Kompetenzbereich Erkenntnisgewinnung national überprüft wird. Im Rahmen der Testvorbereitungen wurden verschiedene Ansätze entwickelt, mit denen die Teilkompetenzen des Problemlösens (cf. Abb. 3.3) reliabel gemessen werden können. Dies gilt sowohl für dreiteilige Modelle (e. g. Mannel et al., 2010) als auch für vierteilige Modelle (e. g. Grube et al., 2007; Wellnitz et al., 2010). In Bezug auf die statistische Unterscheidung von Teilkompetenzen sind verschiedene Ansätze unterschiedlich erfolgreich. Keine Subdimensionen des dreiteiligen Problemlösemodells finden Klos et al. (2008) mit einem Multiple-choice-Test. Hammann et al. (2008) unterscheiden mit einem dreiteiligen Multiple-choice-Test zwei Subdimensionen, wobei die Teilkompetenzen «Suche im Hypothesensuchraum» und «Datenanalyse» zusammen eine Dimension und die Teilkompetenz «Suche im Experimentiersuchraum» eine zweite Dimension bilden. Der von Mayer et al. (2008) auf der Basis des vierteiligen Problemlösemodells entwickelte Papier-und-Bleistift-Test mit offenen Aufgaben differenziert deutlich drei Subdimensionen, wobei der statistische Unterscheide zwischen den Teilkompetenzen «Suche im Hypothesensuchraum» und «Datenanalyse» weniger deutlich ausfällt. Ein Test mit gemischten Formaten differenziert alle vier Subdimensionen hinreichend (Wellnitz et al., 2010). Faktoranalytische Untersuchungen der Tests ergeben jedoch bislang nur Ein-Faktor-Lösungen, die je nach Test bis zu 76% der Varianz der Subdimensionen erklären (Klos et al., 2008; Mayer et al., 2008, 71). Zusammenfassend kann festgehalten werden, dass Teilkompetenzen mit geschlossenen Testformaten mit höherer Reliabilität gemessen werden, mit offenen Testformaten jedoch statistisch besser unterschieden werden können.

B. Kompetenzprogressionen.

vi. *Zwei Ansätze der Progressionsmodellierung*: Ergebnisse aus der deutschen Kompetenzforschung belegen, dass Kompetenzprogressionen auf der Stufe der experimentellen Kompetenz als auch auf der Stufe der Teilkompetenzen modelliert werden können. Dabei können zwei Ansätze unterschieden werden. Ein Ansatz modelliert Kompetenzprogressionen ausschliesslich auf der Ebene der Aufgabenstellung (Mannel et

al., 2010; Walpuski, 2010). Dies entspricht der im Abschnitt 3.1.1 diskutierten Progressionsebene (KP1) (cf. S. 16). Der andere Ansatz bezieht auch die Ebene der Aufgabenlösung in die Modellierung der Itemschwierigkeit ein, i. e. beide Progressionsebenen (KP1) und (KP2) werden mitberücksichtigt (Mayer et al., 2008; Möller et al., 2007).

vii. Einfache Progression (KP1): Die Modellierung der Anforderungsprogression auf der Ebene der Aufgabenstellung erfolgt beim ersten Ansatz mit Hilfe der Konstrukte "Komplexität" und "Kognitive Prozesse" (Walpuski et al., 2008, 324f). Mit der Komplexität wird eine Art Inhaltskomplexität erfasst nach dem Schema «Ein Fakt», «Zwei Fakten», «Ein Zusammenhang», «Zwei Zusammenhänge» und «Konzept». Mit den Kognitiven Prozessen «Reproduzieren», «Selegieren», «Organisieren» und «Integrieren» wird eine Art Problemkomplexität erfasst. Ein large-scale Multiple-choice-Test ergab hochsignifikante Einflüsse der zwei Faktoren Komplexität und Kognitive Prozesse auf die Itemschwierigkeit (Mannel et al., 2010, 302f).

viii. Kombinierte Progressionen (KP1 und KP2): Beim zweiten Ansatz werden die Progressionen für vier Teilkompetenzen separat modelliert, wobei anhand von Aspekten der Problemkomplexität und der Prozessqualität für jede Teilkompetenz eine fünfstufige Skala formuliert wird. Ein Papier-und-Bleistift-Test zeigt, dass die Skalen sowohl mit den Altersstufen als auch mit dem Anforderungsniveau von Schulen auf der Sekundarstufe I korrelieren (Mayer et al., 2008).

3.2.3.2 Empirische Ergebnisse zu Messinstrumenten

Die Suche nach dem geeigneten Messinstrument für die experimentelle Kompetenz stellt sich im Licht der Kompetenzforschung als dreifache Suche heraus. Zum einen wird nach der geeigneten *Testart* gesucht. Zur Debatte stehen Experimentiertests (Hands-on-Tests), Paper-und-Bleistift-Tests (Hands-off-Tests) und Simulationsests, die heutzutage am Computer erfolgen. Die drei Testarten unterscheiden sich hinsichtlich ihrer *Determiniertheit* und der *Notwendigkeit von Manipulationen* (cf. Tab. 3.17). Für die Lösung von indeterminierten Aufgaben werden zusätzliche, in der Aufgabenstellung nicht enthaltene Informationen benötigt, die während des Lösungsprozesses "gewonnen" werden müssen. Manipulative Aufgaben erfordern praktische, handwerkliche Arbeiten, entweder weil die Informationsgewinnung diese erfordert oder weil ein handwerkliches Produkt hergestellt werden soll.

Die zweite Suche der Kompetenzforschung gilt der Frage, mit welcher *Messart* die

[37] Zum Beispiel ist die Plastilin-Aufgabe aus dem TIMSS-Experimentiertest manipulativ und trotzdem determiniert (cf. Labudde & Stebler, 1999, 27).

Testaufgaben ...	**... erfordern Manipulationen**	**.. erfordern keine Manipulationen**
... sind determiniert	gewisse Konstruktionstests[37]	Papier-und-Bleistift-Tests
... sind nicht determiniert	Experimentiertests	Simulationstests

Tabelle 3.17 – Klassifikation von Testarten für die experimentelle Kompetenz

Qualität der Lösungsprozesse bzw. Lösungsprodukte am besten beurteilen kann. Lässt man die Handlungen der Testpersonen durch Rater beobachten und rapportieren (*Fremdrapport* via Audio-, Video- oder Direktbeobachtung) oder vertraut man den Fähigkeiten der Testpersonen, ihre Handlungen selber valide zu rapportieren (*Eigenrapport*)? Dabei stellt sich die Frage, mit welchen Rapportformaten (*Antwortformaten*) sichergestellt werden kann, dass Schülerinnen und Schüler einerseits rapportieren, was sie gemacht haben, und andererseits auch das gemacht haben, was sie rapportieren?

Letztlich beschäftigt sich die Kompetenzforschung auch mit dem Problem der geeigneten *Kodierung*. Diese Frage ist eng verknüpft mit dem Problem der Validität der Tests und dem Problem der Progressionsmodellierung. Beide Probleme werden wir im Abschnitt 4.3.2 aufgreifen.

Die Forschungsliteratur zu den Fragen der geeigneten Testart und der geeigneten Messart ist punktuell und heterogen. Man vergleiche hierzu die Tabelle 3.18, welche die hier benutzte Literaturauswahl nach Test- und Messart ordnet. Im Folgenden werden Ergebnisse zu diesen beiden Frage zusammengestellt.

A. Ergebnisse zu Testarten

Gegen den Einsatz von large-scale Experimentiertests im Rahmen von Evaluationen nationaler Standards werden verschiedene Argumente angeführt: Experimentiertest sind kostspielig, zeitaufwändig und gelten als wenig valide und reliabel, weil die Kodierung auf komplexen menschlichen Urteilen beruht (Shavelson et al., 1993, 215f).

i. *Experimentiertest versus Papier-und-Bleistift-Test*: Experimentiertests korrelieren nur schwach mit Papier-und-Bleistift-Tests. Dies gilt sowohl für Tests, welche die experimentelle Kompetenz als Ganzes messen (Baxter & Shavelson, 1994; Shavelson et al., 1993; Solano-Flores et al., 1999)[38] als auch für Tests, die auf Teilkompetenzen

[38]Ein Vergleich von zwei Experimentieraufgaben, die durch direkte Beobachtung geratet wurden, mit zwei analogen Papier-und-Bleistift-Aufgaben ergab nur schwache Übereinstimmung (*short answer*: $.32 \leq r \leq .53$; *multiple-choice*: $.28 \leq r \leq .41$).

ausgerichtet sind (Hammann et al., 2008)[39]. Die Resultate deuten darauf hin, dass Papier-und-Bleistift-Tests und Experimentiertests unterschiedliches Wissen und unterschiedliche Fähigkeiten messen (Haertel & Linn, 1996, 68; Solano-Flores et al., 1999, 310).

ii. *Experimentiertest versus Simulationstest*: Computer-Simulationen bieten für large-scale Hands-on-Tests, sofern die Experimentieraufgaben sich für eine Simulation eignen[40], einen akzeptablen Ersatz für Experimentiertests (Shavelson et al., 1999). Hands-on-Tests korrelieren mit analogen Simulationstests zwar nur mässig[41], die ungenügende Testübereinstimmung lässt sich jedoch auf ein *instabiles, volatiles Experimentiererhalten* von Schülerinnen und Schülern in äquivalenten, zeitlich separierten Testsituationen zurückführen (Ruiz-Primo et al., 1993; Shavelson et al., 1991, 1992, 1999).[42] Die Strategie-Instabilität bewirkt auch bei beobachteten bzw. rapportierten Experimentiertests, die zeitlich separiert erfolgen, verminderte, mit denen von Computersimulationen vergleichbare Korrelationen (Shavelson et al., 1999, 69f). Die Instabilität des Experimentierverhaltens von Schülerinnen und Schülern fällt somit sowohl bezüglich der Testart als auch bezüglich der Aufgaben negativ ins Gewicht (Baxter & Shavelson, 1994, 294). Der hohe Varianzanteil zwischen Experimentier- und Simulationstests lässt sich auf den Mangel an transferfähigem Wissen (*partial knowledge*) zurückführen, der direkt oder im Zusammenwirken mit der Messmethode hohe Leistungsschwankungen bei Testpersonen verursacht (Rosenquist et al., 2000).

[39] Messungen der Teilkompetenz «Planning experiments» mit einem Experimentiertest und mit einem Multiple-choice-Test korrelieren nur sehr schwache ($r = .33$). Die Korrelation der Messungen der Teilkompetenz «Analysing data» ist sogar negativ ($r = -.27$). Hammann et al. (2008) erklären die Differenzen auf der Ebene der Items wie folgt: " [...] there is a big difference between multiple-choice items that present data from well-designed experiments and a laboratory test that allows pupils to interpret self-generated data (i. e. data that stems from the pupils' own experiments, whose design is often flawed and inconclusive)."

[40] Dies gilt u. a. nicht für die Paper-towel-Aufgabe von NAEP (Brown & Shavelson, 1996, 26ff) oder die Solarzellen-Aufgabe von HarmoS (cf. App. A.2).

[41] Testvergleiche mit zwei Problemlöseaufgaben ergaben bei Schülerinnen und Schülern der 5. und 6. Schulstufe eine Bandbreite bei den Korrelationen von $.47 \leq r \leq .55$, (Baxter & Shavelson, 1994; Shavelson et al., 1993).

[42] Die *Strategie-Instabilität* von Schülerinnen und Schülern korrespondiert zu einem gewissen Grad mit der auf S. 45 beschriebenen *Strategie-Rekurrenz*.

Messart	Hands-on-Test	Papier-und-Bleistift-Test	Computer-Simulation
Beobachtung	[gr.4-5] Baxter et al. (1995) [gr.4-6] Chinn und Malhotra (2002) [gr.5] Baxter et al. (1992) [gr.5-6] Schauble, Klopfer und Raghavan (1991) [gr.5-6] Baxter und Shavelson (1994); Shavelson et al. (1993) [Sek.I,II] C. von Aufschnaiter und Rogge (2010)** [ust.] Schauble et al. (1992) [st.] Ruiz-Primo et al. (1993) [ad.] Ayala et al. (2001, 2002)		[gr.5-6] Baxter und Shavelson (1994); Shavelson et al. (1993) [gr.5-6,13] Rosenquist et al. (2000) [Kl.10] Schreiber et al. (2009) [ust.] Wason (1960) [ust.] Schauble, Glaser et al. (1991) [st.] Klayman und Ha (1989) [ad.] Klahr und Dunbar (1988); Dunbar und Klahr (1989) [ad.] Dunbar (1993)
Rapport (Multiple-choice/-select)	[gr.5-6] Baxter und Shavelson (1994)	[Kl.5] Mannel et al. (2010) [Kl.5-6] Hammann et al. (2007) [gr.5-6] Hammann et al. (2008) [gr.7-9] Tannenbaum (1971) [gr.7-10] Fraser (1980) [Kl.7,12] Klos et al. (2008) [ust.]e Chinn und Brewer (1998)	
Rapport (Kurz- und Langsatz-antworten)	[Kl.2,6,9] HarmoS (2008) [gr.4-6] Chinn und Malhotra (2002) [gr.4,8] TIMSS (1997) [gr.5] Baxter et al. (1992) [gr.5] Solano-Flores et al. (1999) [gr.5] Hammann et al. (2008) [gr.5-6] Baxter und Shavelson (1994); Shavelson et al. (1993) [gr.5-6,13] Rosenquist et al. (2000) [gr.5,8] Siegler und Liebert (1975) [J.11,13,15] APU (1988a, 1988b, 1989)* [Kl.6,9] Labudde, Metzger und Gut (2009); Gut und Labudde (2010) [J.13] Toh und Woolnough (1990); Woolnough und Toh (1990) [Kl.10] Schreiber et al. (2009) [st.] Ruiz-Primo et al. (1993)	[gr.5] Solano-Flores et al. (1999) [Kl.5-10] Grube et al. (2007); Mayer et al. (2008); Möller et al. (2007) [gr.6] Zimmerman und Glaser (2001) [gr.6] Hammann et al. (2008) [Kl.10] Schreiber et al. (2009)	[Kl.10 Schreiber et al. (2009)]
Rapport gemischte / indef. Ant-wortformate		[Kl.9] Wellnitz und Mayer (2008); Wellnitz et al. (2010) [Sek.I] Walpuski (2010) (ind.)	
Interview	[gr.1,3,5] Samarapungavan (1992)** [gr.4-5] Baxter et al. (1995) [gr.4,8] Bass et al. (2002)	[gr.2,4,6/ust.] Tschirgi (1980) [gr.4-5] D. Kuhn und Phelps (1982) [gr.5-6] Schauble, Klopfer und Raghavan (1991) [gr.6,9/ust.] Koslowski (1996)* [Kl.8] Ehmer und Hammann (2007)	[gr.5-6] Schauble (1990)

Tabelle 3.18 – Überblick über ausgewählte Literatur zur Messung experimenteller Kompetenz geordnet nach Testart und Messart (individuelle Messung oder * = auch Gruppenassessment, ** = nur Gruppenassessment; Population: gr. = grade, Kl. = Klassenstufe, ust. = undergraduate students, gst. = graduate students, st. = students, ad. = adults, J. = Alter in Jahren)

B. Ergebnisse zu Messarten bei Experimentiertests

Die direkte Beobachtung von Hands-on-Tests gilt in der Literatur als Benchmark alternativer Messmethoden. Der Einsatz dieses Instruments bedingt jedoch grosse personelle und logistische und letztlich finanzielle Ressourcen, weshalb für large-scale Assessments alternative Messinstrumente evaluiert werden. Um den logistischen Aufwand zu senken, werden Simulationstests und Papier-und-Bleistift-Tests als alternative Testarten eingesetzt. Um den personellen Bedarf an Ratern zu reduzieren, werden Eigenrapporte anstelle von Fremdrapporte bevorzugt. Jedoch kann der Gebrauch gewisser Rapportformate die Fragestellung einer Aufgabe verändern. Dies hängt u. a. vom Grad der *inhaltlichen Offenheit* und *formalen Offenheit* sowie vom Grad der *Strukturiertheit* eines Rapportformats ab (cf. Abb. 7.2.2). Die Skalen der inhaltlichen und formalen Offenheit beziehen sich auf die Frage, inwieweit das Rapportformat den Inhalt und die Darstellungsform (e. g. Text, Tabelle, Zeichnung) vorbestimmt. Die Strukturiertheit betrifft den Grad, mit dem das Antwortformat die Auswahl und die Reihenfolge der zu rapportierenden Aspekte des experimentellen Problemlösens vorgibt. Dies betrifft z. B. spezifische Vorgaben von Elementen, die eine Antwort enthalten sollten.[43]
Gerade hinsichtlich der Eigenrapportierung und dem Ersatz von Hands-on-Aktivitäten durch Hands-off-Probleme stellt sich die Frage nach der Validität der Messungen: Wird mit den verschiedenen Test- und Messarten auch tatsächlich dasselbe gemessen? Nachfolgend eine Zusammenfassung von Antworten empirischer Untersuchungen.

iii. *Beobachtung versus Eigenrapport*: Der Eigenrapport bietet eine reliable Alternative für direkte Beobachtungen von Experimentiertests, wenn ein offenes, partiell strukturiertes Rapportformat gewählt wird. Inhaltlich offene, partiell strukturierte Rapportformat erweisen sich gegenüber minimal und maximal strukturierten Formaten als die Variante mit der höchsten Reliabilität (Toh & Woolnough, 1990).[44] Die Verwendung von vorstrukturierten Laborjournals (*notebook*) ergeben im Vergleich mit direkten Beobachtungen auch hohe Score-Übereinstimmungen ($.75 \leq r \leq .86$, Baxter et al., 1992, 12; Baxter & Shavelson, 1994, 293, Shavelson et al., 1993, 229). Dies korrespondiert mit Ergebnissen der APU-Staffeln (APU, 1985, 212).

[43] Zwei Extrembeispiele zur Offenheit und Strukturiertheit von Antwortformaten: Ein Multiple-choice-Format ist demnach inhaltlich und formal geschlossen und maximal strukturiert. Ein leeres Blatt mit der simplen Aufforderung, den Lösungsweg aufzuschreiben, ist inhaltlich und formal offen und minimal strukturiert.

[44] Als offen, partiell strukturiertes Rapportformat wird eine Serie von offenen Fragen, die auf möglichst viele Aspekte der jeweiligen Experimentieraufgabe (e. g. Vorversuche, Planung des Hauptversuchs, Durchführung, Protokoll, Ergebnisinterpretation, Reflexion und Versuchsmodifikation) ausgerichtet sind (Woolnough & Toh, 1990, 129).

iv. *Performance-Instabilität*: Die Performance von Schülerinnen hängt von der Aufgabe (*task*), von der Test- und Messart (*method*) und vom Zeitpunkt der Testdurchführung (*occasion*) ab (Baxter & Shavelson, 1994; Miller & Linn, 2000; Ruiz-Primo et al., 1993; Ruiz-Primo & Shavelson, 1996; Shavelson et al., 1991, 1992, 1993). Die $person \times task$-Varianz bedingt eine minimale Anzahl von Testaufgaben, um bei der Kompetenzmessung eine akzeptable Reliabilität zu gewährleisten (Gao et al., 1994; Miller & Linn, 2000; Shavelson et al., 1993).[45] Die $person \times method$-Varianz lässt auf die Abwesenheit von transferfähigem Wissen schliessen (Rosenquist et al., 2000, 15ff). Die $person \times occasion$-Varianz bewirkt, dass gleichzeitige Messungen auf verschiedene Arten (e. g. Beobachtung und Eigenrapport beim Hands-on-Test) signifikant höher korrelieren als nicht gleichzeitig stattfindende Messungen, wie dies bei verschiedenen Testarten, e. g. Simulationen und Hands-on-Tests, prinzipiell der Fall ist (Shavelson et al., 1999). Die individuelle Performance bei experimentellen Kompetenzmessungen erweist zudem sich als "volatil" und instabil (Shavelson et al., 1999, 16).

[45] Um bei Experimentiertests eine Reliabilität von .80 sicherzustellen, werden je nach Studie zwischen 8 bis 23 Experimentieraufgaben, um einen Wert von .70 zu erreichen, 2 bis 10 Aufgaben benötigt (Miller & Linn, 2000, 371; Ruiz-Primo & Shavelson, 1996, 1050f).

Kapitel 4

Die Itemschwierigkeit

4.1 Interdependenz von Itemschwierigkeit und Kompetenzausprägung

Ordnungsprinzipien. Eine Aufgabe wird umso leichter eingeschätzt, je besser sie gelöst wird und je weniger kompetent die Personen sind, die sie lösen. Im vorangehenden Kapitel haben wir festgehalten, dass eine Person umso kompetenter ist, je besser sie eine Aufgabe löst und je schwieriger die Aufgabe ist, die sie löst (vgl.die Kompetenzdimensionen auf S. 16). Will man nun die Kompetenzen von Testpersonen messen und setzt die obigen plausiblen Aussagen bei der Kompetenzmessung um, so gerät man in den Zirkel, dass die Kompetenzausprägungen bereits bekannt sein müssen, bevor sie gemessen werden können. Um die Kompetenzausprägung einer Person zu messen, benötigt man nämlich Aufgaben, deren Itemschwierigkeit man bereits kennt. Um die Itemschwierigkeit dieser Aufgaben zu messen, benötigt man jedoch eine Stichprobe von Personen, deren Kompetenzausprägungen bereits bekannt sind. Auch wenn es sich bei diesem Schluss nicht um eine logische Unmöglichkeit handelt – man kann immer noch versuchen, beides, die Kompetenzausprägungen und die Itemschwierigkeiten gleichzeitig zu messen – so wird durch diesen Zirkel deutlich, dass zwischen der Messung von Kompetenzen und der Bestimmung von Itemschwierigkeiten eine wechselseitige Abhängigkeit besteht. Um den beschriebenen Zirkel zu umgehen, müssen den Messungen bestimmte Annahmen zugrunde gelegt werden, die wir im Folgenden als *Ordnungsprinzipien* bezeichnen. Wo diese Ordnungsprinzipien ins Spiel kommen, soll an einem Beispiel veranschaulicht werden.

Um die Schwierigkeit von Aufgaben zu bestimmen, unterzieht man die Aufgaben einem Test mit einer möglichst grossen Schülerstichprobe. Diejenigen Aufgaben, die besser gelöst werden, erweisen sich dadurch als die leichteren Aufgaben. Diejenigen, die schlechter gelöst werden, als die anspruchsvolleren. Nehmen wir z. B. an, dass bei einem vollständigen Test n_I Items mit einer Stichprobe von n_P Testpersonen untersucht wurden und dabei

für jede Testperson i beim Item k die Punktzahl x_{ik} resultierte. Die Itemschwierigkeit σ_k des k-ten Items lässt sich dann als simples negatives Itemscore berechnen (Rost, 2004a, 91f).

$$\sigma_k = -\sum_{i=1}^{n_P} x_{ik} \tag{4.1}$$

Nehmen wir zudem an, dass die Items jeweils nur mit einer Teilstichprobe getestet werden, wie dies bei large-scale Assessments üblich ist. Dann hängt das Itemscore auch davon ab, wie kompetent die Teilstichprobe ist, mit der ein Item getestet wird. Eine weniger kompetente Teilstichprobe führt zu einem tieferen Itemscore und somit zu einer höheren Einschätzung der Itemschwierigkeit. Um in diesem Fall die "wahre" Itemschwierigkeit zu bestimmen, müssen die Kompetenzausprägungen der einzelnen Testpersonen bekannt sein. Nehmen wir nun an, dieses Wissen sei vorhanden und die Fähigkeit der i-ten Person sei durch θ_i gegeben, dann könnte man die Itemschwierigkeit des k-ten Items mit dem negativen, mit der Fähigkeit gewichteten Itemscore gleichsetzen.

$$\sigma_k = -\sum_{i=1}^{n_P} x_{ik}\theta_i \tag{4.2}$$

Die Fähigkeiten der Testpersonen sind nun aber meistens ebenso wenig bekannt wie die Schwierigkeit der Items. Deshalb müssen die Personenfähigkeiten mit denselben Daten berechnet werden. Hierzu bietet sich das mit der Aufgabenschwierigkeit gewichtete Personenscore an.

$$\theta_i = \sum_{j=1}^{n_I} x_{ij}\sigma_j \tag{4.3}$$

Setzt man die beiden Definitionsgleichungen 4.2 und 4.3 zusammen, erhält man ein System von $n_P + n_I$ homogenen Gleichungen.

$$\sigma_k = -\sum_{i=1}^{n_P} x_{ik}\theta_i = -\sum_{i=1}^{n_P}\sum_{j=1}^{n_I} x_{ik}x_{ij}\sigma_j \tag{4.4}$$

$$\theta i = \sum_{j=1}^{n_I} x_{ij}\theta_i = -\sum_{j=1}^{n_I}\sum_{k=1}^{n_P} x_{ij}x_{kj}\theta_k \tag{4.5}$$

Seien $\hat{\sigma} = (\sigma_1, ...\sigma_{n_I})$ und $\hat{\theta} = (\theta_1, ...\theta_{n_P})$ Vektoren im n_I-dimensionalen Itemschwierigkeitsraum und n_P-dimensionalen Personenfähigkeitsraum und X die $n_P \times n_I$-Scorematrix, dann können die Systeme 4.4 und 4.5 als Eigenvektorprobleme von Matrizen dargestellt werden.

$$\hat{\sigma} = -X^T X\hat{\sigma}$$
$$\hat{\theta} = -XX^T\hat{\theta}$$

Der gewählte lineare Ansatz ist problematisch, denn das Gleichungssystem besitzt nebst der trivialen Lösung $\hat{\sigma} = \hat{0}$ und $\hat{\theta} = \hat{0}$ nur in speziellen Fällen eine eindeutige nichttriviale Lösung. Dies kann nur dann der Fall sein, wenn die Ränge der Matrizen der Dimensionen gleich der dazugehörigen Vektorräume sind, i. e. $\text{Rang}(X^T X) = n_I$ und $\text{Rang}(XX^T) = n_P$. Ansonsten gibt es unendlich viele Lösungen – was dann der Fall ist, wenn die Messwerte aufgrund von Messfehlern oder Messschwankungen die Ordnungsprinzipien lokal verletzen. Der lineare Ansatz (Gl. 4.2 und 4.3) erweist sich als nicht stabil und bietet deshalb keine Lösung für den Zirkel der Interdependenz von Kompetenz und Itemschwierigkeit. Dies gilt grundsätzlich für alle deterministischen Ansätze, die Messfehler nicht modellieren.

Wie wir an diesem Beispiel gesehen haben, basiert eine Messung auf Annahmen, nach welchen *Ordnungsprinzipien* die Kompetenzausprägungen und die Aufgabenanforderungen progredieren sollen. Es werden also Progressionsdimensionen für die Kompetenz und die Itemschwierigkeit und deren Zusammenhang a priori festgelegt. In einer klassischen Testtheorie wird die Interdependenz von Itemschwierigkeit und Kompetenzausprägung durch die gewählten Ordnungsprinzipien aufgehoben. Eine Testperson wird umso kompetenter eingeschätzt, je besser sie die Items löst, unabhängig davon, wie schwierig die Items tatsächlich sind. Analog gelten Items einfach dann als schwierig, wenn sie schlechter gelöst werden (cf. Definition 4.1). Die für gewöhnliche Leistungstests an Schulen und Universitäten angebrachten "klassischen" Ordnungsprinzipien bedürfen für large-scale Assessments einer Erweiterung. Diese haben wir bereits in Form von zwei Kompetenzprogressionen (KP) hergeleitet (cf. Abs. 3.1.1, S. 16).

(KP1) Eine Testperson ist umso kompetenter, je schwieriger die Items sind, die sie in einem Test löst.

(KP2) Eine Testperson ist umso kompetenter, je besser sie die Items eines Tests löst.

Die Ordnungsprinzipien für Kompetenzausprägungen werden durch analoge Prinzipien für die Itemschwierigkeitsprogression (IP) ergänzt.

(IP1) Ein Item ist umso leichter, je weniger kompetent die Testpersonen sind, die sie in einem Test lösen.

(IP2) Ein Item ist umso leichter, je besser es in einem Test gelöst wird.

Die vier Ordnungsprinzipien für large-scale Assessments bilden inhaltlich zusammenhängende Paare. In (KP1) und (IP1) ist die oben diskutierte Interdependenz der Messung von Kompetenzausprägungen (θ_i) und der Bestimmung von Itemschwierigkeiten (σ_j) enthalten. (KP2) und (IP2) stellen den Bezug zu den im Test erzielten Scores (x_{ij}) her. De-

terministische Ansätz, wie der oben diskutierte lineare Ansatz (4.2, 4.3), die den vier Ordnungsprinzipien gerecht werden, scheitern an lokalen Verletzungen der Prinzipien durch die Messdaten. Large-scale Assessments werden daher mit probabilistischen Modellen ausgewertet (e. g. Rasch-Modellen), bei denen lokale Ordnungsverletzungen in der Datenmatrix bloss die Wahrscheinlichkeit des Modells senken, nicht aber das Modell unmöglich machen (cf. Rost, 2004a; Wu & Adams, 2007).

Da wir uns für die Auswertung des HarmoS-Experimentiertests ⟨E08|69df⟩ auf das ordinale Rasch-Modell (Partial credit model) stützen werden, wollen wird kurz darauf eingehen, wie das Partial-credit-Modell die vier Ordnungsprinzipien umsetzt (Masters & Wright, 1997). Rasch-Modelle nehmen die Interdependenz (KP1 und IP1) mit der Modellannahme auf, dass die Wahrscheinlichkeit p_{ij}, mit der eine Testperson i ein Item j korrekt löst[1], durch eine allgemeingültige Funktion (*Itemfunktion*) in Abhängigkeit der Differenz von Kompetenzausprägung θ_i und Itemschwierigkeit σ_j gegeben ist. Bei einem dichotomen Test ($x_{ij} \in \{0, 1\}$) wird dieser allgemeingültige A priori-Zusammenhang durch eine logistische Funktion angesetzt.

$$p(x_{ij}) = f(\theta_i - \sigma_j) = \frac{exp(x_{ij}(\theta_i - \sigma_j))}{1 + exp(\theta_i - \sigma j)}$$

Das Produkt aller Itemfunktionen ergibt dann die Wahrscheinlichkeit der ganzen Datenmatrix $X = (x_{ij})_{i \in P, j \in I}$.

$$L = f(\theta_1, ...\theta_{n_P}, \sigma_1, ...\sigma_{n_I}) = \prod_{i=1}^{n_P} \prod_{j=1}^{n_I} p(x_{ij}) = \prod_{i=1}^{n_P} \prod_{j=1}^{n_I} \frac{exp(x_{ij}(\theta_i - \sigma_j))}{1 + exp(\theta_i - \sigma j)}$$

Diese Wahrscheinlichkeit ist eine Funktion (*Likelihoodfunktion*) aller Personen- und Itemparameter, die bei gegebener Datenmatrix X derart gewählt bzw. geschätzt werden können, dass die Datenwahrscheinlichkeit L maximal wird. Lokale Verletzungen der Ordnungsprinzipien in der Datenmatrix haben nun nicht zur Folge, dass dieses Schätzverfahren keine Lösung liefert. Solange eine ausreichend grosse Datenmenge zur Verfügung steht, gibt es immer eine Lösung. Lokale Ordnungsverletzungen führen lediglich zu einer kleineren maximalen Wahrscheinlichkeit der Daten.

Die Itemfunktion genügt auch den Ordnungsprinzipien im einzelnen. Dies lässt sich leicht mit Hilfe der Monotonie der Itemfunktion zeigen. Die Lösungswahrscheinlichkeit ist umso grösser, je grösser die Differenz $(\theta_i - \sigma_j)$ ist.[2] Bei polytomen Codes wird dieses

[1] Die Wahrscheinlichkeit p_{ij} ist eine Propensität und beschreibt eine Leistungsdisposition. Sie kann nicht mit einer realisierbaren Häufigkeit gleichgesetzt werden, sondern entspricht einer imaginären Häufigkeit, mit der die Testperson i das Item j korrekt löste, würde sie bei unverändertem Vorwissen und gleichbleibenden Vorkenntnissen unendlich oft vor die Aufgabe gestellt, das Item zu lösen (Rost, 2004a, S.155ff).

[2] Bei gegebener Itemschwierigkeit ist somit die Kompetenzausprägung einer Testperson umso grösser, je höher die Lösungswahrscheinlichkeit ist (KP2). Umgekehrt gilt bei gegebenem Personenparameter, dass

Modell verallgemeinert, indem jeder Kategorie (*Codes*) $k = 1, ...m_j$ eines Items j eine separate *Kategorienfunktion* mit separatem Schwellenparameter τ_{jk} zugeordnet wird. Die Wahrscheinlichkeit, dass die Testperson i beim Item j den Code x erreicht, ist dann durch die Kategorienfunktion

$$p_{xij} = f_x(\theta_i, \tau_{j1}, ...\tau_{jm_j}) = \frac{exp \sum_{k=0}^{x}(\theta_i - \tau_{jk})}{\sum_{h=0}^{m_j} exp \sum_{k=0}^{h}(\theta_i - \tau_{jk})}$$

gegeben, wobei die Summe über alle Kategorienfunktionen 1 ergeben muss (Masters & Wright, 1997, 103).

$$\sum_{x=1}^{m_j} p_{xij} = 1$$

Ein Schwellenparameter τ_{jk} gibt an, ab welcher Kompetenzausprägung θ_i einer Testperson i der Code k beim Item j wahrscheinlicher ist als der Code $(k-1)$. Die Itemschwierigkeit wird nun als Mittelwert aller Schwellenwerte berechnet. Für alle Items j gilt somit

$$\sigma j = \frac{1}{m_j} \sum_{k=1}^{m_j} \tau_{jk}.$$

Erwartet wird, dass die Schwellenparameter die normale Ordnung der Codes wiedergeben: $\tau_{j1} < \tau_{j2} < \tau_{j3} < ... < \tau_{jm_j}$. In diesem Fall gibt es auf dem Kontinuum der Kompetenzausprägungen für jeden Code einen Bereich, in dem der Code am wahrscheinlichsten auftritt. Sind die Schwierigkeiten benachbarter Codes jedoch nahe beieinander, kann der Fall eintreten, dass ein Code für keine Kompetenzausprägung das wahrscheinlichste Testresultat ist. Die Kategorienfuktion dieses Codes wird überall durch die Kategorienfunktionen der benachbarten Codes übertroffen. Dies zeigt sich daran, dass die Schwellenparamater die angenommene Ordnung verletzen. Im Beispiel $\tau_{j2} < \tau_{j1} < \tau_{j3} < ... < \tau_{jm_j}$ ist der Code 0 bis zur Schwelle τ_{j1} wahrscheinlicher als der Code 1. Bereits ab der tiefer liegenden Schwelle τ_{j2} ist aber der Code 2 wahrscheinlicher als der Code 1. Der Code 1 ist daher nie der wahrscheinlichste Code. Teilt man nun das Kontinuum der Kompetenzausprägungen θ in Bereiche ein, wo jeweils ein Code der wahrscheinlichste ist, so wird der Code 1 auf dem Kontinuum nicht abgebildet. Rost (2004a, 205) erkennt in der Ordnung der Schwellenparameter daher ein Kriterium, mit welchem "nachgeprüft werden kann, *ob* die Kategorien *überhaupt geordnet* sind, d. h. ob die Itemantworten Ordinalskalenqualität

---

die Itemschwierigkeit umso kleiner sein muss, je höher die Lösungswahrscheinlichkeit ist (IP2). Erzielt eine Person j bei einem Item i ein hohes Score ($x_{ij} = 1$), so ist ihre Lösungswahrscheinlichkeit $p(x_{ij} = 1)$ hoch, was wiederum eine grosse Differenz $(\theta_i - \sigma_j)$ bedeutet. Für die Bestimmung des Personenparameters heisst das: Je schwieriger die Aufgabe ist, je grösser muss die Kompetenzausprägung sein (KP1). Für die Bestimmung der Itemschwierigkeit heisst das: Je weniger kompetent die Testperson ist, desto leichter muss die Aufgabe sein (IP1).

haben.“ Masters und Wright (1997, 105f) betonen demgegenüber, dass Kompetenzbereiche, in denen jeweils ein Code der wahrscheinlichste ist, nur eine von verschiedenen Möglichkeiten darstellen, die Antwortkategorien (Codes) auf dem Kontinuum der Personenparameter abzubilden. Eine alternative Abbildung, die auch für die Überprüfung von Standards sinnvoll erscheint, teilt das Kontinuum der Personenparameter in Bereiche ein, in denen ein höherer Code wahrscheinlicher ist als ein tieferer Code (Masters & Wright, 1997). Diese Interpretation der Schwellenparameter ist invariant gegenüber einer Vertauschung der normalen Ordnung der Schwellenparameter. Insofern hängt die Beurteilung der Ordinalskalenqualität von der jeweiligen Interpretation ab.

4.2 Modellierung von Itemschwierigkeit

Die Frage der Validität eines Experimentiertests hängt auch mit der Frage zusammen, was eine Testaufgabe leicht bzw. schwierig macht und welche bzgl. der experimentellen Kompetenz relevanten und -irrelevanten Fähigkeiten dabei eine Rolle spielen (Messick, 1994; Miller & Linn, 2000). In Bezug auf den Einfluss von *kompetenzirrelevanten Anforderungen*[3] auf die Itemschwierigkeit erweisen sich Experimentieraufgaben als besonders komplex.

> The ancillary skill requirements of performance assessments are likely to exceed those of more conventional tests, although, as with any test, they may be minimized through careful test design. The materials and instructions provided are more complex and varied and the required modes of responding are more demanding. Perhaps the most obvious threat to the validity of most performance assessments is their dependence on reading and writing (Haertel & Linn, 1996, 63).

Um die Einflüsse von kompetenzirrelevanten Anforderungen einer Experimentieraufgabe zu modellieren, schlagen wir das in der Abbildung 4.1 dargestellte, idealisierte Aufgabenmodell vor, gemäss welchem die Bearbeitung der Aufgabe in drei Schritten und die Festlegung der Itemschwierigkeit in vier Schritten erfolgt (für die Darstellung des Modells vgl. auch Gut, Labudde & Ramseier, 2010). Der Problemlöseprozess wird gemäss dem Schema in drei Arbeitsschritte zerlegt: Zuerst muss die Aufgabe erfasst (Arbeitsschritt ‹Aufgabe erfassen›), dann das Problem gelöst (Arbeitsschritt ‹Problem lösen›)

[3]Mit kompetenzirrelevanten Anforderungen bezeichnen wir im Folgenden Schwierigkeiten der Aufgabenstellung, wie z. B. grammatikalisch anspruchsvolle Texte, die sich nicht direkt auf die experimentelle Kompetenz beziehen. Dementsprechend betreffen kompetenzrelevante Schwierigkeiten direkt die experimentelle Kompetenz.

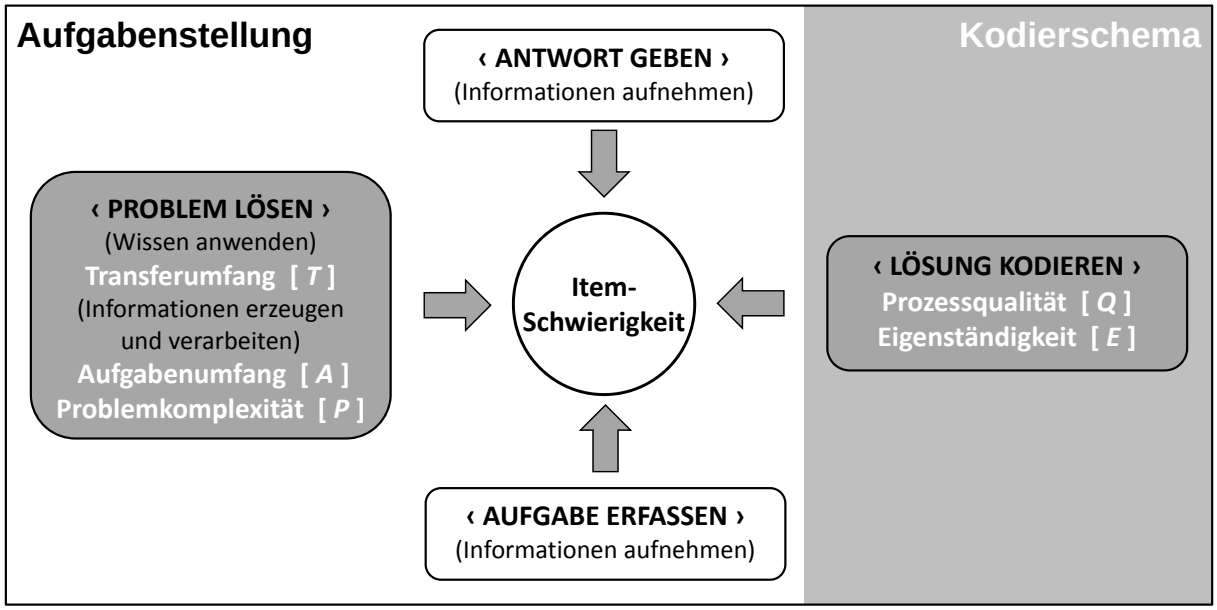


Abbildung 4.1 – Arbeitsschrittmodell zur Beschreibung und Erklärung der Itemschwierigkeit: Die Erfassung der Arbeitsschritte basiert – in Korrespondenz mit den Kompetenzprogressionen (KP1) und (KP2), cf. S.16 – auf der Analyse einerseits der Aufgabenstellung und andererseits des Kodierschemas.

und schliesslich die Antwort gegeben (Arbeitsschritt ‹Antwort geben›) werden. Diese Arbeitsschritte werden von den Testpersonen in der Regel nacheinander "durchgearbeitet", wobei jeweils die erfolgreiche Bearbeitung eines Arbeitsschritts die Bewältigung der vorangehenden Arbeitsschritte bedingt. Bei Experimentiertests, die nicht via Beobachtung, sondern via Eigenrapport geratet werden, stellen das ‹Aufgabe erfassen› und das ‹Antwort geben› zusätzliche kompetenzirrelevante Schwierigkeiten dar, welche die Kompetenzmessung verfälschen. In diesen Fällen wird man bei der Itemkonstruktion versuchen, die nicht kompetenzrelevanten Schwierigkeiten zu minimieren oder zumindest über alle Testaufgaben hinweg zu egalisieren.[4] Es gibt jedoch auch Kompetenzmessungen, bei denen das ‹Aufgabe erfassen› und/oder das ‹Antwort geben› zur Kompetenz dazugehören, weil die Kompetenzbeschreibung diese zusätzlichen Arbeitsschritte umfasst[5] oder weil der

[4]Bei der Entwicklung der Experimentieraufgaben für den HarmoS-Experimentiertest wurden nicht kompetenzspezifische Schwierigkeiten der Arbeitsschritte ‹Aufgabe erfassen› und ‹Antwort geben› mittels globaler und individueller Massnahmen zu minimieren versucht. In der Pilotierungsphase wurden zu schwierige Items vereinfacht, indem Informationen einfacher dargestellt wurden oder einfachere Antwortformate gewählt wurden. Bei den Schülerinnen des 6. Jahrgangs wurde der ganze Test vereinfacht, indem die Aufgabenstellung durch die Testaufsicht laut vorgelesen wurde (cf. Abschnitt 7.1).

[5]Experimentelle Standardbeschreibungen enthalten oft explizit den Arbeitsschritt, Ergebnisse einer Untersuchung korrekt und adäquat zu beschreiben und darzustellen. Es wird jedoch nicht immer klar, ob damit implizit auch der Arbeitsschritt ‹Antwort geben› mitgemeint ist. Hier ist eine Schärfung des

Kern der zu messenden Kompetenz gerade im Erfassen bzw. Beantworten der Aufgabe liegt[6]. Notwendig ist dann die kontrollierte Variation der Anforderungen, die von diesen Arbeitsschritten auszugehen vermutet werden. In beiden Fällen ist es jedoch von wissenschaftlichem Interesse, die Varianzanteile des Arbeitsschritts ‹Aufgabe erfassen› und ‹Antwort geben› an der Gesamtvarianz zu bestimmen.

Während die Lösung einer Aufgabe in drei Arbeitsschritten erfolgt, macht die Bestimmung deren Itemschwierigkeit einen vierten Arbeitsschritt notwendig. Nebst den schwierigkeitsrelevanten Anforderungen, die durch die Arbeitsschritte ‹Aufgabe erfassen›, ‹Problem lösen› und ‹Antwort geben› an die Testperson gestellt werden, erfolgt durch das abschliessende ‹Lösung kodieren› die Bewertung der Qualität der Lösung und des Lösungswegs – bei beobachteten Experimentiertests würde man mutatis mutandis an dieser Stelle die ‹Prozesse kodieren›. Unter der Annahme, dass das ‹Aufgaben erfassen› und das ‹Antwort geben› nicht Teil der experimentellen Kompetenz ist, die man messen möchte, erfordert die Modellierung von kompetenzrelevanten Itemschwierigkeiten die Analyse von schwierigkeitserzeugenden Merkmalen der Arbeitsschritte ‹Problem lösen› und ‹Lösung kodieren›. Die Analyse der beiden Arbeitsschritte entspricht der Untersuchung der auf der S. 16 diskutierten Progressionsebenen der Aufgabenstellung (KP1) und der Aufgabenlösung (KP2). Die zu untersuchenden Merkmale des Arbeitsschritts ‹Problem lösen› betreffen somit in Korrespondenz zu unserem "fünfdimensionalen" Progressionsmodell (cf. S. 16) den Aufgabenumfang $[A]$, die Problemkomplexität $[P]$ und den Transferumfang $[T]$. Dementsprechend korrespondieren die schwierigkeitserzeugenden Merkmale des Arbeitsschritts ‹Lösung kodieren› mit den Progressionsdimensionen Prozessqualität $[Q]$ und Eigenständigkeit $[E]$, sofern unter dem Kodieren der Antwort auch die Kodierung der während des Tests geleisteten Assistenz verstanden wird.

Somit ist die Möglichkeit angesprochen, Anforderungsmerkmale der Kompetenzbeschreibung in schwierigkeitserzeugende Merkmale des Messinstruments zu übersetzen. Hierbei werden Anforderungsmerkmale der fünf Progressionsdimensionen $[A,E,P,Q,T]$ in Itemmerkmale der zwei Arbeitsschritte ‹Problem lösen› und ‹Lösung kodieren› übertragen. Die adäquate Übersetzung – wie sie in der Abbildung 4.2 schematisch dargestellt ist – bildet eine notwendige Bedingung für die valide Kompetenzmessung. Die Übersetzung entpuppt sich jedoch als nicht trivial: Dies aus zwei Gründen (*Übersetzungsproblemen*).

ÜP1 *Modellierung der Schwierigkeit auf unterschiedlichen Ebenen des Messinstruments*: Die Progressionsdimensionen sprechen verschiedene Ebenen des Messinstrumen-

Standraddiskurses angebracht.

[6]Dies gilt z. B. für die Kompetenz, gegebene Informationen zu erfassen und in einer vorgegebenen Repräsentationsform wiedergeben zu können. Man vergleiche hierzu die Kompetenzbeschreibungen des Handlungsaspekts «Informationen erschliessen» in EDK (2011, 42f).

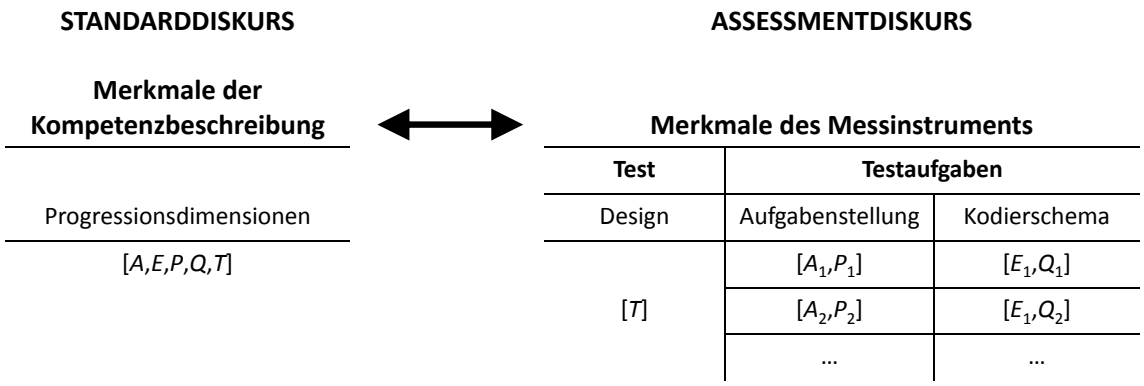


Abbildung 4.2 – Übersetzungsmodell zwischen Anforderungsmerkmalen der Kompetenzbeschreibung und schwierigkeitserzeugenden Merkmalen von Messinstrumenten

tes an. Während der Aufgabenumfang $[A]$, die Problemkomplexität $[P]$, die Prozessqualität $[Q]$ und die Eigenständigkeit $[E]$ Merkmale der einzelnen Testaufgaben (Aufgabenstellung und Kodierschema) sind, ist der Transferumfang $[T]$ ein Merkmal des Tests, sprich ein Merkmal der Gesamtheit aller Testaufgaben. Erst die Gesamtheit aller Aufgaben bestimmt den Umfang der Kontexte, in denen die experimentelle Kompetenz getestet wird. Der Aufgabenumfang und die Problemkomplexität betreffen zudem die Aufgabenstellung, während die Prozessqualität und die Eigenständigkeit sich auf das Kodierschema beziehen (vgl. die Spalten der rechten Tabelle in Abb. 4.2).

ÜP2 *Kompetenzzerlegung in Teilaufgaben*: Aufgrund der Zerlegung einer Kompetenz in Teilaufgaben (S. 29) ergeben sich global über alle Testaufgaben $i = 1, ...n$ variierende Werte für den Aufgabenumfang $[A_i]$ und die Problemkomplexität $[P_i]$ (vgl. die Zeilen der rechten Tabelle in Abb. 4.2). Demgegenüber wird es für den Transferumfang $[T]$ global keine Variation geben, während die Variation der Prozessqualität $[Q]$ und Eigenständigkeit $[E]$ auf einer rein individuellen Ebene stattfindet.

Aufgrund der oben erfolgten Analyse erfordert eine valide Kompetenzmessung einerseits die adäquate Übersetzung der Anforderungsmerkmale der Kompetenzbeschreibung in korrespondierende Schwierigkeitsmerkmale der Testaufgaben und andererseits die Modellierung von kompetenzirrelevanten Schwierigkeiten, die von den Arbeitsschritten ‹Aufgabe erfassen› und ‹Antwort geben› ausgehen. Für die weitere Diskussion von Itemschwierigkeiten folgt aus dem Übersetzungsproblem die Einschränkung des Fokus auf Merkmale des Aufgabenumfangs, der Problemkomplexität, der Prozessqualität und der Eigenständigkeit. Nicht weiter verfolgt werden Merkmale des Transferumfangs, die ganze Tests betreffen.

4.2.1 Fallbeispiel ⟨Balkenwaage⟩

Das Zusammenspiel der vier Arbeitsschritte ‹Aufgabe erfassen›, ‹Problem lösen›, ‹Antwort geben› und ‹Lösung kodieren› wollen wir am Beispiel der HarmoS-Testaufgabe ⟨Balkenwaage⟩ exemplarisch diskutieren. Die Besonderheit dieser Testaufgabe ist, dass mittels einer Standardisierung der Aufgabenformate und der Kodierschemen für verschiedene Teilaufgaben die Einflüsse der Arbeitsschritte ‹Aufgabe erfassen›, ‹Antwort geben› und ‹Lösung kodieren› konstant gehalten werden und der Einfluss des Arbeitsschritts ‹Problem lösen› auf die Aufgabenschwierigkeit isoliert untersucht werden kann. Diese Bedingung wurde mit einer Gruppe von vier unabhängigen Teilaufgaben mit jeweils drei Items realisiert. Jede Teilaufgabe enthält den formal gleichen Auftrag, eine gegebene Hypothese zum Gleichgewicht einer Balkenwaage, die mehrfach belastet werden kann, empirisch zu überprüfen. In der ersten Teilaufgabe wurden die Schülerinnen und Schüler aufgefordert, die Behauptung

H_1 Eine symmetrisch belastete Waage befindet sich immer im Gleichgewicht.

in der zweiten Teilaufgabe deren logische Umkehrung

H_2 Wenn eine Waage im Gleichgewicht ist, dann ist sie immer symmetrisch belastet.

zu untersuchen (cf. App. A.1). Zu jeder Hypothese mussten die Schülerinnen und Schüler mit einer Balkenwaage und sechs gleichen Schraubenmuttern (Abb. 4.3) zwei Tests durchführen und protokollieren (Item 1 und 2) und dann die Gültigkeit der Hypothese aufgrund der Resultate einschätzen (Item 3). Mit den ersten beiden Items wurde die Teilkompetenz «Suche im Experimentierraum» gemessen (man vergleiche hierzu das Problemlösemodell auf S. 43). Für die weitere Diskussion werden diese zwei Items zu einem

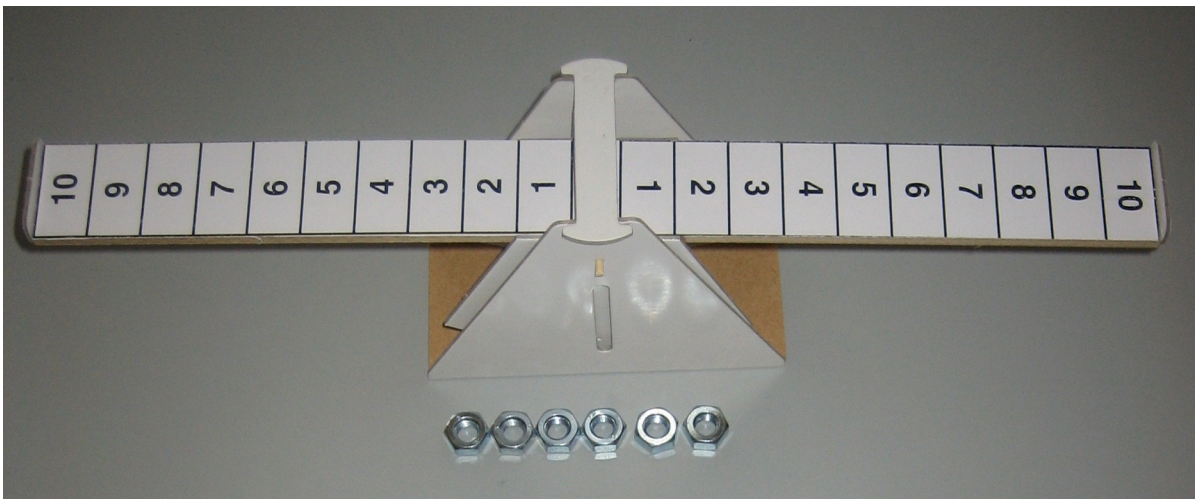


Abbildung 4.3 – Experimentiermaterial der Aufgabe ⟨Balkenwaage⟩

			6. Schuljahr		9. Schuljahr	
Teilkompetenz	**Aufgabe**	**Item**	σ **[logit]**	**error**	σ **[logit]**	**error**
«Suche im Experimentierraum»	H_1	⟨N1E23i01⟩	-1.566	0.108	-1.746	0.108
«Suche im Experimentierraum»	H_1	⟨N1E23i02⟩	-0.431	0.092	-0.603	0.092
«Datenanalyse»	H_1	⟨N1E23i03⟩	-0.566	0.094	-0.844	0.094
«Suche im Experimentierraum»	H_2	⟨N1E23i04⟩	0.629	0.090	-1.063	0.090
«Suche im Experimentierraum»	H_2	⟨N1E23i05⟩	0.673	0.088	-0.379	0.088
«Datenanalyse»	H_2	⟨N1E23i06⟩	0.863	0.086	-0.187	0.086

Tabelle 4.1 – Aufgabe ⟨Balkenwaage⟩: Schätzung der Itemschwierigkeit σ für ausgewählte Items zu den Teilkompetenzen «Suche im Experimentierraum» und «Datenanalyse»

Item «Experimentieren» zusammengezogen. Mit dem dritten Item wurde die Teilkompetenz «Datenanalyse» analog zum Problemlösemodell erfasst.

Die äusseren Strukturen der Aufgaben H_1 und H_2 sind in Bezug auf das ‹Antwort geben› identisch. In Bezug auf das ‹Aufgabe erfassen› bestehen Unterschiede bei der grammatikalischen Formulierung der Hypothesen und bei der Relevanz der im gemeinsamen Aufgabenstamm enthaltenen Zusatzinformationen. Dies betrifft die Experimentierbeispiele im Aufgabenstamm, die gerade Lösungsbeispiel der Aufgabe H_1 sind (cf. App. A.1). Trotz dieser Unterschiede darf davon ausgegangen werden, dass Differenzen in der Itemschwierigkeit zwischen den beiden Aufgaben vor allem auf unterschiedliche Anforderungen der Teilaufgaben ‹Problem lösen› zurückzuführen sind.

Im Test erweist sich die Hypothese H_2 als deutlich schwieriger zu überprüfen als die Hypothese H_1. Eine eindimensionale Rasch-Analyse mit einem reduzierten HarmoS-Experimentiertest ⟨E08| $(6 \wedge 9)$d⟩ mit den 42 Experimentieritems, die gleichzeitig im 6. und 9. Schuljahr eingesetzt wurden, und einer dementsprechend reduzierten Stichprobe von 467 Schülerinnen und Schülern der Deutschschweiz beider Schulstufen zeigt, dass die Items der Aufgabe H_1 in beiden Schulstufen leichter sind als die Items der Aufgabe H_2, wobei die Diskrepanz in der 6. Schulstufe besonders stark ausgeprägt ist (cf. Tab. 4.1). Alle sechs Items sind zudem für Sechstklässlerinnen und -klässler schwieriger zu lösen als für Schulabgängerinnen und -abgänger. Der Item-DIF übersteigt hingegen nur für die Items ⟨N1E23i03-06⟩ die Fehlertoleranz.[7]

Eine qualitative Analyse der Antworten bestätigt dieses Resultat und gibt darüber hinaus Hinweise auf die Ursachen der ungleichen Schwierigkeiten. Die Tabellen 4.2 und

[7] Bei der Interpretation der Ergebnisse ist der suboptimale Fit der sechs Items zu berücksichtigen: $0.86 < \text{wMNSQ} < 1.38$; $\text{T} < 2.8$. Die Fitparameter wMNSQ (Weighted mean square) und T-Werte übersteigen die für einen guten Fit vorgeschlagenen Grenzwerte, i. e. $0.87 < \text{wMNSQ}_{expected} < 1.13$ und $\text{T}_{proposed} < 1.96$ (Wu und Adams (2007), 80ff; Wright & Linacre, 1994; Bond & Fox, 2001, 176ff). Vor allem die Items der Aufgabe H_1 zeigen starken Unter-Fit ($\text{wMNSQ} < 1.27$), i. e. sie diskriminieren nur schwach.

4.3 zeigen die Prozessqualität $[Q]$ der Performance der 155 Deutschschweizer Schülerinnen und Schüler, welche die Balkenwaage-Aufgabe bearbeitet haben, für die Teilkompetenzen «Suche im Experimentierraum» und «Datenanalyse», aufgeschlüsselt nach den Aufgaben H_1 und H_2 den Teilkompetenzen. Aus den Daten ist ersichtlich, dass die Aufgabe H_1 von beiden Schuljahrgängen qualitativ gleich gut gelöst wird. Ein deutlicher und hoch signifikanter Unterschied der Performance gilt es demgegenüber bei den zwei Experimentieritems der Aufgabe H_2 zu verzeichnen.[8] Für die Teilkompetenz «Datenanalyse» zeigt sich hingegen kein signifikanter Unterschied.[9] Die Aufgabe H_2 stellt zudem bezüglich beider Teilkompetenzen «Suche im Experimentierraum» und «Datenanalyse» zumindest in der 6. Schulstufe signifikant höhere Anforderungen als die Aufgabe H_1.[10] Auch die Gesamtperformance differiert deutlich: Während bei der Hypothese H_1 64% ($N = 155$) der Deutschschweizer Schülerinnen und Schüler die Balkenwaage zweimal symmetrisch belastetet und den korrekten Schluss zogen, haben bei der Hypothese H_2 nur 30% ein Gegenbeispiel, i. e. eine asymmetrisch belastetet Waage im Gleichgewicht, gefunden und die Hypothese verworfen.

Die signifikante Asymmetrie bei der Itemschwierigkeit kann nur mit Asymmetrien in den Arbeitsschritten ‹Aufgabe erfassen›, ‹Problem lösen›, ‹Antwort geben› und ‹Lösung kodieren› erklärt werden. Wie bereits erwähnt, besteht beim Arbeitsschritt ‹Aufgabe erfassen› eine Asymmetrie in Bezug auf relevante Informationen im Aufgabenstamm, welche die Anforderungsasymmetrie erklären könnte. Eine mögliche Ursache könnten sprachliche Differenzen sein, die mit den Formulierungen der Hypothesen zusammenhängen. Der Arbeitsschritt ‹Antwort geben› ist absolut symmetrisch und fällt als Ursache weg. Der Arbeitsschritt ‹Lösung kodieren› ist insofern asymmetrisch, als es bei der Aufgabe H_2 mehr Möglichkeiten gibt, einen mittleren Code (Code 1) zu erreichen (cf. App. A.1 auf S.221).

[8] Mann-Whitney-U-Test, $p < .001$; Kolmogorov-Smirnov-Test: $p < .001$

[9] Die Kodierung der Datenanalyse-Items ⟨N1E23i03⟩ und ⟨N1E23i06⟩ berücksichtigt auch die Experimentierstrategien, da nicht nur der korrekte Schluss, sondern auch das korrekte Ergebnis bewertet wird. Es besteht somit eine Itemabhängigkeit zwischen den Experimentier- und den Datenanalyse-Items. Mit einer Rekodierung (Code 1 ⇒ Code 2) der Datenanalyse-Items wurde die Abhängigkeit beseitigt. Während die ursprünglichen Items noch einen hoch signifikanten Performanceunterschied zwischen den Schulstufen beweisen, ist der Unterschied mit den rekodierten Items ⟨N1E23i03R⟩ und ⟨N1E23i06R⟩ nicht mehr signifikant.

[10] Der Unterschied der Performance von Schülerinnen und Schülern der 6. Schulstufe ist sowohl zwischen den Summenitems ⟨N1E23i01⊕02⟩ und ⟨N1E23i04⊕05⟩ als auch zwischen den Items ⟨N1E23i03⟩ und ⟨N1E23i06⟩ hoch signifikant (Wilcoxon-Test: $p < .001$). Die Unterschiede in der 9. Schulstufe sind hingegen nicht signifikant.

[11] Entgegen der Darstellung in Tabelle 4.2 basiert der Mittelwertvergleich auf einer fünfwertigen Skala. Die Basisitems ⟨N1E23i01-05⟩ sind dreiwertig (Code 0, Code 1, Code 2). Die Summierung ergibt fünfwertige Summenitems ⟨N1E23i04⊕05⟩ und ⟨N1E23i04⊕05⟩ (Code 0, Code 1, Code 2, Code 3, Code 4).

	6. Schuljahr		9. Schuljahr	
Prozessqualität [Q] von	H_1	H_2	H_1	H_2
H_1:⟨N1E23i01⊕02⟩ und H_2:⟨N1E23i04⊕05⟩	N (%)	N (%)	N (%)	N (%)
kein Experiment, Experiment nicht adäquat oder widersprüchlich	6 (7%)	16 (19%)	5 (7%)	4 (6%)
nur semi-adäquate oder falsch qualifizierte Experimente	3 (4%)	42 (50%)	2 (3%)	8 (11%)
ein adäquates Experiment	20 (24%)	12 (14%)	16 (23%)	20 (28%)
zwei adäquate Experimente	55 (65%)	14 (17%)	48 (68%)	39 (55%)
	84	84	71	71
Mittelwertvergleich (Wilcoxon-Test)[11]	$p < .001$		ns.	

Tabelle 4.2 – Aufgabe ⟨Balkenwaage⟩: Prozessqualität der Teilkompetenz «Suche im Experimentierraum», Häufigkeiten der Codes für die Items ⟨N1E23i01⊕02⟩ und ⟨N1E23i04⊕05⟩ (*falsch qualifiziertes Experiment*: Angaben zum Experiment und Gleichgewicht passen nicht zusammen; *semi-adäquates Experiment*: Experiment eignet sich nur als Bestätigung der Hypothese; *adäquates Experiment*; Experiment eignet sich als Gegenbeispiel der Hypothese)

Diese Asymmetrie kann jedoch ambivalent beurteilt werden, so dass deren Wirkung auf die Itemschwierigkeit schlecht abzuschätzen ist (cf. S.70f). Das Kodierschema fällt deshalb als Ursache ebenfalls weg. Beim letzten Arbeitsschritt ‹Problem lösen› bestehen auf verschiedenen Ebenen Asymmetrien, welche die Anforderungsunterschiede plausibel begründen können. Im Folgenden gehen wir auf vier Asymmetrien ein.

Asymmetrie der Geltung. Hinsichtlich des Arbeitsschritts ‹Problem lösen› besteht zwischen den beiden Aufgaben eine gewisse Symmetrie. Beide Hypothesen thematisieren denselben physikalischen Kontext und enthalten formallogisch die gleiche Operation, mit der dieselben Eigenschaften einer Balkenwaage – nämlich die Eigenschaft der symmetrischen Belastung S und die Eigenschaft des Gleichgewichts G – verknüpft werden. Grammatikalisch sind die Behauptungen zwar unterschiedlich formuliert. Logisch entsprechen aber beide Allaussagen der Form «Alle A sind B», wobei A und B Elemente der Menge S der symmetrisch belasteten Balkenwaage bzw. der Menge G der sich im Gleichgewicht befindenden Balkenwaagen kennzeichnen. Epistemologisch können beide Behauptungen falsifiziert, jedoch nicht verifiziert werden. Durch das Auffinden von immer mehr Beispielen kann sie jedoch empirisch immer besser bestätigt werden. Ein wesentlicher Unterschied besteht hingegen in der *Geltungsasymmetrie* der zwei Behauptungen: Die Behauptung H_1 ist wahr und kann durch beliebig viele Experiment bestätigt werden. Die Behauptung H_2 ist falsch und kann durch beliebig viele Gegenbeispiele widerlegt werden. Entscheidend ist

	6. Schuljahr		9. Schuljahr	
Prozessqualität [Q] von	H_1	H_2	H_1	H_2
H_1:⟨N1E23i03⟩ bzw. H_2:⟨N1E23i06⟩	N (%)	N (%)	N (%)	N (%)
keine Antwort, nicht korrekter Schluss oder widersprüchliche Antwort	20 (24%)	32 (38%)	11 (15%)	17 (24%)
korrekter Schluss aufgrund nicht adäquater Experimente	6 (7%)	36 (43%)	5 (7%)	13 (18%)
korrekter Schluss aufgrund adäquater Experimente	58 (69%)	16 (19%)	55 (77%)	41 (58%)
	84	84	71	71
Mittelwertvergleich (Wilcoxon-Test)	$p < .001$		ns.	

Tabelle 4.3 – Aufgabe ⟨Balkenwaage⟩: Prozessqualität der Teilkompetenz «Datenanalyse», Häufigkeiten der Codes für die Items ⟨N1E23i03⟩ und ⟨N1E23i06⟩

nun aber der Sachverhalt, dass es auch Beispiele gibt, die H_2 bestätigen! Alle symmetrisch belasteten Balkenwaagen bestätigen die Behauptung $S \supset G$, dass alle Balkenwaage im Gleichgewicht symmetrisch belastet sind. Diese Beispiele sind zwar bezüglich der Hypothese H_2 informativ, eignen sich aber nicht als Gegenbeispiel (cf. Tab. 4.4). Sie werden daher im Weiteren als *semi-adäquate Experimente* bezeichnet. Die Geltungsasymmetrie entwickelt nun bei Testpersonen, die nicht wissen, wie man einen logischen Satz der Form «Alle A sind B» falsifiziert, eine einseitig schwierigkeitserzeugende Wirkung. Die Experimentierstrategie, H_2 zu widerlegen, wird mangels logischen Verständnisses blockiert. Dies im Gegensatz zur Experimentierstrategie, H_1 zu bestätigen. Insofern ist die erfolgreiche Überprüfung der Hypothese H_2 aufgrund ihrer Geltung anspruchsvoller als die Überprüfung der Hypothese H_1.

	symmetrische Belastung: S	**asymmetrische Belastung:** $\neg S$
im Gleichgewicht: G	$S \cap G$ bestätigt H_1 und H_2	$\neg S \cap G$ widerlegt H_2
im Ungleichgewicht: $\neg G$	$S \cap \neg G$ nicht realisierbar	$\neg S \cap \neg G$ irrelevant

Tabelle 4.4 – Handlungsvarianten bei den Aufgaben H_1 und H_2

Asymmetrie der heuristischen Fehlermöglichkeiten versus Asymmetrie der relevanten Experimentiervarianten. Die Matrix der Tabelle 4.4 kann auf zwei Arten interpretiert werden. Aus der Perspektive möglicher Fehlstrategien erscheint die Aufgabe H_2 schwieriger als die Aufgabe H_1. Während die Geltung von H_2 durch mangelhaftes Experimentieren (mittels Experimente vom Typ $S \cap G$) falsch eingeschätzt werden kann und gemäss Kodierschema auch mit einem Abzug von einem Punkt (Code 1 statt Code 2) sanktioniert wird,

ist dies bei H_1 empirisch nicht möglich, solange die Experimente sorgfältig durchgeführt werden. Aus der Perspektive der relevanten Experimentiervarianten erscheint die Aufgabe H_2 hingegen als vorteilhafter. Während nur eine von vier Experimentierstrategien für die Aufgabe H_1 relevante Informationen liefert ($S \cap G$), sind dies bei der Aufgabe H_2 zwei Varianten ($S \cap G$ und $\neg S \cap G$). Somit erhöht sich beim planlosen Experimentieren die Wahrscheinlichkeit, eine höhere Bewertung als Code 0 zu erzielen. Wir stellen also zwei Wirkungen W1 und W2 der Geltungsasymmetrie zwischen H_1 und H_2 fest.

W1: Die Asymmetrie bei den Fehlermöglichkeiten, den maximalen Code 2 nicht zu erreichen, erschwert die Aufgabe H_2.

W2: Die Asymmetrie bei den Strategiemöglichkeiten, einen höheren Code als 0 zu erzielen, erleichtert die Aufgabe H_2.

Um die Frage zu entscheiden, welche Wirkung der Geltungsasymmetrie überwiegt, werden die Daten der 6. Schulstufe aus der Tabelle 4.2 derart rekodiert, dass einmal die Wirkung W1 und einmal die Wirkung W2 unterdrückt wird. Im ersten Fall werden alle Code 1 als Code 2 und im zweiten Fall alle Code 1 als Code 0 gewertet. Der Performancevergleich zeigt trotz unterdrückter Wirkungen weiterhin signifikante Unterschiede zwischen den Aufgaben. Bei der Unterdrückung von W1 fällt zwar, wie zu erwarten war, das Signifikanzniveau die Unterschiede bleiben jedoch immer noch sehr signifikant. Wir ziehen daraus den vorläufigen Schluss, dass die Wirkungen der Geltungsasymmetrie nur eine untergeordnete Rolle bei der Erklärung der Anforderungsunterschiede zwischen den zwei Aufgaben H_1 und H_2 spielen.

Asymmetrie des doxastischen Status. Die Hypothesen H_1 und H_2 unterscheiden sich im doxastischen Status. Während die Hypothese H_1 vertraut und plausibel erscheint[13], wirkt der Inhalt der Hypothese H_2 fremd und konstruiert. Mit einer Symmetrie-Argumentation lässt sich die Gültigkeit von H_1 zudem anschaulich "beweisen". Aus einer reductio ad absurdum folgt, dass eine symmetrische Ursache (die Belastung der Waage) keine asymmetrische Wirkung (das Senken einer Seite der Waage) haben kann. Wäre nämlich das

[12] Entgegen der Darstellung in Tabelle 4.5 basiert der Mittelwertvergleich auf einer dreiwertigen Skala. Durch die Rekodierung resultieren zunächst zweiwertige Items ⟨N1E23i01-05⟩ (Code 0, Code 2). Die Summierung ergibt dann dreiwertige Summenitems ⟨N1E23i04⊕05⟩ und ⟨N1E23i04⊕05⟩ (Code 0, Code 2, Code 4).

[13] Die Hypothese H_1 wurde bereits in der ältesten bekannten Schrift zu den Hebelgesetzen – dem so genannten *Book on the balance*, dessen Autorenschaft Euklid zugeschrieben wird – als Folgesatz aus basalen Axiomen zum Gleichgewicht der Balance abgeleitet (Clagett, 1959, 24ff). Bei Archimedes wird die Hypothese in vereinfachter Form als Korollar in seiner Schrift *Vom Gleichgewicht ebener Figuren* wiederholt (Clagett, 1959; Stein, 1930).

	W1 unterdrückt (Code 1 ⇒ Code 2)		**W2 unterdrückt** (Code 1 ⇒ Code 0)	
Prozessqualität [Q] von	H_1	H_2	H_1	H_2
H_1:⟨N1E23i01⊕02⟩ und H_2:⟨N1E23i04⊕05⟩	N (%)	N (%)	N (%)	N (%)
kein Experiment, Experiment nicht adäquat oder widersprüchlich	6 (7%)	16 (19%)		
nur semi-adäquate oder falsch qualifizierte Experimente			29 (35%)	70 (83%)
ein adäquates Experiment	78 (93%)	68 (81%)		
zwei adäquate Experimente			55 (65%)	14 (17%)
	84	84	84	84
Mittelwertvergleich (Wilcoxon-Test)[12]	$p < .01$		$p < .001$	

Tabelle 4.5 – Aufgabe ⟨Balkenwaage⟩: Prozessqualität der Teilkompetenz «Suche im Experimentierraum» im 6. Schuljahr, Auswertung der Summenitems ⟨N1E23i01⊕02⟩ und ⟨N1E23i04⊕05⟩ mit Unterdrückung der Wirkungen der Geltungsasymmetrie

Gegenteil der Fall und es würde sich bei symmetrischer Belastung zum Beispiel die rechte Seite der Waage senken, so müsste man die Waage nur um 180° in der Horizontalen drehen und man hätte bei identischer Ursache die gegenteilige Wirkung, i. e. die linke Seite der Waage würde sich senken. Die Annahme einer asymmetrischen Wirkung führt also zum Widerspruch, da dieselbe Ursache gegenteilige Wirkungen hätte, weshalb die Annahme der asymmetrischen Wirkung verworfen wird (Genz, 1996, 128; Mach, 1963 [1883], 10f). Weil die beiden Hypothesen H_1 und H_2 nun formallogisch äquivalent sind, lässt sich das reductio-Argument prinzipiell auch auf die Hypothese H_2 anwenden, weshalb aus der symmetrischen Prämisse (Gleichgewicht) eigentlich die symmetrische Konklusion (symmetrische Belastung) folgen sollte. Dieser Schluss ist jedoch weder plausibel noch empirisch richtig. Obwohl die beiden Hypothesen formallogisch äquivalent sind, ergibt das Symmetrie-Argument nur bei einer der Hypothesen einen plausiblen Schluss. Der tiefere Grund für diese Asymmetrie liegt darin, dass zwischen der Belastung und dem Gleichgewicht einer Waage nur in einer Richtung eine kausale Abhängigkeit plausibel erscheint. Die Vorstellung, dass das Gleichgewicht einer Waage deren Belastung beeinflusst, erscheint absurd und widerspricht auch gängigen Alltagserfahrungen. Die Hypothese H_2 verwirrt, weil sie gängigen kausalen Überzeugungen widerspricht, die Hypothese H_1 ist vertraut, weil sie alltäglichen Erfahrungen entspricht.

Asymmetrie der adäquaten Manipulationsstrategien. Um eine Hypothese vom Typ «Alle A sind B» zu falsifizieren, müssen Fälle gefunden werden, welche die beiden Eigenschaften A und $\neg B$ vereinen. Die Suche nach solchen Gegenbeispielen kann auf zwei Weisen erfolgen: Entweder werden im Experimentiersuchraum zuerst Fälle mit der Eigenschaft A realisiert, die man dann nach Beispielen mit der Eigenschaft $\neg B$ durchforstet. Oder es werden zuerst Fälle mit der Eigenschaft $\neg B$ produziert, welche nach Beispielen mit der Eigenschaft A durchsucht werden. In der konkreten Experimentiersituation wird man am gegebenen Objekt zuerst eine Eigenschaft realisieren und dann versuchen, am Objekt die andere Eigenschaft zu produzieren, ohne dabei die erste Eigenschaft zu verändern. Der beschriebene Zweischritt besteht also aus einer *freien Realisation* $\mathrm{R}(E_1)$ der ersten Eigenschaft E_1 und einer *bedingten Realisation* $\mathrm{R}(E_2 \mid E_1)$ der zweiten Eigenschaft E_2. Die Schwierigkeit besteht nun nicht nur darin zu wissen, welche Eigenschaften man korealisieren muss, um eine Hypothese zu widerlegen, sondern auch zu wissen, wie man diese Eigenschaften am Objekt realisiert. Das Know-how umfasst praktisches und theoretisches Wissen und bestimmt die Anforderung einer freien Realisation massgeblich mit. Eine bedingte Realisation erfordert die Verknüpfung des Know-hows für zwei Eigenschaften, sofern die beiden Eigenschaften nicht kausal unabhängig sind. In diesem Fall erweist sich eine bedingte Realisation ebenfalls als frei. Die Schwierigkeit verschiedener *Manipulationsstrategien* hängt somit vom erforderten Know-how sowie dem Vorliegen einer kausalen Verknüpfung der Eigenschaften ab.

Wenden wir nun diese kurze Theorie auf das Beispiel der Balkenwaage-Aufgaben H_1 und H_2 an, um die Anforderungen verschiedener Experimentierstrategien zu vergleichen. Hierfür gehen wir von einer Testperson aus, die über hinreichendes Konzeptwissen verfügt, um korrekt zu entscheiden, ob eine Balkenwaage symmetrisch oder asymmetrisch belastet bzw. ob sie im Gleichgewicht oder im Ungleichgewicht ist. Nehmen wir zudem an, dass die Testperson kein hinreichendes Vorwissen zu den Gesetzmässigkeiten am Hebel besitzt und deshalb die Wahrheit und Falschheit einer Hypothese als gleich wahrscheinlich einschätzt. Im Experiment wird diese Testperson daher sowohl die *Bestätigungsstrategie* BS als auch die *Widerlegungsstrategie* WS verfolgen, i. e. sie wird zu einer Hypothese sowohl bestätigende als auch widerlegende Fälle suchen. Zu jeder Experimentierstrategie (Bestätigung bzw. Widerlegung) kann die Testperson aus zwei möglichen Manipulationsstrategien auswählen: Entweder manipuliert sie zuerst die Belastung der Balkenwaage und dann den Gleichgewichtszustand oder sie beginnt mit dem Gleichgewichtszustand und stellt dann die Belastung ein. Insgesamt stehen der Testperson pro Hypothese vier relevante Manipulationsstrategien zur Verfügung (cf. Tab. 4.6).

Eine theoretische Analyse möglicher Hürden, die den Erfolg einer Manipulationsstrategie beeinträchtigen, lassen fünf Strategieprofile unterscheiden, für diese haben wir eine

Aufgabe	Experimentierstrategie		Manipulationsstrategie			
Hypothese	**Strategie**	**Experiment**	**Variante**		**Hürden**	**Stufe**
H_1: $G \supset S$	Bestätigung:	$S \cap G$	BS_{11}:	$\frac{R(S)}{R(G \mid S)\ !}$	keine keine	I
			BS_{12}:	$\frac{R(G)}{R(S \mid G)}$	Wissen "stat. Hebelgesetz" Wissen "dyn. Hebelgesetz"	IV
	Widerlegung:	$S \cap \neg G$	WS_{11}:	$\frac{R(S)}{R(\neg G \mid S)\ \lightning}$	keine Wissen "stat. Hebelgesetz"	II
			WS_{12}:	$\frac{R(\neg G)}{R(S \mid \neg G)\ \lightning}$	Wissen "stat. Hebelgesetz" Wissen "dyn. Hebelgesetz"	V
H_2: $S \supset G$	Bestätigung:	$S \cap G$	BS_{21}:	$\frac{R(S)}{R(G \mid S)\ !}$	keine keine	I
			BS_{22}:	$\frac{R(G)}{R(S \mid G)}$	Wissen "stat. Hebelgesetz" Wissen "dyn. Hebelgesetz"	IV
	Widerlegung:	$\neg S \cap G$	WS_{21}:	$\frac{R(\neg S)}{R(G \mid \neg S)}$	keine Wissen "stat. Hebelgesetz"	III
			WS_{22}:	$\frac{R(G)}{R(\neg S \mid G)}$	Wissen "stat. Hebelgesetz" Wissen "dyn. Hebelgesetz"	IV

Tabelle 4.6 – Aufgabe ⟨Balkenwaage⟩: Manipulationsstrategien bei den Aufgaben H_1 und H_2 (!: Die Realisation ist prinzipiell immer erfolgreich; ↯: Die Realisation ist prinzipiell nie erfolgreich)

empirisch noch zu überprüfende Anforderungsstufung vorgenommen, die neben den Hürden auch die Erfolgswahrscheinlichkeit einer Strategie berücksichtigt (cf. Tab. 4.7). Die Analyse zeigt, dass die Bestätigung (BS_{21}) der Hypothese H_2 eine weniger anspruchsvolle Manipulationsstrategie erfordert als deren Widerlegung (WS_{21}, cf. Tab. 4.6). Der einfachste Versuch, die Hypothese H_1 zu widerlegen (WS_{11}), führt zudem automatisch zur Bestätigung derselben (BS_{11}). Hinsichtlich der Manipulationsstrategien ist die Hypothese H_1 einfacher erfolgreich zu testen als die Hypothese H_2.

Zusammenfassung. Der signifikante Unterschied zwischen den Anforderungen der beiden diskutierten Aufgaben lässt sich auf die Geltungsasymmetrie (H_1 kann bestätigt, aber nicht widerlegt werden; H_2 kann sowohl bestätigt als auch widerlegt werden), auf den asymmetrischen doxastischen Status (H_1 erscheint logisch, H_2 verwirrt), auf die Asymmetrie der Experimentierstrategien (die Widerlegung ist logisch anspruchsvoller als die Bestätigung) sowie auf die Asymmetrie der Manipulationsstrategien (die Realisation einer bestimmten Belastungssymmetrie $R(Bel)$ ist einfacher als die Realisation eines bestimmten Gleichgewichtszustandes $R(Gew)$) zurückführen. Einflüsse der Kodierung auf

	Realisation	Beschreibung der Hürde
	R(Bel)	Die erfolgreiche Realisation einer bestimmten Belastung erfordert kein empirisches Wissen.
	R(Gew)	Die erfolgreiche Realisation eines Gleichgewichtszustandes bedingt das Wissen über Hebelwirkungen bei statischer Belastung.
	R($Gew \mid Bel$)	Die erfolgreiche Suche eines Gleichgewichtszustandes unter einer Belastungsbedingung erfordert das Wissen über Hebelwirkungen bei statischer Belastung.
	R($Bel \mid Gew$)	Die erfolgreiche Suche eines Belastungszustandes unter einer Gleichgewichtsbedingung erfordert das Wissen über Hebelwirkungen bei dynamischer Belastung.

Stufe	Strategie	Beschreibung der Hürde
I	R(Bel) / R($Gew \mid Bel$) !	Der Erfolg der ersten Realisation erfordert kein empirisches Wissen. Der Erfolg der zweiten Realisation stellt sich mit dem Erfolg der ersten Realisation automatisch ein. Er stellt daher keine weitere Hürde dar.
II	R(Bel) / R($Gew \mid Bel$)	Der Erfolg der ersten Realisation erfordert kein empirisches Wissen. Der Erfolg der zweiten Realisation bedingt Wissen über Hebelwirkungen bei statischer Belastung.
III	R(Bel) / R($Gew \mid Bel$) ↯	Der Erfolg der ersten Realisation erfordert kein empirisches Wissen. Der Erfolg der zweiten Realisation bedingt aber Wissen über Hebelwirkungen bei statischer Belastung. Die Unmöglichkeit der Realisation kann den Experimentierprozess erschweren.
IV	R(Gew) / R($Bel \mid Gew$)	Der Erfolg der ersten Realisation bedingt das Wissen über Hebelwirkungen bei statischer Belastung. Der Erfolg der zweiten Realisation bedingt zudem das Wissen über Hebelwirkungen bei dynamischer Belastung. Die Unmöglichkeit des zweiten Schrittes kann den Erkenntnisprozess erleichtern.
V	R(Gew) / R($Bel \mid Gew$) ↯	Der Erfolg der ersten Realisation bedingt das Wissen über Hebelwirkungen bei statischer Belastung. Der Erfolg der zweiten Realisation bedingt zudem das Wissen über Hebelwirkungen bei dynamischer Belastung. Die Unmöglichkeit des zweiten Schrittes kann den Erkenntnisprozess erschweren.

Tabelle 4.7 – Aufgabe ⟨Balkenwaage⟩: Theoretische Stufung der Manipulationsstrategien (!: Die Realisation ist prinzipiell immer erfolgreich; ↯: Die Realisation ist prinzipiell nie erfolgreich)

den Anforderungsunterschied schliessen wir aufgrund der diskutierten Analysen aus.

Beobachtungen der Testleiterinnen und -leiter des Experimentiertests stimmen darin überein, dass die Schülerinnen und Schüler während der Tests praktisch ununterbrochen

damit beschäftigt waren, die Gewichte auf der Balkenwaage durch Hinzusetzen, Wegnehmen oder Verschieben zu manipulieren. Grundsätzlich liessen sich die zwei Aufgaben durch reines Denken lösen, trotzdem wurde vor allem experimentiert. Welche Manipulationsstrategien die Schülerinnen und Schüler dabei mit Erfolg anwendeten, wissen wir nicht. Theoretische Überlegungen lassen aber den begründeten Schluss zu, dass Manipulationen einfacher sind, wenn sie ausgehend von der Belastungssituation der Balkenwaage gedacht und realisiert werden. Geht man nun davon aus, dass diese Strategien auch bevorzugt werden, führt jeder Versuch, H_1 zu widerlegen, automatisch in eine Bestätigung von H_1. Dass H_1 durch die Experimente bestätigt wird, wird hingegen auch im 6. Schuljahr nur wenige Schülerinnen und Schüler überraschen und erklärt den hohen Anteil von 69% korrekter Lösungen zur Hypothese H_1. Defizite in der Entwicklung des logischen Verständnisses von Aussagen wie «Alle A sind B» könnten letztlich die Mühe der 6. Klässer mit der Widerlegungsstrategie und den Gegensatz zur Performance in der 9. Schulstufe erklären: Während im 6. Schuljahr 69% der Schülerinnen und Schüler kein Experiment vorschlagen, dass sich für die Falsifizierung von H_2 eignet, sind es im 9. Schuljahr nur noch 17% der Schülerinnen und Schüler. Gegen diese Erklärung spricht hingegen das Ergebnis, dass nur die Teilkompetenz «Suche im Experimentierraum» signifikante Stufenunterschiede aufweist. Bei der reinen «Datenanalyse» sind die Unterschiede nicht signifikant.[14]

4.3 Modellierung kompetenzrelevanter Itemschwierigkeit

Zur Modellierung kompetenzrelevanter Itemschwierigkeit gibt es grundsätzlich drei Ansätze: Die Modellierung kann durch die Differenzierung der Aufgabenstellung (KP1), durch die Differenzierung der Aufgabenlösungen (KP2) oder durch eine kombinierte Differenzierung beider Ebenen erfolgen. Der erste Ansatz wird zum Beispiel von Mannel et al. (2010) verfolgt (sofern dies anhand der bisher veröffentlichten Unterlagen beurteilt werden kann). Die Itemschwierigkeit der Papier-und-Bleistift-Aufgaben wird durch das *erforderliche Vorwissen*, die angesprochenen *kognitiven Prozesse* und die *konzeptuelle Komplexität* modelliert (S. 49f). Die Beurteilung der Prozessqualität geschieht anhand einer einheitlichen Kodierung. Der zweite Ansatz entspricht der Modellierung von Wellnitz et al. (2010): Jede Teilkompetenz wird mit Hilfe eines einheitlichen Aufgabenformats gemessen, wobei die Differenzierung durch die Kodierung der Prozessqualität anhand eines ordinalen Kodierschemas erfolgt (cf. die Stufenmodelle von Mayer et al. (2008); Tab. 3.11, 3.12 und 3.13 auf S. 38ff). Die Kombination beider Ansätze ist u. W. bislang noch nicht realisiert

[14] cf. Fussnote 9

worden.

4.3.1 Arbeitsschritt ‹Problem lösen›

Mit dem Begriff des Problemlösens betreten wir in den folgenden Abschnitten ein breites Feld der psychologischen Forschung. Wir werden auf einige wenige Resultate dieses Forschungsgebiets zurückgreifen, die uns bei der Erfassung von Experimentierproblemen als zweckdienlich erscheinen.

Eine Theorie des experimentellen Problemlösens? Der Arbeitsschritt ‹Problem lösen› umfasst alle Handlungsprozesse einer Testperson, die nicht mit dem Erfassen der Aufgabe oder mit dem Geben der Antwort in Verbindung stehen. Die Modellierung dieser Prozesse setzt voraus, dass vorerst alle relevanten Informationen aus der Aufgabenstellung zum Gegebenen (inhaltliche Inputs, Verwendungszweck des Experimentiermaterials) und zum Gesuchten (Aufgabenziele) korrekt erfasst und interpretiert vorliegen. Der Arbeitsschritt ‹Problem lösen› bezieht sich sodann auf die Transformation der gegebenen Information in die gesuchte Information, wobei in diesen Informationsverarbeitungsprozess zusätzliches Vorwissen und speziell bei Experimentier- und Simulationsaufgaben zusätzlich generierte Informationen einfliessen (cf. Abb. 4.1). Je nach Aufgabentyp tragen die verschiedenen Informationsquellen (Aufgabeninput, Vorwissen und generierte Information) unterschiedlich viel zur gesuchten Information bei. Und je nach Anteil der generierten Information an der gesuchten Information steht bei einer Aufgabe die Informationsgenerierung oder die Informationstransformation im Vordergrund. Probleme beim ‹Problem lösen› sind somit entweder ein *Problem des Vorwissens*, ein *Problem der Informationsgenerierung* oder ein *Problem der Informationstransformation*.

Zur experimentellen Informationsgenerierung, verstanden als Experimentieren im Sinne eines *discovery process*, gibt es in der Literatur zwei Auffassungen. Die eine Sichtweise erkennt das Experimentieren als Prozess, bei dem neue Konzepte auf der Basis experimentell generierter Evidenz gebildet werden (Bruner et al., 1956). Diese Sichtweise betont den induktiven Aspekt der Informationsgenerierung. Die zweite Sichtweise fasst das Experimentieren als Problemlöseprozess auf, der als Suchprozess in einem komplexen Suchraum charakterisiert wird (H. A. Simon, 1977) und der strukturell mit den Suchprozessen von nicht experimentellen Problemlöseaufgaben vergleichbar ist (Klahr & Dunbar, 1988, 4). Der Suchprozess kann dabei als parallele Suche in verschiedenen Suchräumen modelliert werden (e. g. in einem Hypothesen- und in einem Experimentiersuchraum (Klahr & Dunbar, 1988)), die auch den Aspekt der Konzeptbildung- bzw. Konzeptentdeckung umfassen kann (Dunbar, 1993). Diese zweite Sichtweise fängt somit nicht nur den deduktiven Aspekt des Experimentierens ein, sondern ermöglicht auch die Integration induktiver Aspekte ins

Modell. Wir werden daher für die weitere Diskussion die Konzeption des Experimentierens als Problemlöseprozess übernehmen (Garrett, 1986).

Die Psychologie hält eine Vielzahl von Klassifikationen und Taxonomien für das Problemlösen bereit. Meistens beziehen sich die Beschreibungen auf das problemlösende Denken und können nicht ohne weiteres auf die Situation des Experimentierens übertragen werden (cf. Funke, 2003). Einige Ansätze schliessen jedoch problemlösendes Handeln mit ein, wie z. B. die Problemtypologie von Jonassen (2000). Unter den theoretischen Ansätzen, die ebenfalls vorwiegend dem kognitiven Aspekt des Problemlösens gewidmet sind, erhält vor allem die oben erwähnte Zwei-Räume-Theorie von Klahr und Dunbar (1988) in der Naturwissenschaftsdidaktik Beachtung. Trotz der theoretischen Erfolge fehlt bislang eine Theorie des experimentellen Problemlösens, welche die Schwierigkeit einer Experimentieraufgabe erklärt. Eine solche Theorie müsste neben der kognitiven Ebene auch die manipulative Problemebene, sprich die Experimentier- und Manipulationsstrategien (cf. Fallbeispiel ⟨Balkenwaage⟩, S. 66ff), adäquat erfassen. Kognitive Problemtaxonomien und Problemlösetheorien lösen daher das Theorieproblem des experimentellen Problemlösens nicht, sie können aber als Ansätze für eine solche Theorie dienen. Bevor wir Problemlösetheorien zurückgreifen, wenden wir uns einer Grundidee der Kognitionspsychologie zu: der Unterscheidung von internen und externen Schwierigkeitsfaktoren.

Interne versus externe Schwierigkeitsfaktoren. Die Schwierigkeit einer Experimentieraufgabe ist ein Resultat aus dem Verhältnis von Aufgabenkompliziertheit und Problemlösefähigkeit. Für die Schwierigkeit einer Aufgabe gibt es daher immer zwei Gründe. Eine Testaufgabe ist schwierig, weil sie ein kompliziertes Problem stellt oder weil die Fähigkeit der Testperson nicht genügt. Deshalb werden Aufgabenschwierigkeiten sowohl mit internen Faktoren – so genannten Problemmerkmalen, die nur die Aufgabe betreffen – als auch mit externen Faktoren – so genannten Personenmerkmale, die sich auf die Fähigkeiten des Problemlösers beziehen – erklärt (Funke, 2003, 32). Gelegentlich werden neben den Problemmerkmalen auch Merkmale der Problemdarstellung differenziert (Jonassen, 2000, 66). Externe Faktoren sind Anforderungsmerkmale im Sinne von H. E. Fischer und Draxler (2007, 648ff) und beschreiben das Wissen und die allgemeinen Fähigkeiten, die für die erfolgreiche Bearbeitung der Aufgabe notwendig sind. (cf. Abs. 3.1.3, S. 20). Die internen Faktoren beschreiben den Umfang und die Komplexität der Aufgabe. Die Unterscheidung von externen und internen Faktoren korrespondiert mit der Unterscheidung von psychometrischen Schwierigkeitsmodellen und Fähigkeitsmodellen (Rost, 2004b, 667ff). Die Beschreibung von Itemschwierigkeit mit Hilfe von externen Faktoren erfordert

die Unterscheidung von mehreren Fähigkeiten bzw. Teilfähigkeiten der Testpersonen.[15] Die Verwendung von internen Faktoren bedingt hingegen die Unterscheidung von verschiedenen Itemkomponenten.[16] Gemäss unserer Auffassung der Kompetenzprogression betreffen der Aufgabenumfang [A] und die Problemkomplexität [P] interne Problemfaktoren, während die Eigenständigkeit [E] und die Beschreibung von Prozessqualität [Q] letztlich nur mit Hilfe von externen Personenfaktoren gelingt.

Aufgabenumfang [A]

Der Aufgabenumfang (Umfang und Abfolge der Teilaufgaben) ist bei Assessments selten ein Thema. Gemessen wird meist mit standardisierten Testaufgaben nach vorgegebenem Design und fixem Umfang. Ansätze, den Aufgabenumfang zu "variieren", wurden in den APU-Staffeln umgesetzt.

i. *Umfang der Teilaufgaben*: Grundsätzlich darf angenommen werden, dass mit zunehmendem Umfang der Teilaufgaben auch die Schwierigkeit der ganzen Aufgabe zunimmt. Denn mit jeder Teilaufgabe mehr, gibt es auch mehr Fehlermöglichkeiten.

ii. *Abfolge der Teilaufgaben.* In der APU-Staffeln 1982/83 wurde untersucht, welchen Einfluss die Position des "Planungs"-Schrittes in der Abfolge der verschiedenen Teilaufgaben eines Experiments auf die Schülerperformance hat. Entweder wurde das Experiment vor dem Experimentieren geplant oder nach dem Experimentieren rapportiert oder man wurde aufgefordert, post hoc eine Wiederholung des Experiments zu planen (APU, 1985, 209ff). Die Resultate zeigen, dass unterschiedliche Positionen des "Planungs"-Elementes in der Aufgabenabfolge qualitativ unterschiedliche Aufgaben ergeben. In der konkreten Experimentiersituation zeigen sich Schülerinnen und Schüler kompetenter als anhand ihrer Planung angenommen werden darf. Einen positiven Einfluss hat die vorgängige Planung einzig auf die Darstellung der Ergebnisse (APU, 1985, 209). Das Experimentierverhalten ändert sich jedoch kaum, ob vorgängig ein Planung verlangt wird oder nicht. Hingegen stimmen die "Planungsrapporte" mit dem tatsächlichen Experimentierverhalten gut überein, was zur folgenden Interpretation Anlass gibt: "It would seem that the process of writing is not itself an obstacle; rather it is the need in a non-practical context to carry out 'thought' experiments without any form of feedback to stimulate further thought" (APU, 1989, 130).

Der Aufgabenumfang [A] wirkt also sowohl direkt als auch indirekt auf die Itemschwierigkeit. Die Anzahl Teilaufgaben bzw. Teilprozesse eines Items haben einen direkten Einfluss

[15] u. a. kann das Testmodell MULTIRA von Rost und Carstensen (2002) zu diesem Zweck verwendet werden.

[16] Diese Anforderung erfüllt das linear-logistische Testmodell von G. H. Fischer (1997).

auf die Itemschwierigkeit. Die Abfolge der Teilaufgaben erzeugt qualitativ andere Aufgaben und beeinflusst somit die Prozesskomplexität und letztlich auch die Itemschwierigkeit.

Problemkomplexität [P]

Wir haben eingangs dieses Abschnitts das Experimentieren als Informationsverarbeitungsprozess beschrieben, bei dem Informationen erzeugt und mit bereits vorhandenen Informationen zur Zielinformation verarbeitet werden. Aus der Sicht der Problemlöseforschung stellen sich drei Fragen:

1. Welche Problemmerkmale erhöhen den Verarbeitungsaufwand?
2. Welche Problemmerkmale behindern die Informationsverarbeitung?
3. Welche Problemmerkmale senken den Verarbeitungsaufwand?

Die erste Frage zielt auf allgemeine Merkmale, aus denen sich spezifische Anforderungen ergeben (Dörner, 2006, 59). Unter ihnen wird die *Komplexität* der Problemsituation (Anzahl beteiligter Variablen), die *Vernetztheit* der Variablen, *Intransparenz* von Informationen im Hinblick auf die beteiligten Variablen und die Zielstellung, welche die Informationsbeschaffung erfordert, die *Vielzieligkeit* (Anzahl und Vereinbarkeit der Ziele) sowie die Abstraktheit des Kontexts genannt (Funke, 2003, 126; Jonassen, 2000). Problemmerkmale, welche die zweite Frage beantworten, stellen so genannte Barrieren dar (Dörner, 1976, 10). Als Barrieren gelten Unklarheit über die Zielkriterien (*Zielklarheit*) sowie ein tiefer *Bekanntheitsgrad der Mittel* (Dörner, 1976, 14). Die dritte Frage betrifft Hilfestellungen, welche die Lösung der Aufgabe wahrscheinlicher machen. Im Fall von Experimentieraufgaben sind dies der Grad der Anleitung (*Strukturiertheit*) und die Zahl der akzeptablen Lösungen und Lösungswege (*Problemoffenheit*). Im Folgenden sollen anhand der acht genannten Problemmerkmale (*Komplexität*, *Vernetztheit*, *Intransparenz*, *Vielzieligkeit*, *Zielklarheit*, *Mittelbekanntheit*, *Strukturiertheit*, *Problemoffenheit*) bekannte Ansätze, die Problemkomplexität [P] zu modellieren, vorgestellt werden.

i. *Komplexität und Vernetztheit*: Die Idee, die Anforderung eines Problems über die Anzahl der zu verarbeitenden Elemente und Beziehungen zu erfassen, bedingt, dass die relevanten Elemente und Beziehungen beschränkt und eindeutig gegeben sind. Dies trifft z. B. auf das Problem zu, kausale Abhängigkeiten zwischen bekannten Variablen experimentell zu untersuchen. Für diesen Problemtyp kann die Anforderungsprogression mit Hilfe der Komplexität der Experimentierstrategien modelliert werden (Gott & Duggan, 1995; Mannel et al., 2010). Bislang keine Operationalisierung existiert für die Komplexität der Manipulationsstrategien. Keine sinnvolle Entsprechung für das Komplexitätskonstrukt gibt es weiter bei Beobachtungs- oder

Konstruktionsaufgaben. Die Inhalte einer Beobachtungsaufgabe sind nicht eindeutig und vorgängig auch nicht immer bekannt. Bei Konstruktionsaufgaben übersteigt die Zahl der Elemente, Zusammenhänge und Funktionen rasch das bewältigbare Mass. Mit dem psychometrischen Erfolg von Komplexitätskonstrukten besteht daher die Gefahr, experimentelle Kompetenz einseitig zu definieren, wie dies Gott und Duggan (1995, 48) auf den Punkt bringen: "If we take a restricted view of investigations as being solely to do with variables and numerical data, than large swathes of science, particularly chemistry and those elements of science bordering on technology, can become neglected."

ii. *Abstraktheit des Kontextes*: Das Niveau einer Kompetenzausübung hängt vom spezifischen Kontext ab (Jonassen, 2000). Für dieses Phänomen werden unterschiedliche Aspekte des Kontextes verantwortlich gemacht. Kontexte wirken jedoch nicht unbedingt nur über die Itemschwierigkeit auf die Performance. Sie können auch die Interpretation des Aufgabenziels beeinflussen. Ein erster relevanter Aspekt ist die Kontextfremdheit (cf. S. 17): Kontexte können Testpersonen mehr oder weniger fremd sein. Alltägliche Kontexte sind ihnen vertrauter als wissenschaftliche. Trotzdem zeigen Schülerinnen und Schüler bei Experimentieraufgaben mit wissenschaftlichem Kontext bessere Experimentierstrategien und verwenden mehr Messdaten als bei Experimentieraufgaben mit alltäglichem Kontext (Gott & Duggan, 1995, 59). Alltägliche Kontexte müssen jedoch nicht unbedingt schwieriger sein als wissenschaftliche. Folgende alternative Interpretation erscheint ebenso plausibel: " [...] an everyday context can lead some pupils to the idea that an everyday answer is all that is required and the notion that 'we should only behave scientifically when the task looks scientific'." (Gott & Duggan, 1995, 59). Kontexte enthalten auch Konzepte, die wiederum mehr oder weniger abstrakt sind (Kontextabstraktheit, cf. Gott & Duggan, 1995, 56f). Abstrakte Konzepte mögen eine Experimentieraufgabe schwierig machen. Dieser Effekt kann aber durch andere Einflüsse auf die Itemschwierigkeit verdeckt werden (Gott & Duggan, 1995, 56f).

iii. *Intransparenz*: Die Intransparenz einer Aufgabe erhöht den Bedarf an Informationsbeschaffung. Experimentieraufgaben sind per definitionem intransparent, da die Erzeugung zusätzlicher Information im wahrsten Sinne des Wortes der Zweck des Experimentierens ist. Die Menge und die Ambivalenz der zu beschaffenden Information können für die Schwierigkeit eines experimentellen Problems eine Rolle spielen. Beobachtungen enthalten eher ambivalente Informationen, da die Sinneswahrnehmung stets eine gewissen Interpretation des Subjekts erfordert. Messungen liefern dagegen eher eindeutige Informationen.

iv. Vielzieligkeit und Zielklarheit: Vielzieligkeit bezieht sich auf die Vielzahl an Zielvorstellungen, die eine Lösung eines Problems erfüllen soll. Konstruktionsaufgaben sind vielzielig, wenn die gesuchte Konstruktion verschiedenen funktionalen Anforderungen genügen soll. Large-scale Experimentieraufgaben sind vielzielig, wenn der Inhalt der gesuchten Antwort verschiedene Bedingungen erfüllen soll. Unklarheit über das Ziel besteht, wenn die Zielkriterien, welche die gesuchte Lösung von alternativen Antworten unterscheiden, nicht klar oder nicht bekannt sind.

v. Bekanntheitsgrad der Mittel, Aufgabenstrukturiertheit und Problemoffenheit: Der Bekanntheitsgrad der Mittel wird hier im Zusammenhang mit Problemmerkmalen behandelt, obwohl er eher ein Personenmerkmal darstellt. Der Bezug zu Problemmerkmalen ist gegeben, indem der Bekanntheitsgrad der Mittel durch das reine Problemmerkmal der Strukturiertheit[17] beeinflusst wird. Die Idee des medialen Bekanntheitsgrades ist, dass erfolgreiches Experimentieren umso wahrscheinlicher wird, je besser die Handhabung der benötigten Werkzeuge (Messinstrumente und Materialien) und die notwendigen Verfahren und Vorgehensweisen (Experimentier- und Manipulationsstrategien) bekannt sind. Bei vollkommener Unbekanntheit der Mittel erscheint der Testperson jeder mögliche Lösungsansatz mit gleicher Wahrscheinlichkeit zum Ziel zu führen. Mit steigendem Bekanntheitsgrad der Mittel werden immer mehr mögliche Lösungsansätze unwahrscheinlich und verworfen. Der sinnvolle Experimentiersuchraum schrumpft, wodurch die Wahrscheinlichkeit steigt, dass die Testperson die richtigen Entscheidungen fällt und den korrekten Lösungsweg wählt. Mit angeleiteten und durchstrukturierten Experimentieraufgaben werden den Testpersonen Entscheidungen abgenommen, wodurch der Experimentiersuchraum ebenfalls schrumpft und die Lösungswahrscheinlichkeit steigt. Gemäss dieser Auffassung haben Aufgaben, die mehrere Lösungswegen zulassen – wie in der Abbildung 4.4 schematisch dargestellt – eine höhere Lösungswahrscheinlichkeit, da die Suche im Experimentiersuchraum mehr richtige Treffer zulässt. Die Problemoffenheit, definiert als die Menge der korrekten Lösungswege, beeinflusst somit die Lösungswahrscheinlichkeit auf dieselbe Weise wir der Bekanntheitsgrad der Mittel.

Die Strukturiertheit als Mass für den Umfang des durch eine Aufgabe aufgespannten Experimentiersuchraums wird in der Literatur im Zusammenhang mit weiteren verwandten Problemmerkmalen diskutiert. H. E. Fischer und Draxler (2007, 646) verbinden Aspekte der Problemoffenheit und der Aufgabenstrukturiertheit zu einer kombinierten vierstufigen Offenheitsskala. S. A. Simon und Jones (1992, 12f) verbin-

[17] Der im psychologischen Problemlösediskurs verwendete Begriff der *Structuredness* entspricht nicht dem von uns verwandten Begriff der Strukturiertheit. Das Konzept der Structuredness vereint Aspekte der Zielklarheit, Problemoffenheit und Intransparenz.

den im Rahmen einer Studie zu "open-ended work in science" die Aufgabenstrukturiertheit mit der Zielklarheit und der Problemoffenheit, zu einer übergeordneten Skala der "openness of tasks". Gott und Duggan (1995, 54ff) stellen fest, dass eine offene Fragestellung (*openness of tasks*) die Performance bei der Teilkompetenz «Datenanalyse» signifikant erhöht, hingegen auf die Teilkompetenz «Experimentierstrategien» nur wenig Einfluss hat. Eine offene Frage erhöht den Interpretationsspielraum der Zielstellung und generiert Lösungsalternativen, sofern diese im Kodierschema berücksichtigt werden. Baxter und Glaser (1998) thematisieren die Strukturiertheit mit Hilfe des Prozessraumes (process space), der mehr oder weniger beschränkt beziehungsweise mehr oder weniger offen sein kann. Das Mass der Strukturiertheit kann auch anhand von Experimentiermustern erfasst werden, die in der Unterrichtspraxis erwünscht sind oder beobachteten werden (H. E. Fischer & Draxler, 2007, 647).

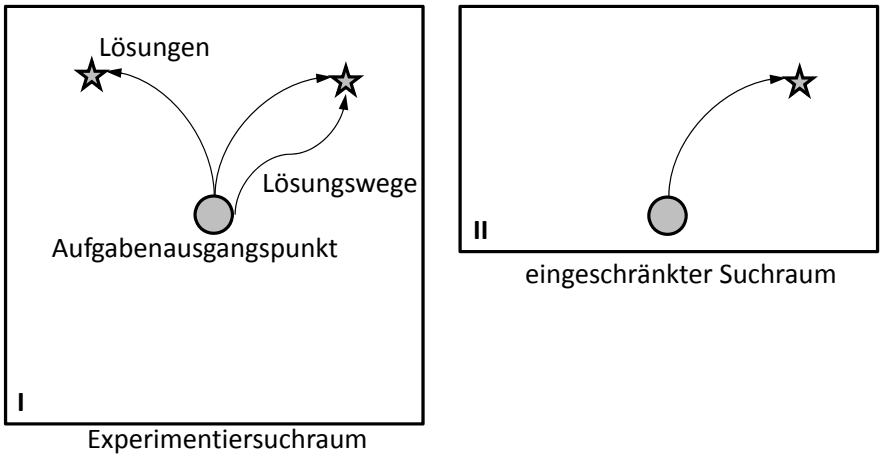


Abbildung 4.4 – Der Experimentiersuchraum I einer Aufgaben mit zwei Lösungen (Zieloffenheit) und drei Lösungswegen (Problemoffenheit) wird durch mehr Strukturierung eingeschränkt (Experimentiersuchraum II)

Transferumfang [T]

Die Fähigkeit, Know-how in verschiedenen Kontexten anzuwenden und auf strukturverwandte Inhalte zu transferieren, stellt ein wichtiger Kompetenzaspekt und ein massgebender Faktor für die Kompetenzprogression dar. Die Schwierigkeit der Transferfähigkeit zeigt sich an der hohen empirischen Performance-Variabilität in Bezug auf die Testaufgaben bei Assessments (cf. S. 55) und lässt sich durch die Abhängigkeit eines erfolgreichen Transfers von spezifischem Inhaltswissen erklären (Millar & Driver, 1987, 51ff). Um die Transferleistung einer Testperson zu messen, werden Tests benötigt, bei denen die Testperson möglichst viele Aufgaben mit einem breiten Umfang an Kontexten bearbeitet. Der

dabei von der Testperson bewältigte Transferumfang [T] ist als einzige Progressionsdimension keine Eigenschaft der einzelnen Testaufgabe, sondern ein Merkmal des ganzen Tests (cf. S. 65). Ein Kompetenzzuwachs in Bezug auf die Transferfähigkeit zeigt sich demnach nicht darin, dass eine Testperson schwierigere Aufgaben besser löst, sondern dass sie längere und umfangreichere Tests besteht. Im Rahmen von Rasch-Analysen fliesst der Transferumfang in die Kompetenzbeurteilung über die Suffizienzbedingung ein, dass bei vollständiger Datenmatrix und Modell-Geltung das Personenscore gleichviel Information in Bezug auf die Fähigkeit der Person enthält wie der geschätzte Personenparameter (cf. Rost, 2004a, 122f, Borg & Staufenbiel, 2007, 350ff).

4.3.2 Arbeitsschritt ‹Lösung kodieren›

Die Beurteilung einer Kompetenzausprägung erfordert nebst der Analyse der Aufgabenstellung auch die Kodierung der Aufgabenlösung, welche Aussagen über die Qualität [Q] und die Eigenständigkeit [E], mit welchen eine Testperson eine Aufgabe bearbeitet, erlaubt. Da bei einem gegebenen Messinstrument die Anforderungen durch die Testaufgaben für alle Testpersonen die gleichen sind, basieren individuelle Unterschiede bei der Kompetenzbeurteilung in erster Linie auf Unterschiede bei der Beurteilung der Prozessqualität und der Eigenständigkeit. Die Frage, welche Aufgaben mit welchem Aufgabenumfang [A] und welcher Problemkomplexität [P] eine bestimmte Testperson nun tatsächlich löst, spielt bei der Bestimmung des individuellen Fähigkeitsparameters im Rahmen einer bei large-scale Assessment üblichen Rasch-Analyse keine Rolle. Eine Gewichtung der Aufgaben aufgrund von Merkmalen der Aufgabenstellung ist bei der Rasch-Analyse nicht vorgesehen und widerspricht dem Modellansatz. Die Analyse des Aufgabenumfangs und der Problemkomplexität der Testaufgaben findet letztlich vor allem darin Verwendung, die Messinstrumente zu verorten, e. g. bei der Analyse von schwierigkeitsinduzierenden Itemmerkmalen.

Eigenständigkeit [E]: Individuelle Assistenz

Die Kodierung der Eigenständigkeit basiert auf der Idee, dass die Testpersonen bei einem Assessment je nach Bedarf unterschiedlich viel Support und Assistenz erhalten. Die Nutzung zusätzlicher Unterstützungsangebote während der Tests muss natürlich individuell festgehalten werden und in irgendeiner Form in die Auswertung einfliessen. Im Rahmen des TIMSS-Experimentiertests wurde bei der Handhabung von Messinstrumenten Hilfe angeboten und separat kodiert.[18] Beim HarmoS-Experimentiertest wurden bei anspruchs-

[18] Bei der Aufgabe ⟨Solutions⟩ wurde bei der Handhabung des Thermometers bei Bedarf Unterstützung angeboten (TIMSS, 1994).

vollen Geräten wie Mikroskop und Leistungsmesser individuell Bedienungshilfen gegeben. Auf eine separate individuelle Kodierung wurde jedoch verzichtet. Letztlich hat die Eigenständigkeit gerade im Unterricht einen grossen Stellenwert und wird bei der Beurteilung durch die Lehrperson mitberücksichtigt. Dies gilt u. a. auch bei den standardisierten englischen Abschlussprüfungen der Sekundarschule (man vergleiche die Kodierungsanleitungen beim General Certificate of Secondary Education; R. W. Fairbrother, 1988, 53f).

Prozessqualität [Q]: Kodiermassstäbe

Validität. Die Beurteilung der Prozessqualität erfolgt ex post anhand von Kodierschemen, welche möglichst genaue Beschreibungen von Kriterien für die Vergabe von Punkten (Credits) enthalten sollten. Mit den Kriterien wird festgelegt, was und mit welchen Massstäben gemessen wird, wobei gerade bei Partial credits-Kodierungen a priori Kompetenzprogressionen vorgegeben werden, die posteriori nicht überprüft werden können (cf. Abschnitte auf Seiten 57 und 90). Um entscheiden zu können, welche Prozesse und welche Qualitäten nun tatsächlich mit einer Experimentieraufgabe gemessen werden, muss das entsprechende Kodierschema analysiert werden. Messick (1996, 11) bringt die Analysen von Kodierschemen mit der allgemeinen Beurteilung der Validität von Assessments in Zusammenhang und formuliert die Notwendigkeit solcher Analysen wie folgt: "Validity is not a property of the test or the assessment as such, but rather of the meaning of the test scores. These scores are a function not only of the items or stimulus conditions, but also of the *persons* responding as well as the *context* of the assessment. In particular, what needs to be valid is the meaning or interpretation of the scores [...]" (vgl. hierzu auch Cronbach (1971)). Im folgenden Abschnitt fassen wir die Ergebnisse einer qualitativen Analyse von Kodierschemen unterschiedlicher Experimentiertests zusammen. Die Analyse stützt sich auf einer Untersuchung sämtlicher Aufgaben des HarmoS- und des TIMS-Experimentiertests (TIMSS, 1994) sowie ausgewählter Aufgaben der NAEP-Assessments (Brown & Shavelson, 1996; Shavelson, Solano-Flores & Ruiz-Primo, 1998; Solano-Flores, 1994; Solano-Flores & Shavelson, 1997; Solano-Flores, Shavelson, Ruiz-Primo, Schultz & Wiley, 1997; Solano-Flores et al., 1999), der APU-Staffeln (APU, 1981, 1984, 1985, 1989), des QuiP-Experimentiertests (Zberg, preprint) und aus weiteren Forschungsarbeiten zur Messung experimenteller Kompetenz (e. g. Hammann et al., 2006, 2008).

Qualitätskategorien, Massstäbe und Kriterien. Beim Kodieren werden üblicherweise unterschiedliche Aspekte einer Lösung beurteilt, die bestimmten Qualitätskategorien zugeordnet werden können. Dabei werden an eine Aufgabenlösung verschiedene Kodiermassstäbe angelegt, von denen jeder eine bestimmte Qualitätskategorie "vermisst". Die Massstäbe werden in den Kodierschemen mit Hilfe von Kriterien beschrieben. Im Idealfall teilt

jedes Kriterium (K) die Menge aller möglichen Antworten Ω einer Aufgabe in eindeutiger Weise in eine Menge K mit Antworten, die K erfüllen, und in eine Menge $\bar{K}$ mit Antworten, die K nicht genügen ($K \cup \bar{K} = \Omega$). In realen Kodierschemen kommen jedoch auch *konditionale Kriterien* zur Anwendung, die nur auf Antworten sinnvoll angewandt werden können, die bereits ein anderes Kriterium erfüllen.

Oft werden dichotome Massstäbe verwandt, die aus einem einzigen Kriterium bestehen und die für sich alleine nur eine zweiwertige Kodierung erlauben. Bei höherwertiger Kreditierung werden entweder polytome Massstäbe eingesetzt oder es werden dichotome Massstäbe kombiniert. Dabei werden mehrere Massstäbe mit Hilfe logischer und rechnerischer Operationen zu einer mehrwertigen Skala unter dem Prinzip verknüpft, dass eine Antwort qualitativ um so besser eingestuft wird, je mehr und anspruchsvoller die Kriterien sind, denen sie genügt. Da jedoch zwei Massstäbe auf unterschiedliche Arten zu einer mehrwertigen Skala kombiniert werden können, unterliegt die Verwendung kombinierter Massstäbe stets einer gewissen Beliebigkeit.

Kompetenzspezifische Massstäbe: Beurteilungsgegenstände versus Ideale Grundsätzlich bieten sich zwei Arten an, kompetenzspezifische Massstäbe zu klassifizieren. Zum Einen können die Gegenstände unterschieden und klassifiziert werden, die mit den Massstäben beurteilt werden. Im Fall der experimentellen Kompetenz lassen sich dann *ergebnisorientierte* und *prozessorientierte Kodierungen* unterscheiden (cf. Hodson, 1992, 118f). Ergebnisorientierte Kodierungen richten den Fokus auf ein Produkt wie die Beschreibung oder Skizze einer Beobachtung, die Daten von Messungen, das Ergebnis einer Ableitung oder das Resultat einer Berechnung. Demgegenüber sind bei prozessorientierten Kodierungen die Handlungen und Vorgehensweisen, die zu diesen Produkten führen, Gegenstand der Bewertung. Während bei Rechenaufgaben zwischen Prozess und Produkt hinreichend scharf unterschieden werden kann – Resultat und Lösungsweg lassen sich meist klar unterscheiden – verschwimmen die Grenzen bei Experimentieraufgaben, bei denen die Prozesse auf der Basis von Eigenrapporten beurteilt werden. Werden nämlich Prozesse rapportiert, werden einerseits die Rapportinhalte, also die Prozesse, und andererseits der Rapport als Produkt und Ergebnis von Lösungshandlungen bewertet. Dies trifft u. a. auf Planungsaufgaben zu. Die Planung eines Experiments ist einerseits ein Produkt, das als Ergebnis aus der Bearbeitung der entsprechenden Aufgabe entstanden ist. Die Bewertung der Planung anhand von formalen und inhaltlichen Qualitätsvorstellungen, die allgemein an eine Planung gestellt werden, ist demzufolge ergebnisorientiert. Andererseits enthält eine Planung Beschreibungen von (beabsichtigten) Handlungen. Die Bewertung der Adäquatheit dieser Handlungen kann als prozessorientiert gedeutet werden, sofern die Planung nachfolgend im realen Experiment umgesetzt wird. Die Überlegungen, die zur Planung eines Experi-

ments geführt haben – sprich der Lösungsweg bei Planungsaufgaben – werden hingegen praktisch nicht bewertet. In diesem Sinne erfolgt die Kodierung von Planungsaufgaben nicht prozessorientiert. Zu unterscheiden gilt daher prozessorientiert als handlungsbezogen und prozessorientiert als auf den Lösungsweg bezogen.

Die zweite und ergiebigere Klassifikation von Massstäbe, die wir vorschlagen, basiert auf der Unterscheidung von Wissensarten, die benötigt werden, um eine Aufgabe korrekt zu bearbeiten und die mit den Massstäben beurteilt werden. Zu jedem Wissen definieren wir Ideale für deren korrekte Anwendung. Die Analyse der Kodierschemen ergab in Bezug auf die experimentelle Kompetenz die Unterscheidung von sechs "*Korrektheitsidealen*".

- *Theoretische Korrektheit*: Viele Resultate von Experimentieraufgaben können aufgrund von fachlichem Vor- und Hintergrundwissen hergeleitet werden, ohne die beschriebenen Experimente tatsächlich durchführen zu müssen. Dies gilt z. B. für die im Abschnitt 4.2.1 diskutierte Balkenwaage-Aufgabe. Wer die Hebelgesetze kennt, weiss, aus welchen Experimenten welche Ergebnisse resultieren, und richtet sein experimentelles Handeln danach aus. Sollte ein Experiment nicht den theoretisch erwarteten Ausgang ergeben, wird die theoriekundige Testperson das Experiment wiederholen, während die unkundige Testperson das Resultat wahrscheinlich als korrekt akzeptieren wird. Mit der Balkenwaage-Aufgabe wird somit vor allem theoretisches Wissen geprüft, die Kodierung erfolgt anhand des Standards der theoretischen Korrektheit.[19]
  Theoriebasierte Kodierungen von Mess- und Beobachtungsergebnissen sind problematisch, wenn natürlichen Schwankungen von Messergebnissen keine Beachtung geschenkt wird. Korrektes Experimentieren kann aufgrund zufällig "falscher" Resultate sanktioniert werden. Dies kann fallweise durch die Kombination theorie- und evidenzbasierter Kodierung vermieden werden.

- *Evidenzielle Korrektheit/Präzision*: Daneben gibt es Resultate von Experimentieraufgaben, die man theoretisch nicht herleiten kann, entweder weil zu wenig kontingente Informationen zur Verfügung stehen oder weil die passende Theorie nicht vorhanden ist. Dies gilt beispielsweise für die ersten zwei Items der HarmoS-Aufgabe ⟨Solarzellen⟩ (cf. App. A.2). Die Frage des zweiten Items ⟨N9E84i02⟩, wie weit eine Tischlampe von einer Schaltung mit einer Solarzelle und einem Motor entfernt werden darf, ohne dass der Motor aufhört zu drehen, kann nur experimentell bestimmt werden. Bei der Kodierung des Resultats orientiert man sich in diesem Fall nicht an der Theorie, sondern

[19] Theoriebasierte Kodierung gibt es nicht nur bei der Bewertung von Mess- oder Beobachtungsergebnissen, sondern auch bei der Beurteilung von Experimetieransätzen mit Kontrollvariablen. Die Ausscheidung einer Kontrollvariable erfolgt aufgrund von theoretischem Vorwissen über mögliche kausale Wirkungen. Eine "korrekt" identifizierte Kontrollvariable ist daher immer nur im Bezug auf eine gegebene Theorie korrekt.

am Standard von Testmessungen, die durch "Experten" ausgeführt wurden. Diese Art der Kodierung erfolgt evidenzbasiert. Typisch für evidenzbasierte Kodierungen ist, dass meist nicht die "Korrektheit" der Resultate, sondern die Präzision der Messungen bzw. Beobachtung bewertet wird.

- *Technisch-praktische Korrektheit*: Neben dem theoretischen Vorwissen spielt bei Experimentieraufgaben auch technisch-praktisches Wissen eine Rolle. Die Aufgabe, die Stärke zweier Magnete mit Hilfe von Büroklammern zu vergleichen (cf. TIMSS-Aufgabe ⟨Magnets⟩, TIMSS, 1997, 21ff) wird nicht aufgrund von theoretischem Vorwissen gelöst, sondern vor allem mit Hilfe von praktischem Hintergrundwissen. Dies gilt ganz allgemein für Entwicklungsaufgaben von Messverfahren oder technische Konstruktionen, die eine bestimmte Funktion erfüllen sollen. Die "Korrektheit" eines Messverfahrens oder einer technischen Konstruktion erweist sich im praktischen Test. Das für die Entwicklung notwendige Wissen ist jedoch nicht evidentieller oder theoretischer Natur, sondern umfasst technisches und praktisches Wissen, das teilweise nicht explizit vorliegt.[20]

- *Heuristische Korrektheit*: Bei der Planung- und Durchführung von Experimentieraufträgen interessiert nicht nur die "Korrektheit" der Ergebnisse, sondern auch das methodisch adäquate Vorgehen. Erwartet wird z. B. bei Aufgaben, bei denen ein kausaler Zusammenhang untersucht werden soll, dass Test- und Kontrollvariablen unterschieden werden, mit den Testvariablen systematisch umgegangen wird und die Kontrollvariablen konsequent kontrolliert werden (Hammann et al., 2006, 293). Bei der Beurteilung der heuristischen Korrektheit wird das methodische Vorgehen bei einer Experimentieraufgabe bewertet. Dabei wird letztlich das heuristische Wissen der Testpersonen überprüft.

- *Schlusslogische Korrektheit*: In Experimentiertests ebenfalls bewertet wird die Fähigkeit, gegebene Daten logisch korrekt in Bezug zu einer gegebenen Fragestellung bzw. Hypothese zu stellen. Beispielsweise wird mit dem dritten Item ⟨N9E23i03⟩ der HarmoS-Aufgabe ⟨Solarzellen⟩ (cf. App. A.1, S. 214ff) kodiert, ob aus den Ergebnissen korrekt auf die Wahrheit der zu überprüfenden Behauptung geschlossen wird. Dabei wird das Wissen über die korrekte Schlussweisen der Testpersonen geprüft.

[20] Woolnough und Allsop (1985, 33f) verweisen auf den enormen Stellenwert des "tacit knowledge" bei der Entwicklung von Kompetenzen, wie sie Experten auszeichnen: "This form of learning is very important in the development of scientists, who, through experience, 'get a feel for' or an 'awareness of' phenomena. When making a device, or solving a problem, they will 'know' what material to use and which lines of attack will work, not because they have developed a formal understanding of the properties of materials or the contents of the problem, but because they have developed a feel for them; through experience." Der Begriff des tacit knowledge geht zurück auf eine Differenzierung zwischen explizitem und implizitem Wissen von Polanyi (1969).

- *Mathematisch-logische Korrektheit*: Sofern in Experimentieraufgaben mathematische Teilaufgaben gelöst werden, wird auch mathematisch-logisches Wissen getestet.

Mit Ausnahme der evidentiellen Korrektheit beziehen sich alle vorgestellten Massstäbe auf Vorwissen, das eine Testperson zum Test mitbringt. Das theoretische, heuristische, praktische und logische Vorwissen wird gebraucht, um neue vertrauenswürdige Informationen – Evidenz – zu erzeugen. Demgegenüber spricht die evidentielle Korrektheit kontingentes Wissen an, das eine Testperson erst beim Lösen der Testaufgabe erwirbt. Der Erwerb dieses Wissens macht jedoch im Wesentlichen geraden den Inhalt der zu messenden experimentellen Kompetenz aus, weshalb evidentielle Massstäbe bei der Kompetenzmessung eine gesonderte Stellung haben.

Korrespondenz von Aufgabentypen und Kodierschemen. Nicht jeder Massstab kann bei jeder Aufgabe sinnvoll angewendet werden. Im Gegenzug erfordern gewisse Aufgabentypen (Prozesse) natürlicherweise die Verwendung bestimmter Massstäbe. Auf diesem Zusammenhang beruht die von Ruiz-Primo und Shavelson (1996) festgestellte Korrespondenz zwischen Aufgabentypen und Kodiersystemen (Shavelson & Ruiz-Primo, 1999; Solano-Flores & Shavelson, 1997). Die Tabelle 4.8 enthält eine Darstellung dieser Korrespondenz und eine Übersetzung in das System der oben eingeführten Kodiermassstäbe. Die Korrespondenz legt die Idee nahe, für verschiedene Aufgabentypen unterschiedliche Kodiersysteme zu entwickeln. Für Vergleichs-, Klassifikations- und Beobachtungsaufgaben wurden Grundraster für Kodiersysteme entwickelt und evaluiert (Shavelson et al., 1998; Solano-Flores, 1994; Solano-Flores et al., 1997).

Allgemeine, kompetenzunspezifische Massstäbe. Neben diesen fünf wissensspezifischen Massstäben werden in Experimentiertests weitere kompetenzunspezifische Massstäbe angewandt, die bei jeder Kodierung teils notwendigerweise, teils implizit berücksichtigt werden.

- *Vollständigkeit*: Von einer Lösung wird erwartet, dass sie vollständig ist. Einerseits sollen alle Antwortformate, im Sinn von Lückenformaten, bearbeitet werden (e. g. beim ersten Item der Balkenwaage-Aufgabe soll sowohl das Experiment in die Abbildung eingezeichnet als auch der Gleichgewichtszustand der Waage mit Ankreuzen qualifiziert werden, cf. App. A.1). Andererseits sollen inhaltliche Vorgaben in der Aufgabenstellung bei der Antwort berücksichtigt werden (wenn beispielsweise für die Planung eines Experiments sowohl eine Beschreibung der Vorgehensweise als auch eine Skizze der Experimentieranordnung verlangt werden, wird das Fehlen eines von beidem sanktioniert).

Task typ*	**Scoring system***	**Massstab**
Comparative investigation: "Student conducts an experiment to compare two or more objects on some property."	**Procedure-based**: "[...] it focuses on the scientific defensibility of the procedures used by the student to compare the objects."	*heuristische Korrektheit*
Component identification: "Student tests objects to determine their components or how those components are organized."	**Evidence-based**: "[...] it focuses on the quality of the evidence used to confirm or disconfirm the presence of components."	*technisch-praktische und evidentielle Korrektheit*
Classification: "Student classifies objects according to critical attributes to serve a practical or conceptual purpose."	**Dimension-based**: "[...] it focuses on how well the classification system constructed uses attributes that are relevant to the purposes of classification."	*theoretische Korrektheit*
Observation: "Student performs observations and/or models a process that cannot be manipulated."	**Accuracy-based**: "[...] it focuses on the accuracy of the observations performed and the models constructed."	*evidentielle Korrektheit*

Tabelle 4.8 – Korrespondenz zwischen Aufgabentyp und Kodierschema: Übersetzung der Korrespondenz gemäss Ruiz-Primo und Shavelson (1996)* in das Massstabsmodell; Zitate aus Solano-Flores und Shavelson (1997, 19)

- *Darstellung*: Neben inhaltlichen Vorgaben werden auch formale Vorgaben gemacht, beispielsweise wenn verlangt wird, Messwerte in tabellarischer Form darzustellen. Sofern die verwandte Kompetenzdefinition nicht explizit solche Fähigkeiten beinhaltet, sollte die Darstellungsqualität separat kodiert und dem Arbeitsschritt ‹Antwort geben› zugeordnet werden, wie dies u. a. bei der TIMSS-Aufgabe ⟨Solutions⟩ gemacht wurde (TIMSS, 1997, 43ff).

- *Konsistenz*: Von einer Lösung wird letztlich erwartet, dass sie konsistent ist. Widersprechen sich Teile einer Antwort, wird das Prinzip angewandt "In dubio contra reo". Der Wert einer korrekten Antwort wird durch eine weitere konträre Antwort herabgesetzt. Dies gilt insbesondere bei widersprüchlichen Antworten von gemeinsam kodierten Teilfragen.

Ordinalität polytomer und kombinierter Massstäbe. Mehrstufige Kodierskalen werden durch Verknüpfen mehrerer Kriterien erzeugt. Beziehen sich die Kriterien hierbei ausschliesslich auf eine Qualitätskategorie, sprechen wir von *einem polytomen Massstab*. Beziehen sie sich auf verschiedene Qualitätskategorien handelt es sich um *eine Kombination*

verschiedener Massstäbe. Der Unterschiede zwischen einem polytomen Massstab und einer Kombination zeigt sich in der Art und Weise, wie die Ordinalität der Skalen aufgebaut ist.

Die Stufung polytomer Massstäbe bildet eine qualitative oder quantitative Zunahme einer Qualitätskategorie ab (R. W. Fairbrother, 1988, 51f). Eine natürliche Ordinalität liegt vor, wenn die verwandten Kriterien logisch geordnet sind. Dies gilt zum Beispiel für die Kodierung der Messgenauigkeit mit Hilfe der Kriterien (K1) und (K2) (siehe Kodierschema in Tab. 4.9). Die Codes hängen logisch voneinander ab. Alle Antworten, die das Kriterium (K1) erfüllen, genügen notwendigerweise auch dem Kriterium (K2). Die Ordinalität des Kodierschemas ist daher logisch gegeben. Liegt zwischen den Massstäben keine logische Ordnung vor, wird durch das kombinierte Kodierschema a priori eine künstliche Ordnung festgelegt.

Die Stufung kombinierter Massstäbe entsteht durch die Art, wie die Kriterien verknüpft werden. Die Ordinalität baut dabei auf dem Prinzip auf, dass eine bessere Antwort mehr oder anspruchsvolleren Qualitätskriterien genügt. Da Kriterien zu unterschiedlichen Qualitätskategorien logisch unabhängig voneinander sind, ist die Ordinalität kombinierter Massstäbe stets künstlich. Im Folgenden werden vier Verknüpfungsarten vorgestellt und verglichen.

Code 2	Die Messergebnisse weichen höchstens 10% von der Testmessung ab (K1).
Code 1	Die Messergebnisse weichen höchstens 20% von der Testmessung ab (K2).
Code 0	Die Messergebnisse weichen mehr als 20% von der Testmessung ab (nicht K2).

Tabelle 4.9 – Beispiel eines polytomen Massstabes für die Präzision von Messergebnissen: Ergebnisorientierte, evidenzbasierte Kodierung mit natürlicher Ordinalität

Gleichwertige Verknüpfung von Kriterien. In einer von Hammann et al. (2008, 68f) beschriebenen Papier-und-Bleistift-Aufgabe werden die Testpersonen aufgefordert, eine Untersuchung zu planen, um den Einfluss von Körperaktivität und Tageszeit auf den Blutpuls festzustellen. Der vierwertige Massstab verknüpft folgende drei Kriterien, mit welchen die heuristische Qualität der Planung beurteilt wird.

(ExpK) *Experimentelle Kontrolle*: Beide Testvariablen Körperaktivität und Tageszeit werden einer experimentellen Kontrolle unterzogen, d. h. es werden Experimente vorgeschlagen, bei welchen diese Variablen variiert werden.

(KonA) *Kontrollansatz*: Bei beiden Testvariablen wird ein Kontrollansatz vorgeschlagen. Während die eine Variable variiert wird, wird die andere Variable kontrolliert.

(Verg) *Vergleich*: Bei beiden Testvariablen wird ein Vergleich von zwei (oder mehreren) Experimenten (Messungen) gemacht.

Im verwandten Kodierschema (siehe Tab. 4.10) werden die drei Kriterien gleichwertig zu einem polytomen Massstab verknüpft. Eine natürliche Ordnung zwischen den drei Kriterien (ExpK), (KonA) und (Verg) ist nicht gegeben. Im Einzelfall lässt die Frage, welches der drei Kriterien nun nicht erfüllt ist, mehrere Antworten zu. Die gleichwertige Behandlung der drei Kriterien im Kodierschema ist daher nicht nur sinnvoll, sondern vor allem kodiertechnisch notwendig.

Code 3	Alle drei Kriterien (ExpK), (KonA) und (Verg) erfüllt.
Code 2	Zwei von drei Kriterien (ExpK), (KonA) und (Verg) erfüllt.
Code 1	Eines von drei Kriterien (ExpK), (KonA) und (Verg) erfüllt.
Code 0	Keines der drei Kriterien (ExpK), (KonA) und (Verg) erfüllt.

Tabelle 4.10 – Beispiel eines polytomen Massstabes für die heuristische Korrektheit einer Planung einer Untersuchung mit gleichwertiger Verknüpfung der Kriterien (Hammann et al., 2008, 68).

Hierarchische Verknüpfung von Kriterien. Ein Beispiel einer hierarchischen Verknüpfung zweier Massstäbe stellt das Kodierschema des dritten Items der HarmoS-Aufgabe ⟨Balkenwaage⟩ dar (cf. App. A.1, S. 221). Mit dem Item ⟨N9E23i03⟩ wird bewertet,

(SchK) ob aus den Evidenzen in Bezug auf die Hypothese der logisch korrekte Schluss gezogen wird (*Schlusskorrektheit*) und

(TheK) ob das Resultat theoretisch korrekt ist (*Theoriekorrektheit*).

Die zwei Kriterien (SchK) und (TheK) werden – wie in Tabelle 4.11 beschrieben – hierarchisch zu einem kombinierten Massstab verknüpft. Der Vorteil der hierarchischen Verknüpfung von Kriterien gegnüber der gleichwertigen Verknüpfung ist, dass jeder Code nur eine Interpretation zulässt und daher weitergehende Analysen ermöglicht. Der Nachteil ist, dass künstlich eine Hierarchie zwischen den Massstäben geschaffen wird. In diesem Beispiel ist eine theoretisch korrekte, aber aus den Daten falsch hergeleitete Antwort gegenüber einer theoretisch falschen, aber aus den Daten korrekt abgeleitete Antwort minderwertig.

Semi-hierarchische Verknüpfung von Kriterien. Werden mehr als zwei Kriterien oder polytome Massstäbe kombiniert, sind hierarchische Verknüpfungen möglich, bei denen bestimmte Codes mehrere Interpretationen haben. Als Beispiel dient das Kodierschema

Code 2	Antwort genügt (SchK) und (TheK).
Code 1	Antwort genügt (SchK).
Code 0	Antwort genügt weder (SchK) noch (TheK).

Tabelle 4.11 – Balkenwaage-Aufgabe: Beispiel einer hierarchischen Verknüpfung zweier dichotomer Massstäbe.

zur Aufgabe ⟨Paper Towel⟩, die von Brown und Shavelson (1996, 26ff) ausführlich beschrieben wird. Bei dieser Experimentieraufgabe müssen Schülerinnen und Schüler von drei verschiedenen Haushaltpapieren dasjenige herausfinden, das die grösste, und dasjenige, das die kleinste Saugfähigkeit besitzt. Kodiert werden folgende fünf Kriterien.

(MetA) *Adäquatheit der Methode*: Die verwandte Experimentiermethode eignet sich, um die Fragestellung zu beantworten. Kodiert wird: A) Das Kriterium ist erfüllt, B) das Kriterium ist nicht erfüllt.

(MesS) *Adäquatheit der Messmethode*: Bei den Messungen wird darauf geachtet, dass das Testpapier mit Wasser saturiert ist. Kodiert wird: A) Die Saturierung der Papiertücher wird beachtet, B) bei allen Papiertüchern wird dieselbe Wassermenge verwendet, C) die Saturierung der Papiertücher wird nicht beachtet.

(MesA) *Adäquatheit der Messmethode*: Die verwandte Methode, Wasser abzumessen, ergibt korrekte Ergebnisse und lässt einen adäquaten Vergleich zu. Kodiert wird: A) das Kriterium ist erfüllt, B) das Kriterium ist nicht erfüllt.

(MesP) *Sorgfalt/Präzision der Messung*: Die handwerklichen Handlungen werden mit Sorgfalt durchgeführt. Kodiert wird: A) das Kriterium ist erfüllt, B) das Kriterium ist nicht erfüllt.

(EviK) *Evidenzkorrektheit*: Das Resultat ist korrekt, d. h. es entspricht den Testmessungen. Kodiert wird: A) das saugfähigste und das am wenigsten saugfähige Papier wird korrekt identifiziert, B) mindestens das saugfähigste oder das am wenigsten saugfähige Papier wird korrekt identifiziert.

Aus dem von Brown und Shavelson (1996, 43) vorgeschlagenen Kodierschema (siehe Tab. 4.12) wird zwar teilweise eine Hierarchie unter den fünf Kriterien ersichtlich, zu den Codes 1, 2 und 3 gibt es jedoch mehrere Interpretationen. Die dargestellte Kodierung wird unter dem Begriff des Procedure-based scoring bei Baxter et al. (1992) ausführlich beschrieben und evaluiert.

	(MetA)	(MesS)	(MesA)	(MesP)	(EviK)
Code 5	A	A	A	A	A
Code 4	A	A	A	B	B
Code 3	A	B	A	A oder B	B
Code 2	A	C *oder*	B	A oder B	B
Code 1	B *oder*	C	irrelevant	A oder B	B

Tabelle 4.12 – ⟨Paper towel⟩-Aufgabe: Beispiel einer semi-hierarchischen Verknüpfung von fünf Massstäben.

Gewichtete Verknüpfung von Kriterien. Bei der gewichteten Verknüpfung werden Lösungen von Experimentieraufgaben wie eine Klassenarbeit bewertet: Für jedes erfüllte Kriterium werden Punkte vergeben, die am Schluss zusammengezählt und mit Hilfe einer Notenskala zu einer Gesamtnote bzw. zu einem Code "umgerechnet" werden. Diese Methode erlaubt, wesentliche Kriterien stärker zu gewichten als unwichtige Kriterien. Und sie lässt sich sowohl auf polynome Massstäbe als auf Kombinationen von Massstäben anwenden. Diese Art der Kodierung ist beliebt, weil sich mit ihr auf ökonomische und reliable Weise aus sehr vielen Kriterien ein Gesamtscore ermitteln lässt. Der Nachteil dieser Methode ist, dass ein Code sehr viele Interpretationen zulässt und ein Rückschluss auf einzelne Kriterien unmöglich ist. Zudem unterliegt die Gewichtung einer gewissen Beliebigkeit. Es stellt sich daher auch die Frage, welche Verknüpfung und welche Gewichtung in einem konkreten Fall reliable und valide Ergebnisse gewährleisten.

Dieser Frage geht Solano-Flores (1994) in seiner Dissertation nach. Anhand der oben diskutierten ⟨Paper towel⟩-Aufgabe vergleicht er verschiedene Arten, Kriterien zu gewichten und zu einem Gesamtscore zusammenzuzählen, und stellt diese dem so genannten p Procedure-based scoring (siehe Tab. 4.12) gegenüber. Seine Resultate zeigen, dass alle Gewichtungen im Wesentlichen zu denselben hohen Verallgemeinerungskoeffizienten führen. Verschiedene Gewichtungen produzieren jedoch unterschiedliche Score-Verteilungen. Die Verwendung von Schüsselkriterien, von deren Beurteilung die Art der Verknüpfung der anderen Kriterien abhängig gemacht wird – d. h. wie und ob andere Kriterien berücksichtigt werden, ergibt signifikante Unterschiede z. B. in Bezug auf die Sensitivität der Testergebnisse auf vorgängige Instruktion im Unterricht.

A priori-Ordinalität von Kodiermassstäben. Wenn in einem Test nicht nur Richtig und Falsch gewertet, sondern auch Stufungen der Korrektheit oder Güte einer Antwort unterschieden werden, besteht die Gefahr, dass a priori "falsche" Progressionen geschaffen werden. Tests bilden stets die Ordinalität der zugrundeliegenden Kodiersysteme ab: Was als anspruchsvollere Antwort kodiert wird, wird auch als Resultat anspruchsvoller her-

auskommen. Problematisch sind daher Schwierigkeitsvergleiche, die auf künstlichen Progressionen abstützen. Dazu gehören einerseits Vergleiche von Massstäben, die in einem Kodiersystem kombiniert werden. Die in einem Kodiersystem gewählte Hierarchie der Massstäbe ist a priori und muss immer kritisch hinterfragt werden. Andererseits gaukeln polytome Massstäbe zuweilen eine Progression vor, die logischerweise nicht gilt. Dies vor allem dann, wenn eine falsche Antwort besser eingestuft wird als eine andere falsche Antwort. Die Graduierung der Falschheit von Antworten erfolgt meist willkürlich.

Bei der Analyse der Itemschwierigkeit im HarmoS-Experimentiertest (siehe Kapitel 7) werden wir die in den Kodiersystemen verwandten Massstäbe als schwierigkeitsrelevante Itemfaktoren miteinbeziehen. Sollten sich signifikante Unterschiede der Wirkungen einzelner Massstäbe ergeben, müsste mit einer zusätzlichen Analyse der Kodiersysteme geklärt werden, inwieweit die Kodierung diese Effekte erzeugt.

4.3.3 Inhaltsproblem der Messung experimenteller Kompetenz

Inhaltswissen versus Kompetenz. Das für den Standarddiskurs zentrale Motiv der Kompetenzorientierung in der Unterrichtspraxis wird in der öffentlichen Meinung gemeinhin mit der Abwertung des Fachwissens gleichgesetzt. Im entsprechenden wissenschaftlichen Assessmentdiskurs wird jedoch deutlich gemacht, dass experimentelle Kompetenzen nur an konkreten fachlichen Inhalten, sprich in einem fachlichen Kontext, getestet werden können (Hodson, 1992; Millar & Driver, 1987). Insofern ist nicht zu befürchten, dass mit der Ausrichtung auf prozessurale Kompetenzen Fachinhalte im Unterricht gänzlich in den Hintergrund träten. Trotzdem wird im Modelldiskurs erwartet, dass sich Kompetenzbeurteilungen von Fachwissenstests unterscheiden. Die äussere Abgrenzung der experimentellen Kompetenz ist sodann auch eine wesentliche Anforderung, die an ein valides Kompetenzmodell gestellt wird. Unklar bleibt nun, inwieweit fachliche Inhalte bei der Ausübung experimenteller Kompetenzen tatsächlich eine Rolle spielen bzw. inwieweit mit der Messung experimenteller Kompetenz Fachwissen getestet wird.

In der Fachliteratur werden verschiedene Aspekte der Abhängigkeit der Kompetenzbeherrschung vom Inhaltswissen beschrieben. Anhand unseres Progressionsmodells (Abschnitt 3.1.1) lassen sich diese Abhängigkeiten wie folgt zusammenfassen:

- *Theorieabhängigkeit der Problemkomplexität [P]*: Ein experimentelles Problem ist stets in einen Kontext eingebettet, dessen Inhalte die Komplexität des Problems beeinflussen.

- *Theorieabhängigkeit der Prozessqualiät [Q]*: Das Verständnis des in einer Aufgabe thematisierten Kontextes setzt ein gewisses Vor- und Hintergrundwissen voraus. Die Qualität, mit der eine Aufgabe gelöst wird, hängt deshalb von der Qualität dieses Wissens

ab.

- *Theorieabhängigkeit des Transferumfangs [T]*: Der Transfer einer Kompetenz auf verschiedene Kontexte hängt vom verfügbaren Inhaltswissen ab.

Nachfolgend eine detaillierte Diskussion der drei Theorieabhängigkeiten.

Theorieabhängigkeit der Problemkomplexität. Die Abhängigkeit von den Theorien und Konzepten der Testperson betrifft verschiedene Teilaspekte der Problemkomplexität, die im Kapitel 4.3.1 diskutiert wurden.

- *Abstraktheit der Kontexte*: Die Einschätzung der Abstraktheit und Fremdheit eines Aufgabenkontextes hängt massgeblich von den theoretischen und praktischen Vorkenntnissen der Testperson ab.

- *Bekanntheitsgrad der Mittel*: Experimentiermaterial und Instrumente können nur dann zielgerichtet eingesetzt werden, wenn die Funktionen der Objekte bekannt sind. Das Verständnis der Funktionen bedingt die Kenntnisse von Konzepten (e. g. die Grösse, die mit einem bestimmten Messinstrument gemessen wird) oder Theorien (e. g. um die Wirkungsweise einer experimentellen Anordnung zu verstehen). Die Performance bei der Planung von Manipulationsstrategien oder bei der Entwicklung von Messverfahren hängt wesentlich von den genannten Kenntnissen ab.

- *Komplexität und Vernetztheit*: Die Anzahl relevanter Faktoren und deren Wechselwirkungen hängt von der Theorie ab, die dem Problem zugrunde gelegt wird. Mangelhaftes theoretisches Vorwissen führt oft dazu, dass komplexe Aufgaben auf einfache Weise zu lösen versucht werden. D. h. Laien sehen oft die Schwierigkeit einer Aufgabe nicht, weil sie die komplexen Zusammenhänge nicht kennen.

Theorieabhängigkeit der Prozessqualität. Die Qualität, mit welcher experimentelle Prozesse ausgeführt werden, hängt wesentlich von den Theorien und den Konzepten der Testperson ab. Diese Abhängigkeit der Prozessqualität muss bei der Verwendung von entsprechenden Kodiermassstäben berücksichtigt werden.

- *Prozess des Beobachtens*: Millar und Driver (1987) verweisen darauf, dass sich der Prozess des wissenschaftlichen Beobachtens von alltäglichen Beobachtungsprozessen vor allem durch die involvierten Konzepte unterscheidet. Die Qualität einer Beobachtung hängt daher auch von der Korrektheit der verwandten Konzepte ab. “Children’s ability to observe involves their learning of a conceptual framework which identifies the elements of a complex situation which are scientifically ’worth observing’. Only when the framework has been grasped is the observation possible. In fact, what we are doing in

those science lessons where accurate observation plays a major role is not to develop *observation*, but rather to train pupils in *scientific observation*, i.e. to help them to come to see the situation in the way that scientists find it useful and productive to see it“ (Millar & Driver, 1987, 43).

- *Prozess des Klassifizierens*: Was fürs Beobachten gilt, trifft auch auf den Prozess des Klassifizierens zu. Ob eine Klassifikation wissenschaftlich ist oder nicht, hängt vor allem von den verwandten Konzepten ab. Millar und Driver (1987, 43f) bringen diesen Zusammenhang mit folgendem Beispiel auf den Punkt: “[...] pupils' difficulties in recognising the appropriate classifications for certain plants or animals [...] reflect not a failure of the 'process' of classifying, but an incomplete understanding, or appreciation of the usefulness of that particular classification system which is used by scientists in these cases.“

- *Prozess des Formulierens von Hypothesen*: Auch das Aufstellen von Vermutungen muss wie das Beobachten und Klassifizieren im eigentlich Sinne nicht gelehrt werden: Diese Prozesse tun wir alle ständig. Wissenschaftliche Hypothesen unterscheiden sich von alltäglichen Hypothesen einfach darin, dass in Ersteren wissenschaftliche Konzepte und in Letzteren Alltagskonzepte vorkommen (Millar & Driver, 1987).

- *Planungen und Untersuchungen*: Wie bereits im Abschnitt 3.2.3.1 aufgezeigt wurde, hängt die Fehlerhaftigkeit einer Experimentierstrategie von der Theorie ab, von der eine Testperson in einer konkreten Planungs- und Experimentieraufgabe ausgeht. Dies gilt beispielsweise für die Untersuchung von kausalen Zusammenhängen. Die Frage, ob eine Variable in einem Experiment kontrolliert werden soll, hängt davon ab, ob eine kausale Wirkung dieser Variablen vermutet wird (Koslowski, 1996). Ganz allgemein spielt die Vertrautheit mit einem Kontext für die Performance bei experimentellen Problemlöseaufgaben eine massgebliche Rolle (Millar & Driver, 1987, 48f). Eine mangelhafte Experimentierstrategie kann daher Ausdruck von falschem oder fehlendem theoretischen und konzeptuellen Wissen sein. Diese Ambiguität sollte bei der Verwendung heuristischer Kodiermassstäbe bedacht werden.

- *Datenanalyse*: Ebenfalls beschrieben wurde bereits die Abhängigkeit der Datenanalyse von theoretischem Vorwissen (Abschnitt 3.2.3.1). Die Analyse eines Faktors, der im Zusammenspiel mit anderen Faktoren kausal für einen Effekt sein könnte, fällt anders aus je nachdem, ob der Faktor nur als notwendig oder sogar als hinreichend für den untersuchten Effekt angenommen wird (Ehmer & Hammann, 2007). Dieser Einfluss der zugrundeliegenden kausalen Modelle sollte bei der Beurteilung der Schlusskorrektheit berücksichtigt werden.

Theorieabhängigkeit des Transferumfangs. So wie die Qualität von Prozessen wie Planen, Untersuchen, Beobachten, Klassifizieren oder Analysieren von Daten zu einem guten Teil vom verfügbaren Theorie- und Konzeptwissen abhängt, so bestimmen auch die Vertrautheit eines Kontextes und das dazugehörige Fachwissen den Transfer dieser Prozesse (Hodson, 1992, 124f). Für Messick (1994, 19) stellt die Abhängigkeit des Kompetenztransfers vom Inhaltswissen Testentwickler vor ein Trilemma: Entweder werden Kompetenzen kontextabhängig definiert und entsprechend gemessen[21]. Oder es wird versucht, den Einfluss von Inhalten zu minimieren. In diesem Fall können entweder möglichst "kontextarme" oder besonders kontextübergreifende Tests entwickelt werden. Ersteres gelingt nur, wenn auf authentische Testaufgaben wie beispielsweise Experimentieraufgaben verzichtet wird. Letztere Variante basiert auf der Idee, die Unterschiede beim Vor- und Hintergrundwissen der Testpersonen durch die Vielfalt von Aufgabenkontexten auszugleichen.

4.4 Modellierung kompetenzirrelevanter Itemschwierigkeit

Eingrenzung des Begriffs der experimentellen Kompetenz. Misst man die experimentelle Kompetenz, ist man stets mit dem Problem konfrontiert, dass die Messung durch kompetenzirrelevante Schwierigkeiten der Testaufgaben verfälscht werden kann. Dies gilt ganz allgemein für jede Test- und Messart, im besonderen Masse trifft dies aber auf Large-scale Assessments zu, deren Kodierung auf Eigenrapporten basiert. Welche Schwierigkeiten als kompetenzirrelevant gelten, hängt aber letztlich von der konkreten Kompetenzdefinition ab. Üblicherweise wird das Erfassen des Auftrags, sprich das Text- und das Hörverständnis, nicht zur experimentellen Kompetenz gezählt. Beim Antwortgeben verhält sich dies anders. Je nach Auffassung wird die Fähigkeit, Messresultate in angemessener Form darzustellen, als Teilkompetenz zur experimentellen Kompetenz dazu gezählt.[22] Im Folgenden werden wir den Kompetenzbegriff derart eng fassen, dass das Erfassen der Aufgabe und das Geben der Antwort nicht Teil der experimentellen Kompetenz sind.

Die nachfolgend vorgestellten Kategorien von Aufgabenmerkmalen und Ergebnissen basieren allesamt auf Analysen von deutschsprachigen Papier-und-Bleistift-Aufgaben der TIMS-Studie III oder von einer der drei PISA-Staffeln 2000, 2003 und 2006.

[21] Diesen Weg geht man derzeit bei der Formulierung des Deutschschweizer Lehrplans.

[22] In den Schweizer Grundkompetenzen für den Fähigkeitsbereich «Fragen und untersuchen» ist die Fähigkeit, Ergebnisse in geeigneter Form darzustellen, explizit enthalten (EDK, 2011). Dementsprechend wurde die Darstellung der Ergebnisse beim HarmoS-Experimentiertest teilweise mit kodiert. Auch bei den TIMS-Studien wurde die Darstellungsqualität erfasst. Im Gegensatz zu HarmoS wurden die Codes jedoch separat erfasst und nicht mit der Beurteilung der Experimentierfähigkeit vermengt (TIMSS, 1994).

Itemmerkmale der Problemstruktur und der Aufgabenstellung Prenzel, Häußler, Rost und Senkbeil (2002) unterscheiden drei Arten von Aufgabenmerkmalen:

1. *Konzeptkategorien* beschreiben "Merkmale der für das Lösen der Aufgabe erforderlichen Wissensbasis".

2. *Prozesskategorien* beschreiben "kognitive Anforderungen beim Lösen der Aufgabe".

3. *Äussere Aufgabenstrukturmerkmale* beziehen sich auf äussere Merkmale der Aufgabenstellung. "Sie bestimmen, wie der 'Input' der Aufgabenstellung und die Art des geforderten 'Outputs' aussieht" (Prenzel et al., 2002, 124ff).

Im Sinne einer möglichst präzisen Klassifikation von Aufgabenmerkmalen wollen wir den drei Merkmalsarten eine vierte hinzufügen:

4. *Innere Aufgabenstrukturmerkmale* sind Merkmale der Aufgabenstellung, die weder eine äussere Eigenschaft darstellen noch den Bezug auf eine erforderliche Wissensbasis oder erforderliche Prozesse herstellen (e. g. die Stimulanz eines Aufgabentextes oder der Zweck einer Information).

Konzept- und Prozesskategorien beschreiben die Struktur des zu lösenden experimentellen Problems. Sie gehören daher zur *Problemstruktur* und nehmen Bezug auf Ursachen für kompetenzrelevante Itemschwierigkeiten. Die äusseren und inneren Itemmerkmale betreffen Strukturmerkmale der meist gedruckten Aufgabenstellung (*Aufgabenstruktur*). Sie beschreiben mögliche kompetenzirrelevante Barrieren beim Aufnehmen und Mitteilen von Informationen. Während jedoch bei Konzept- und Prozesskategorien der Bezug zu einer Wissensbasis bzw. einer Fähigkeit offensichtlich ist, muss dieser Zusammenhang bei inneren und äusseren Aufgabenstrukturmerkmalen mit Hilfe von "Hypothesen über psychologische Prozesse", die das Leistungsverhalten von Schülerinnen und Schülern erklären, konstruiert und begründet werden (TIMSS, 2000b, 102). Zum Beispiel kann die Verwendung des inneren Merkmals der Text-Rekurrenz, i. e. die substantivistische Anknüpfung an den vorangehenden Satz, mit der Theorie von Kintsch und van Dijk (1978) begründet werden (cf. Wellenreuther, 2005, 188ff), wonach ein Text am einfachsten zu verstehen ist, wenn neu zu erfassende Propositionen in einem Text ohne zusätzliche Überbrückungsschlüsse mit früheren, bereits erfassten Propositionen verknüpft werden können (Kulgemeyer & Schecker, 2007, 202). Die Bestimmung von inneren Itemmerkmalen, Konzept- und Prozesskategorien erfordert eine inhaltliche Strukturanalyse der Aufgabe und erfolgt ausschliesslich per Expertenrating. Demgegenüber beziehen sich äussere Itemmerkmale auf das Vorhandensein von bestimmten Elementen in der gedruckten Aufgabenstellung (e. g.

das Vorhandensein von Abbildungen und Grafiken in der Aufgabenstellung). Ihre Bestimmung erfordert daher vor allem eine möglichst objektive Definition der Beurteilungskriterien.

4.4.1 Arbeitsschritt ‹Aufgabe erfassen›

Abgrenzung der Arbeitsschritte ‹Aufgabe erfassen› und ‹Problem lösen›. Eine Aufgabe zu erfassen und sie dann zu lösen, sind nicht unabhängige Arbeitsschritte. Dasselbe Fachwissen, das benötigt wird, um eine Aufgabe zu verstehen, ist auch notwendig, um sie zu lösen. Vordringlich ist daher zunächst eine klare Abgrenzung der zwei Arbeitsschritte.

Die Unterscheidung der zwei Arbeitsschritte im Rahmen einer Kompetenzmessung ist nur dann gerechtfertigt, wenn die zwei Arbeitsschritte unterschiedliche Fähigkeiten bzw. unterschiedliches Wissen ansprechen. Für H. E. Fischer et al. (2003, 194) trifft dies auf die *Lesekompetenz* zu, indem mit einer nicht zufriedenstellenden Lesekompetenz "die erste Voraussetzung für ein erfolgreiches Bearbeiten naturwissenschaftlicher Aufgaben nicht gegeben" ist, wie die bei PISA nachgewiesene hohe Korrelation zwischen Lesekompetenz und den Leistungen im Bereich der naturwissenschaftlichen Grundbildung nahelegt. "Es ist somit unerlässlich, im Zusammenhang mit naturwissenschaftlichen Aufgaben auch Lesekompetenz zu berücksichtigen" (H. E. Fischer & Draxler, 2007, 649).

Arbeitsschrittspezifische Itemmerkmale. Aus einem Vergleich verschiedener Untersuchungen von Itemmerkmalen mit TIMSS- und PISA-Aufgaben (TIMSS, 2000a; Kulgemeyer & Schecker, 2007; Kulgemeyer, 2009; Prenzel et al., 2002) ergeben sich für uns drei Merkmalsbereiche, die wir ausschliesslich für den Arbeitsschritt ‹Aufgaben erfassen› als relevant erachten.

- *Begriffswissen*: Kenntnis von Definitionen (TIMSS, 2000a); terminologisches Wissen (Prenzel et al., 2002). Eine Untersuchung von Schmiemann (2011) im Rahmen des Projekts "Biologie im Kontext" zeigt auf, dass die Repräsentation von Begriffen (wissenschaftliche Begriffe versus Alltagsbegriffe) die Schwierigkeit von Aufgaben beeinflusst.

- *Textverständnis*: Erfassen von Textinformationen (TIMSS, 2000a) mit unterschiedlicher Länge (Prenzel et al., 2002), Repräsentationsform (kontinuierliche / diskontinuierliche Texte) und unterschiedlich ausgeprägten Textbarrieren (Gliederung, Prägnanz(Satzbau), Stimulanz und Kohärenz) (Kulgemeyer, 2009).

- *Verständnis von Symbolsprachen.* Verständnis von formalisierten Gesetzen, symbolischen Zeichnungen und Diagrammen (TIMSS, 2000a); Vorhandensein von grafischen, bildlichen und numerischen Inputs in der Aufgabenstellung (Prenzel et al., 2002).

Empirische Resultate. Die hier wiedergegebenen Ergebnisse zu schwierigkeitserzeugenden Itemmerkmalen beziehen sich ausschliesslich auf Papier-und-Bleistift-Aufgaben. Prenzel et al. (2002) haben die Aufgaben der PISA-Staffel 2000 analysiert und einen Katalog von Merkmalen eruiert, mit denen man die Aufgabenschwierigkeit vorhersagen kann. In Bezug auf den Arbeitsschritt ‹Aufgabe erfassen› hat die Analyse folgende signifikanten Ergebnisse ergeben.

i. Terminologisches Wissen: Ist terminologisches Wissen erforderlich, ist die Aufgabe schwieriger.

ii. Textlänge: Die Länge des Aufgabentextes beeinflusst die Itemschwierigkeit nicht.

iii. Grafische und bildliche Inputs: Grafiken und Informationen in Bildform machen eine Aufgabe leichter.

4.4.2 Arbeitsschritt ‹Antwort geben›

Abgrenzung der Arbeitsschritte ‹Antwort geben› und ‹Problem lösen› Bei einem Testitem gilt es zwischen zwei Aufgaben zu unterscheiden: die Aufgabe, das formulierte experimentelle Problem zu lösen, und die Aufgabe, die richtige Antwort zu geben. Da bei large-scale Assessments die Kodierung meist auf Eigenrapporten basiert, nennen wir das ‹Problem lösen› die *intendierte Aufgabe* und das ‹Antwort geben› die *kodierte Aufgabe*. Natürlich wird bei jeder Testkonstruktion das Ziel verfolgt, die intendierte und die kodierte Aufgabe optimal in Übereinstimmung zu bringen. Dies kann jedoch prinzipiell nie vollständig gelingen. Beim ‹Antwort geben› können Fähigkeiten und Wissen ins Spiel kommen, die beim ‹Problem lösen› keine Rolle spielen. Wenn zum Beispiel eine Testperson die Antwort auf eine Testfrage nicht weiss, werden bei geschlossenen Antwortformaten (Multiple choice- und Multiple select-Formate) bestimmte Lösungsstrategien wie das Raten oder das intelligente Ausschliessen von Lösungsalternativen wichtig (TIMSS, 2000b, 102). Im Gegenzug erfordern offene Antwortformate Schreibkompetenz und zusätzliches Wissen, da bei ausführlichen Antworten meist nicht nur das Ergebnis, sondern auch der Lösungsweg beschrieben und begründet werden muss (TIMSS, 2000b, 102f).

Arbeitsschrittspezifische Itemmerkmale. Die in der Literatur diskutierten Itemmerkmale zum Arbeitsschritt ‹Antwort geben› beziehen sich im Wesentlichen auf zwei Aspekte einer Antwort: der Repräsentationsform des Antwortinhalts und dem “Antwortformat“ gemäss Definition von TIMSS (2000a, 102ff).[23]

[23]Die Begriffe werden in der Literatur uneinheitlich verwendet. Was TIMSS (2000a) und H. E. Fischer und Draxler (2007) als “Antwortformat“ bezeichnen, wird bei Kulgemeyer (2009) als “Aufgabenformat“

- *Repräsentationsform*: Text, grafischer, numerischer Output erfordert (Prenzel et al., 2002)
- *"Antwortformat"*: geschlossene und offene Formate (Multiple choice-, Multiple select-, Kurzsatz- und Langsatzantwort sowie Aufsatz bzw. ausführliche oder erweiterte Antworten) (H. E. Fischer & Draxler, 2007; Kulgemeyer, 2009; Prenzel et al., 2002; TIMSS, 2000a)

Empirische Resultate. Die folgenden Ergebnisse beziehen sich auf zwei Untersuchungen mit Papier-und-Bleistift-Aufgaben aus dem Bereich der Naturwissenschaften.

iv. Offenheit des Antwortformats: Testaufgaben mit offenen Antwortformaten (Kurzsatz- wie auch Langsatzantworten) waren sowohl bei der TIMS-Studie als auch bei der PISA-Staffel 2000 signifikant schwieriger als Aufgaben mit geschlossenem Format (Multiple choice oder Multiple select) (Prenzel et al., 2002; TIMSS, 2000a).

v. Trennschärfe von Antwortformaten: TIMSS-Aufgaben mit offenem Antwortformat (Kurzsatzantworten und ausführliche Antworten) wiesen eine signifikant höhere Trennschärfe (im Sinne der klassischen Testtheorie) auf als Aufgaben mit geschlossenem Antwortformat (Multiple choice).

vi. Arbeitsschrittspezifische Anforderungen: TIMSS (2000b) wiesen für Aufgaben der TIMS-Studie einen substantiellen Einfluss von arbeitsschrittspezifischen Anforderungen des Arbeitsschritts ‹Antwort geben› auf die Itemschwierigkeit nach. Bei Multiple-choice-Aufgaben bzw. Aufgaben mit ausführlichen Antworten werden 7 bis 19 Prozent der Varianz durch separate Methodenfaktoren erklärt. Trotz dieses Befunds schliessen TIMSS (2000b, 106): "Insgesamt dominieren jedoch Effekte des sich in allen Testaufgaben abbildenden generellen Fähigkeitsfaktors der mathematisch-naturwissenschaftlichen Grundbildung. Dies rechtfertigt die Verwendung eines Gesamtwertes unter Vernachlässigung des Antwortformates."

behandelt. Wie in Kapitel 7 erläutert wird, machen wir die Unterscheidung von *Lücken-* und *Füllformaten*. Lückenformate entsprechen mehr oder weniger den Aufgabenformaten, wie sie TIMSS (2000a) benutzt. Mit den Füllformaten wird der Aspekt der Form aufgenommen, in welcher der Antwortinhalt präsentiert werden soll. Den Begriff "Antwortformat" werden wir später als Oberbegriff benutzen. Er umfasst Lücken- und Füllformate.

Teil II

Analyse eines large-scale Experimentiertests

Kapitel 5

Der HarmoS-Experimentiertest

5.1 HarmoS-Kompetenzmodell

Projekt HarmoS Naturwissenschaften. Im Jahr 2004 erteilte die Schweizerische Konferenz der kantonalen Erziehungsdirektoren (EDK) dem Konsortium HarmoS Naturwissenschaften den Auftrag, gestützt auf ein zu validierendes Kompetenzmodell Basisstandards in den Naturwissenschaften für das Ende des 2., 6. und 9. Schuljahres zu formulieren. Im Rahmen der Validierung des HarmoS-Kompetenzmodells wurden zwischen April 2007 und Juni 2008 in der deutsch- und französischsprachigen Schweiz auf allen drei Schulstufen sowohl Papier-und-Bleistifttests als auch Experimentiertests durchgeführt. Auf der Grundlage der Testergebnisse wurden im Oktober 2008 vom Konsortium HarmoS Naturwissenschaften der EDK Vorschläge für Basisstandards unterbreitet (HarmoS, 2008). Die Ergebnisse wurden zudem verwendet, um das HarmoS-Kompetenzmodell zu validieren (Labudde, Metzger & Gut, 2009; Ramseier, Labudde & Adamina, 2011).

Kompetenzmodell. Im Kompetenzmodell von HarmoS werden die Inhalte (Themenbereiche) und die Fähigkeiten (Handlungsaspekte) als zwei separate Dimensionen behandelt. Auf der Inhaltsachse werden fünf fächerübergreifend formulierte Themenbereiche und auf der Fähigkeitsachse sechs Handlungsaspekte unterschieden. Wie in der Abbildung 5.1 dargestellt, resultieren aus der Überschneidung der zwei Achsen 30 Zellen, die bei der Entwicklung der Validierungstests bestmöglich durch geeignete Testaufgaben zu füllen waren. Die Entwicklung der Niveaus für die drei Zyklen erfolgt bei HarmoS ausschliesslich auf der Achse der Handlungsaspekte.

Validierung des Kompetenzmodells. Das Kompetenzmodell wurde in zwei Schritten validiert. Zum einen wurden die nicht experimentellen Handlungsaspekte «Informationen erschliessen», «Ordnen, strukturieren, modellieren» und «Einschätzen und beurteilen»

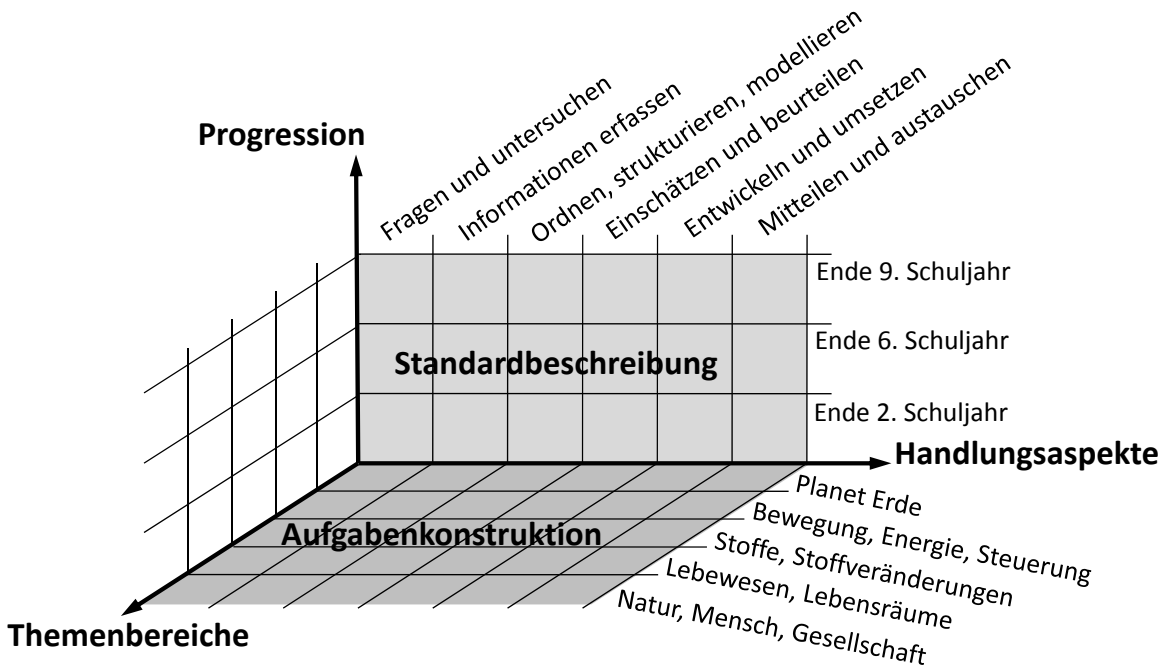


Abbildung 5.1 – HarmoS-Kompetenzmodell

anhand eines Papier-und-Bleistifttests mit rund 8000 Schülerinnen und Schülern des 6. und 9. Schuljahres überprüft. Die Validierung ergab, dass der Test mehrere Dimensionen unterscheidet, dabei jedoch stärker die Themenbereiche als die Handlungsaspekte differenziert (Ramseier et al., 2011, 18ff).

Zum anderen wurde parallel zum HarmoS-Experimentiertest ein reduzierter Papier-und-Bleistifttest mit 35 Aufgaben zu den oben validierten Handlungsaspekten durchgeführt. Der Vergleich der beiden Tests ergab, dass die drei nicht experimentellen Handlungsaspekte untereinander stärker korrelieren als mit dem experimentellen Handlungsaspekt «Fragen und untersuchen» (Labudde, Metzger & Gut, 2009, 315ff). Die Korrelationen sind jedoch bei beiden untersuchten Schulstufen auf hohem Niveau (0.70-0.74) und übersteigen die Reliabilität der gemessenen Dimensionen (0.66-0.71).

5.2 Experimentiertest: Entwicklung, Durchführung und Auswertung

Personenstichprobe. Zwischen April und Juni 2008 wurde mit insgesamt 78 Klassen des 6. und 9. Schuljahres in der deutsch- und französischsprachigen Schweiz ein stufenübergreifender Experimentiertest durchgeführt. Darüber hinaus wurde der Test auch in der italienischen Schweiz bei einer Stichprobe von 6 Klassen des 9. Schuljahres eingesetzt.

	D-CH		F-CH			I-CH	
Kantone/Regionen	**BE/SO**	**ZH/SH/LU/ZG**	**GE**	**NE**	**VD**	**TI**	**CH**
6. Schuljahr	250	149	136	134	0	0	**669**
9. Schuljahr	261	147	0	0	286	128	**822**
beide Schulstufen	511	296	136	134	286	128	**1491**
		807		556			
			1363				

Tabelle 5.1 – Personenstichprobe des HarmoS-Experimentiertests ⟨E08|69dfi⟩. Abkürzungen der Sprachregionen: D-CH = deutschsprachige Schweiz, F-CH = französischsprachige Schweiz, I-CH = italienischsprachige Schweiz; Abkürzungen der Kantone: BE = Bern, SO = Solothurn, ZH = Zürich, SH = Schaffhausen, LU = Luzern, ZG = Zug, GE = Genf, NE = Neuenburg, VD = Waadt, TI = Tessin.

Repräsentativität. Die Stichprobe ist weder hinsichtlich der Schulstufen noch hinsichtlich der Sprachregionen repräsentativ. Die Lehrpersonen der getesteten Klassen wurden aufgrund persönlicher Bekanntschaft direkt vom Konsortium für eine Teilnahme angefragt. Auf diese Weise haben eher Lehrpersonen zugesagt, die für das Thema Experimentieren sensibilisiert sind und bereits im eigenen Unterricht viel experimentieren. Die Stichprobe wird daher eher kompetenter sein als der Durchschnitt in den entsprechenden Schulstufen. Zudem hat man sich bei der Klassenauswahl auf Schulstandorte beschränkt, die im Einzugsgebiet der beteiligten Hochschulen liegen, von wo aus die regionalen Tests organisiert und durchgeführt wurden. So besteht beispielsweise die französischsprachige Stichprobe des 9. Schuljahres ausschliesslich aus Schülerinnen und Schüler des Kantons Waadt (cf. Tab. 5.1). Man vergleiche die Ausführungen im Abschnitt 8.1 (S. 163ff).

Aufgabenkonstruktion und Pilotierung. Mit der Testentwicklung wurde im Sinne von Messick (1994, 17ff) ein bezüglich des Unterrichts authentisches Assessment angestrebt, wobei mit den Experimentieraufgaben einerseits ein breites Spektrum an unterrichtsrelevanten Themenbereichen angesprochen (analog zum Experimentiertest von TIMSS, 1997, 7) und andererseits bedeutsame Lernaktivitäten initiiert werden sollten. Aus der Pilotierung, die zwischen November 2007 und Januar 2008 stattfand und auf die im Kapitel 7.1 detaillierter eingegangen wird, wurden 15 Aufgaben für den Haupttest übernommen. Wie in der Tabelle 5.2 ersichtlich, wurden drei Aufgaben nur im 6. Schuljahr, sieben Aufgaben nur im 9. Schuljahr und fünf Aufgaben in beiden Schulstufen eingesetzt. Die 15 Aufgaben umfassen 145 Items, von denen rund zwei Drittel der experimentellen Kompetenz zugeordnet werden. Die restlichen Items verteilten sich auf andere nicht experimentelle

Themenbereiche	6. Schuljahr	6. & 9. Schuljahr	9. Schuljahr
Planet Erde		⟨Steine⟩ (df)	⟨Rapex⟩ (d)
Bewegung, Energie, Steuerung		⟨Taschenlampe⟩ (df) ⟨Balkenwaage⟩ (dfi)	⟨Murmel⟩ (df)
Stoffe, Stoffveränderungen	⟨Schwimmen & Sinken⟩ (d)	⟨Tabletten⟩ (df)	⟨Seife⟩ (dfi) ⟨Öl⟩ (d)
Lebewesen, Lebensräume	⟨Gänseblümchen⟩ (df) ⟨Laubbäume⟩ (d)	⟨Asseln⟩ (dfi)	⟨Wasserpest⟩ (d)
Natur, Mensch, Gesellschaft			⟨Sparlampe⟩ (d) ⟨Solarzellen⟩ (dfi)

Tabelle 5.2 – Aufgaben HarmoS-Experimentiertest ⟨E08|69dfi⟩ (d: Verwendung in der deutschsprachigen Schweiz, f: Verwendung in der französischsprachigen Schweiz, i: Verwendung in der italienischsprachigen Schweiz)

Kompetenzen (cf. App. B).[1]

Haupttest. Der Haupttest fand zwischen April und Mai 2008 statt. Die am Test beteiligten Schulklassen wurden von einem zwei- bis dreiköpfigen Testteam während insgesamt eines Halbtages besucht: Im 6. Schuljahr wurde der Test auf zwei Tage verteilt; im 9. Schuljahr fand der Test an einem Tag statt. Jede Schülerin und jeder Schüler bearbeitete zwei Experimentieraufgaben, jede während jeweils 30 Minuten. Im 6. Schuljahr wurden die Klassen angeleitet: Die Aufgabenstellung wurde von der Testleitung vorgelesen und die Bearbeitungszeit wurde kontrolliert. Die 9. Klassen arbeiteten ohne Anleitung, mit Ausnahme bei der Aufgabe ⟨Steine⟩, die ebenfalls vorgelesen wurde. Auf beiden Schulstufen wurde für der Handhabung anspruchsvoller Instrumente und Geräte (e. g. Mikroskop) bei Bedarf Unterstützung gegeben. Neben den zwei Experimentieraufgaben wurden von den Schülerinnen und Schülern 4 bis 5 Papier-und-Bleistift-Aufgaben gelöst und ein Fragebogen mit Fragen zu Familie und Interessen ausgefüllt.

Kodierung. Die Testbogen wurden mit zwei Teams an der PHBern und der PH FHNW in Basel im Juni und Juli 2008 kodiert.

[1]Neben den 15 Aufgaben wurden im 6. Schuljahr vier weitere Aufgaben mit insgesamt 34 Items eingesetzt, die jedoch explizit keine praktischen Tätigkeiten erfordern. Im Rahmen der Validierung des HarmoS-Kompetenzmodells wurden diese Aufgaben verwendet, um den Handlungsaspekt «Entwickeln und umsetzen» zu evaluieren. Für die Überprüfung des experimentellen Handlungsaspekts «Fragen und untersuchen» spielen diese Aufgaben keine Rolle (vgl. auch Labudde, Metzger & Gut, 2009).

5.3 Itemstichprobe

1. Reduktion. Im Rahmen der Kodierung und ersten Rasch-Auswertung des HarmoS-Experimentiertests wurden verschiedene Items gestrichen: Entweder wurden sie nicht kodiert oder sie wurden in andere Items integriert.[2] Dies betrifft die sechs Items ⟨N9E14i06⟩, ⟨N1E22i08⟩, ⟨N9E43i08⟩, ⟨N9E84i01⟩, ⟨N9E84i04⟩, ⟨N9E84i06⟩. Die Validierung des Kompetenzmodells wurde somit mit insgesamt 173 Items gerechnet.

Erweiterung. Für die Zwecke dieser Arbeit wurden zwei Items bezüglich der Schulstufen gesplittet. Dies betrifft die Items ⟨N1E43i02⟩ und ⟨N1E43i03⟩, die sich in Bezug auf die gedruckte Aufgabenstellung in einem für unsere Analyse der Itemschwierigkeit wesentlichen Punkt unterscheiden. Aus den zwei Items wurden vier Items gemacht: ⟨N6E43i02⟩ ⟨N9E43i02⟩ sowie ⟨N6E43i03⟩ und ⟨N9E43i03⟩.

Basisstichprobe. Aus der ursprünglichen Stichprobe von 145 Items erhalten wir aufgrund der Reduktion und Erweiterung 141 Items, welche die Basisstichprobe für unsere Auswertungen bilden. Diese Basisstichprobe ist im Appendix B vollständig wiedergegeben.

2. Reduktion. Für unsere Analysen musste die Itemstichprobe auf die experimentellen Items reduziert werden. Die Aufgabe ⟨Rapex⟩ ist wenig experimentell im naturwissenschaftlichen Sinn und enthält nur drei Messitems, die für unsere Zwecke wenig passend sind. Die 12 Items dieser Aufgabe wurden deshalb in unseren Auswertungen nicht berücksichtigt. Weiter wurden alle nicht experimentellen Items aus der Basisstichprobe herausgestrichen. Nach dieser zweiten Reduktion bleiben 96 Experimentieritems übrig, die wir im Weiteren als die *vollständige Itemstichprobe* ⟨E08|69dfi|V⟩ bezeichnen.

5.4 Fragebogen

Die Beantwortung des Fragebogens durch die getesteten Schülerinnen und Schüler war fakultativ. Die Personenstichprobe des Fragebogens ist deshalb kleiner als die Stichprobe des Experimentiertests (man vergleiche die Tabellen 9.1 und 8.1). Die Itemmenge des Fragebogens wird im Weiteren mit der Abkürzung ⟨Q08|69dfi⟩ bezeichnet.

[2] Diese Itemreduktion wurde von Prof. Dr. Erich Ramseier vorgenommen.

Kapitel 6

Test-Analysen

6.1 Fragestellung

Das HarmoS-Kompetenzmodell unterscheidet für den Handlungsaspekt «Fragen und untersuchen» fünf Teilaspekte (cf. Tab. 6.1). Einige Teilaspekte wie «Fragen und Hypothesen formulieren» beziehen sich auf Teilprozesse des experimentellen Problemlösens. Andere wiederum beschreiben bestimmte Aufgabentypen wie das «Bewusst wahrnehmen» (HarmoS, 2008, 46). Bei der Entwicklung der Experimentieraufgaben wurden Teilaufga-

Teilaspekte von «Fragen und untersuchen»
«Bewusst wahrnehmen»
«Fragen, Probleme und Hypothesen aufwerfen»
«Erkundungen, Untersuchungen und Experimente planen, durchführen und auswerten»
«Geeignete Instrumente, Werkzeuge und Materialien auswählen und verwenden»
«Über Ergebnisse und Untersuchungsmethoden nachdenken»

Tabelle 6.1 – HarmoS-Kompetenzmodell: Teilaspekte des Handlungsaspekts «Fragen und untersuchen»

ben gemäss den Teilaspekten differenziert und als separate Items kodiert. Es stellt sich die Frage, ob dem Experimentiertest eine durch die Teilaspekte geprägte mehrdimensionale Struktur unterliegt, die Teilprozesse oder Aufgabentypen differenziert. Zum Vergleich interessiert die Differenzierung der Themenbereiche, da versucht wurde, mit den Experimentieraufgaben die Themenbereiche des HarmoS-Kompetenzmodells möglichst gut abzudecken:

1.1. Unterscheidet der HarmoS-Experimentiertest bezüglich der Teilprozesse zwischen verschiedenen Dimensionen?

⟨E08|69d⟩	«Frage/Hyp.»	«Plan.»		«Durchf.»		«Ausw.»	«Refl.»
	3	**10**	**(16)**	**30**	**(12)**	**19**	**6**
1D-Modell	allgemeine experimentelle Kompetenz «Frage/Hyp. & Plan. & Durchf. & Ausw. & Refl.»						
	96						
2D-Teilprozessmodell	rein kognitive Teilprozesse «Frage/Hyp. & Plan. & Ausw. & Refl.»					manipulative Teilprozesse «Durchf.»	
	38					58	
3D-Teilprozessmodell	prospektive Teilprozesse «Frage/Hyp. & Plan.»			retrospektive Teilprozesse «Ausw. & Refl.»		manipulative Teilprozesse «Durchf.»	
	13			25		58	

Tabelle 6.2 – Itemanzahl der verschiedenen Teilprozessmodelle der vollständigen Itemstichprobe ⟨E08|69d|V⟩; Frage/Hyp.: Fragen/Hypothesen, Plan.: Planung, Durchf.: Durchführung, Ausw.: Auswertung, Refl.: Reflexion

1.2. Unterscheidet der HarmoS-Experimentiertest bezüglich der Aufgabentypen zwischen verschiedenen Dimensionen?

1.3. Unterscheidet der HarmoS-Experimentiertest bezüglich der Themenbereiche zwischen verschiedenen Dimensionen?

In den nachfolgenden drei Abschnitten werden wir die obigen Fragen beantworten, indem wir verschiedene Dimensionsmodelle – so genannte Teilprozessmodelle, Aufgabentypenmodelle und Themenbereichmodelle – untersuchen und vergleichen.

6.2 Teilprozessmodelle

Itemstichprobe. Für die Dimensionsanalysen werden wir uns auf die deutschsprachigen Items ⟨E08|69d|V⟩ der im Abschnitt 5.3 beschriebenen vollständigen Itemstichprobe beschränken.

Die 96 Items der vollständigen Stichprobe verteilen sich, wie der Tabelle 6.2 (erste zwei Zeilen) zu entnehmen ist, auf die fünf Teilprozesse «Frage und Hypothese formulieren», «Untersuchung planen», «Untersuchung durchführen», «Daten auswerten» und «Untersuchung reflektieren». Grundsätzlich bezieht sich jedes Item auf einen Teilprozess. 28 Items werden jedoch zwei Teilprozessen zugeordnet (Zahlen in Klammern), wobei entweder die Prozesse «Planung» und «Durchführung» oder die Prozesse «Durchführung» und «Auswertung» zusammen in einem Item vorkommen. Der Schwerpunkt der Itemstichprobe liegt beim Manipulieren und Experimentieren (Teilprozess «Durchführen»), weshalb wir

	«Frage/Hyp. & Plan. & Durchf. & Ausw. & Refl.»
«Frage/Hyp. & Plan. & Durchf. & Ausw. & Refl.»	$M = 0.350$ (0.024) *Rel* = 0.583 (EAP/PV)

Tabelle 6.3 – 1D-Modell der vollständigen Itemstichprobe ⟨E08|69d|V⟩: Mittelwert der Fähigkeitsparameter M, EAP/PV-Reliabilität der Dimension *Rel*

	«Frage/Hyp. & Pla. & Ausw. & Refl.»	**«Durchf.»**
«Frage/Hyp. & Plan. & Ausw. & Refl.»	$M = 0.354$ (0.028) *Rel* = 0.553 (EAP/PV)	$r = 0.986$
«Durchf.»		$M = 0.357$ (0.022) *Rel* = 0.561 (EAP/PV)

Tabelle 6.4 – 2D-Teilprozessmodell der vollständigen Itemstichprobe ⟨E08|69d|V⟩: Korrelation r, Mittelwert der Fähigkeitsparameter M, EAP/PV-Reliabilität der Dimension *Rel*

diesen Teilprozess in der Dimensionsanalyse auch separat betrachten werden. Nur vereinzelt sind Items dem Formulieren von Fragen und Hypothesen und dem Reflektieren von Untersuchungen gewidmet.

Modelle. Dem eindimensionalen Grundmodell werden wir zwei weitere Modelle gegenüberstellen (cf. Tab. 6.2). Mit einem zweidimensionalen Modell soll untersucht werden, ob die manipulativen Items, welche das Hantieren mit Experimentiermaterial erfordern, eine andere Teilkompetenz ansprechen als die rein kognitiven Items. In einem dreidimensionalen Modell werden die kognitiven Items weiter aufgeteilt in prospektive Teilprozesse, die einer Manipulation vorangehen, und retrospektive Teilprozesse, die einer Manipulation nachfolgen.

Resultate. Für alle drei Modelle wurden separate Rasch-Analysen mit dem Programm ConQuest Version 2.0 gemäss Wu, Adams, Wilson und Haldane (2007) gerechnet. Die Analysen ergeben sowohl für das zwei- als auch für das dreidimensionale Modell durchwegs sehr hohe Korrelationen zwischen den Dimensionen (> 0.95), dies bei ungenügenden EAP/PV-Reliabilität der Dimensionen (< 0.56) (cf. Tab. 6.3, 6.4 und 6.5).

Der Modellvergleich anhand der drei informationstheoretischen Masse AIC, BIC und CAIC ergibt für das zweidimensionale Modell die kleinsten Werte (Masse gemäss Rost, 2004a, 339ff). Wie aus der Tabelle 6.6 ersichtlich, sind die Unterschiede der Modellfits gering.

	«Frage/Hyp. & Plan.»	«Durchf.»	«Ausw. & Refl.»
«Frage/Hyp. & Plan.»	M = -0.252 (0.025) Rel = 0.518 (EAP/PV)	$r = 0.951$	$r = 0.957$
«Durchf.»		M = 0.451 (0.022) Rel = 0.544 (EAP/PV)	$r = 0.990$
«Ausw. & Refl.»			M = 0.426 (0.029) Rel = 0.541 (EAP/PV)

Tabelle 6.5 – 3D-Teilprozessmodell der vollständigen Itemstichprobe ⟨E08|69d|V⟩: Korrelation r, Mittelwert der Fähigkeitsparameter M, EAP/PV-Reliabilität der Dimension Rel

	AIC[1]	**BIC**[2]	**CAIC**[3]
1D-Modell[4]	14324	14468	14627
2D-Teilprozessmodell[5]	14273	14419	14580
3D-Teilprozessmodell[6]	14278	14427	14591

Tabelle 6.6 – Vergleich der verschiedenen Teilprozessmodelle anhand der Informationsmasse AIC, BIC und CAIC

Diskussion. Das zweidimensionale Modell fittet die Daten am besten, das eindimensionale Modell am schlechtesten. Die Unterschiede der Informationsmasse sind gering und die zusätzlichen Dimensionen der mehrdimensionalen Modelle bringen keinen brauchbaren Nutzen. Die Korrelationen zwischen den Dimensionen sind durchwegs sehr hoch, weshalb die Dimensionen statistisch nicht unterschieden werden können. Der HarmoS-Experimentiertest misst Teilprozesse nicht im Sinne von Teilkompetenzen. Dieses Resultat war aufgrund des gewählten Testansatzes mit vollständigen, meist alle Teilprozesse umfassenden Experimentieraufgaben auch zu erwarten. Die Teilprozesse wurden zwar separat kodiert, die Codes sind jedoch nicht unabhängig voneinander. (cf. Tab. 6.2).

[1] AIC $= -2\log(L) + 2n_p$
[2] BIC $= -2\log(L) + \log(N)n_p$
[3] CAIC $= -2\log(L) + \log(N)n_p + n_p$
[4] Devianz: $-2\log(L) = 14006$, Parameteranzahl: $n_p = 159$, Personenstichprobe: $N = 807$
[5] $-2\log(L) = 13951$, $n_p = 161$, $N = 807$
[6] $-2\log(L) = 13950$, $n_p = 164$, $N = 807$

6.3 Aufgabentypenmodelle

Itemstichprobe. Wie in der Tabelle 6.7 (erste Zeile) ersichtlich, wird jedem Item einer der fünf Aufgabentypen ⟨Beobachtung machen⟩, ⟨Zusammenhänge untersuchen⟩, ⟨Objekte vergleichen⟩, ⟨Einfache Messung machen⟩ und ⟨Effekte herstellen⟩ zugeordnet (cf. Tabelle 7.5). Die Verteilung der Aufgabentypen auf die Items ist gleichmässiger als die Verteilung der Teilprozesse.

Modelle. Das eindimensionale Grundmodell wird mit einem zwei- und einem dreidimensionalen Modell (cf. Tab. 6.7) verglichen. Beim zweidimensionalen Modell werden Aufgaben unterschieden, bei denen die Schülerin bzw. der Schüler eine passive Rolle (Beobachtungen/Klassifikationen) und eine aktive Rolle (Untersuchungen, Vergleiche, Messungen oder Herstellungen mit manipulativen Anteilen) gegenüber der Natur einnimmt. Im dreidimensionalen Modell werden die aktiven Aufgaben in rein analysierende Aufgabentypen und Aufgaben mit synthetisierenden Anteilen aufgeteilt.

Resultate. Für alle drei Modelle wurden separate Rasch-Analysen gerechnet. Diese ergeben für beide mehrdimensionalen Modelle hohe Korrelationen zwischen den Dimensionen (0.66-0.74). Die EAP/PV-Reliabilität ist teilweise sehr tief (0.33-0.56) (cf. Tab. 6.8, 6.9).

Die beiden mehrdimensionalen Modelle erreichen praktisch den gleichen Fit mit den Daten (cf. Tab. 6.10). Das zweidimensionale Modell differenziert schlechter zwischen den Dimensionen, erreicht aber bessere Reliabilitäten. Im Vergleich zu den mehrdimensionalen Modellen schneidet das eindimensionale Modell deutlich schlechter ab.

⟨E08|69d|V⟩	**⟨Beob.⟩**	**⟨Unters.⟩**	**⟨Vergl.⟩**	**⟨Mess.⟩**	**⟨Herst.⟩**
	21	**24**	**11**	**14**	**26**
1D-Modell	⟨Beob. & Unters. & Vergl. & Mess. & Herst.⟩				
	96				
2D-Aufgabentypenmodell	passive Rolle ⟨Beob.⟩	aktive Rolle ⟨Unters. & Vergl. & Mess. & Herst.⟩			
	21	75			
3D-Aufgabentypenmodell	passive Rolle ⟨Beob.⟩	aktiv-analysierende Rolle ⟨Unters. & Vergl.⟩		aktiv-synthetisierende Rolle ⟨Mess. & Herst.⟩	
	21	35		40	

Tabelle 6.7 – Itemanzahl der verschiedenen Aufgabentypenmodelle der vollständigen Itemstichprobe ⟨E08|69d|V⟩; Beob.: Beobachtung, Unters.: Untersuchung, Vergl.: Vergleich, Mess.: Messung, Herst.: Herstellung

	⟨**Beob.**⟩	⟨**Unters. & Vergl. & Mess. & Herst.**⟩
⟨**Beob.**⟩	$M = 0.338$ (0.018) $Rel = 0.376$ (EAP/PV)	$r = 0.738$
⟨**Unters. & Vergl. & Mess. & Herst.**⟩		$M = 0.359$ (0.026) $Rel = 0.550$ (EAP/PV)

Tabelle 6.8 – 2D-Aufgabentypenmodell der vollständigen Itemstichprobe ⟨E08|69d|V⟩: Korrelation r, Mittelwert der Fähigkeitsparameter M, EAP/PV-Reliabilität der Dimension Rel

	⟨**Beob.**⟩	⟨**Unters. & Vergl.**⟩	⟨**Mess. & Herst.**⟩
⟨**Beob.**⟩	$M = 0.338$ (0.019) $Rel = 0.331$ (EAP/PV)	$r = 0.719$	$r = 0.704$
⟨**Unters. & Vergl.**⟩		$M = 0.266$ (0.024) $Rel = 0.411$ (EAP/PV)	$r = 0.664$
⟨**Mess. & Herst.**⟩			$M = 0.437$ (0.029) $Rel = 0.409$ (EAP/PV)

Tabelle 6.9 – 3D-Aufgabentypenmodell der vollständigen Itemstichprobe ⟨E08|69d|V⟩: Korrelation r, Mittelwert der Fähigkeitsparameter M, EAP/PV-Reliabilität der Dimension Rel

	AIC	**BIC**	**CAIC**
1D-Modell[4]	14324	14468	14627
2D-Aufgabentypenmodell[7]	14258	14404	14565
3D-Aufgabentypenmodell[8]	14256	14404	14568

Tabelle 6.10 – Vergleich der Aufgabentypenmodelle anhand der Masse AIC, BIC und CAIC

Diskussion. Die Modellanalyse deutet eine mehrdimensionale Struktur des HarmoS-Experimentiertests an, die auf der Unterscheidung von Aufgabentypen basiert und zumindest zwischen Aufgaben unterscheidet, die eine aktive Rolle, und solchen, die eine passive Rolle der Testpersonen erfordern. Das Resultat steht im Einklang mit der Tatsache, dass es im Gegensatz zu den Teilprozessen zwischen den Aufgabentypen praktisch keine Itemabhängigkeiten gibt. Relativiert wird das Ergebnis durch die einseitige Verteilung von Aufgabentypen auf die Schulstufen. Beobachtungs- und Klassifizierungsaufgaben wurden typischerweise vorwiegend im 6. Schuljahr eingesetzt.

[7] $-2\log(L) = 13936$, $n_p = 161$, $N = 807$
[8] $-2\log(L) = 13928$, $n_p = 164$, $N = 807$

6.4 Themenbereichmodelle

Itemstichprobe. Die Items der vollständigen Stichprobe verteilen sich gemäss Tabelle 6.11 auf die im vorangehenden Kapitel vorgestellten fünf HarmoS-Themenbereiche. Der Bereich "Planet Erde" (PE) wird als einziger Themenbereich von nur einer Experimentieraufgabe abgedeckt. Für unsere Analyse werden die sechs Items dieser Aufgabe mit mechanischem Kontext (Bewegung auf schiefer Ebene) zusammen mit den Items des Bereichs "Natur, Mensch, Gesellschaft" (NMG) mit elektrischem Kontext in den Themenbereich "Bewegung, Energie, Steuerung" (BES)integriert.

Modelle. Mit dem dreidimensionalen Modell wird die Struktur der Grundfächer Physik, Chemie und Biologie übernommen. Alternativ wird ein reduziertes zweidimensionales Modell gerechnet, bei dem den mechanischen und elektrischen Themen (Physik) die stofflichen und biologischen Themen (Chemie, Biologie) gegenübergestellt werden (cf. Tab. 6.11).

Resultate. Die Korrelationen zwischen den physikalischen und den nicht physikalischen Themenbereichen sind überraschend tief (0.45-0.53) (cf. Tab. 6.12 und 6.13). Die EAP/PV-Reliabilität sind schwach.

Beide mehrdimensionalen Modelle erreichen vergleichbare Modellfits, die deutlich besser sind als der Fit des eindimensionalen Modells (cf. Tab. 6.14).

⟨**E08|69d|V**⟩	**PE**	**BES**	**NMG**	**St**	**LL**
	6	**29**	**15**	**21**	**25**
1D-Modell	Physik & Chemie & Biologie				
	96				
2D-Themenbereichmodell	Physik: "Bewegung, Kraft, Energie, Steuerung"			Chemie & Biologie: "Stoffe & Lebewesen, Lebensräume"	
	50			46	
3D-Themenbereichmodell	Physik: "Bewegung, Kraft, Energie, Steuerung"			Chemie: "Stoffe"	Biologie: "Lebewesen, Lebensräume"
	50			21	25

Tabelle 6.11 – Itemanzahl der verschiedenen Themenbereichmodelle der vollständigen Itemstichprobe ⟨E08|69d|V⟩; PE: Planet Erde; BES: Bewegung, Energie, Steuerung; NMG: Natur, Mensch, Gesellschaft; St: Stoffe; LL: Lebewesen und Lebensräume

	“Bewegung, Kraft, Energie, Steuerung“	**“Stoffe & Lebewesen, Lebensräume“**
“Bewegung, Kraft, Energie, Steuerung“	$M = 0.431$ (.031) $Rel = 0.495$ (EAP/PV)	$r = 0.490$
“Stoffe & Lebewesen, Lebensräume“		$M = 0.307$ (.023) $Rel = 0.405$ (EAP/PV)

Tabelle 6.12 – 2D-Themenbereichmodell der vollständigen Itemstichprobe ⟨E08|69d|V⟩: Korrelation r, Mittelwert der Fähigkeitsparameter M, EAP/PV-Reliabilität der Dimension *Rel*

	“Bewegung, Kraft, Energie, Steuerung“	**“Stoffe“**	**“Lebewesen, Lebensräume“**
“Bewegung, Kraft, Energie, Steuerung“	$M = 0.431$ (.031) $Rel = 0.499$ (EAP/PV)	$r = 0.459$	$r = 0.527$
“Stoffe“		$M = 0.133$ (.030) $Rel = 0.374$ (EAP/PV)	$r = 0.847$
“Lebewesen, Lebensräume“			$M = 0.436$ (.018) $Rel = 0.350$ (EAP/PV)

Tabelle 6.13 – 3D-Themenbereichmodell der vollständigen Itemstichprobe ⟨E08|69d|V⟩: Korrelation r, Mittelwert der Fähigkeitsparameter M, EAP/PV-Reliabilität der Dimension *Rel*

	AIC	**BIC**	**CAIC**
1D-Modell[4]	14324	14468	14627
2D-Themenbereichmodell[9]	14202	14348	14509
3D-Themenbereichmodell[10]	14197	14446	14510

Tabelle 6.14 – Vergleich der Themenbereichmodelle anhand der Masse AIC, BIC und CAIC

[9] $-2\log(L) = 13880$, $n_p = 161$, $N = 807$
[10] $-2\log(L) = 13869$, $n_p = 164$, $N = 807$

Diskussion. Von allen untersuchten Modellen passen die Themenbereichmodelle am besten zu den Daten. Die Fitmasse erreichen beim 2D- und 3D-Modell die tiefsten Werte (cf. Tab. 6.14). Aufgrund der optimalen Korrelation zwischen physikalischen Themen und chemisch-biologischen Themen ist das zweidimensionale Modell zu bevorzugen. Dies bedeutet, dass der Transfer zwischen "physikalischen" Kontexten und "nicht physikalischen" Kontexten nur gering ist.

Die Situation des Experimentiertests ist somit vergleichbar mit der Situation beim HarmoS-Papier-und-Bleistifttest 2007, mit welchem die nicht experimentellen Handlungsaspekte validiert wurden. Der Papier-und-Bleistifttest differenziert nämlich hauptsächlich zwischen den Themenbereichen und praktisch nicht zwischen den Handlungsaspekten (Ramseier et al., 2011, 19).

Kapitel 7

Item-Test-Analysen

Dieses Kapitel ist der Analyse von schwierigkeitserzeugenden Merkmalen der im HarmoS-Experimentiertest verwandten Experimentieraufgaben gewidmet. Die Fragestellung der Analyse wird durch konkrete Probleme bei der Aufgabenentwicklung begründet, die im Abschnitt 7.1 erläutert werden. Im Abschnitt 7.2 wird das Kodiersystem vorgestellt, mit dem die Einflüsse der vier Bearbeitungsschritte ‹Aufgabe erfassen›, ‹Problem lösen›, ‹Antwort geben› und ‹Lösung kodieren› auf die Itemschwierigkeit unterschieden und separat modelliert werden. Die Ergebnisse der Modellierung und die Diskussion folgen im Abschnitt 7.3.

7.1 Fragestellung

Schwierigkeitsreduzierende Korrekturen von Experimentieraufgaben. An gute Experimentiertests werden besonders hohe Anforderungen gestellt. Wie jeder standardisierte Test sollten Experimentiertests Schülerleistungen reliabel, objektiv und valide messen. Je nach Zweck und Kontext eines Tests können die klassischen Validitätskriterien der Inhalts- Vorhersage- und Konstruktgültigkeit um entsprechende Kriterien (Baker et al., 1993; Linn et al., 1991; Stebler et al., 1998) erweitert bzw. zu umfassenden Dimensionen von Zielkriterien ausgebaut (e. g. *assessment development dimensions* in Solano-Flores & Shavelson, 1997) oder auf der Ebene von Teilkriterien ausdifferenziert (Miller & Linn, 2000) werden. Die Zielkriterien können sich gegenseitig ausschliessen oder sich derart bedingen, dass bei der Aufgabenentwicklung eine Priorisierung von Kriterien notwendig ist (Solano-Flores & Shavelson, 1997). Bei der Konstruktion der HarmoS-Experimentieraufgaben wurden verschiedene Zielkriterien (Z.01 - Z.13) berücksichtigt (Labudde, Metzger & Gut, 2009, 315ff), von denen folgende dreizehn als prioritär behandelt wurden.

Z.01 *Anforderungsniveau*: Das Anforderungsniveau der Aufgaben sollte auf die schwächsten Schülerinnen und Schüler der jeweiligen Schulstufe ausgerichtet sein, damit

der Test auf dem Niveau von Basisstandards gut differenziert.

Z.02 *Testdauer*: Die Experimente sollten in 30 Minuten eigenständig bearbeitet werden können.

Z.03 *Inhalts-Validität*: Die Aufgaben sollten die Teilkompetenzen des Handlungsaspekts «Fragen und untersuchen» hinreichend abdecken (cf. Tab. 6.1; HarmoS, 2008, 46).

Z.04 *Curriculums-Validität*: Die Aufgaben sollten die Themenbereiche des HarmoS-Kompetenzmodells, die aufgrund einer vergleichenden Analyse der kantonalen Lehrpläne der Schweiz festgelegt wurden, hinreichend abdecken (Szlovak, 2005; HarmoS, 2008, 67f).

Z.05 *Authentizität*: Die Aufgaben sollten authentische Experimentiersituationen im Unterricht darstellen und zu einem gewissen Grad auch die Unterrichtspraxis widerspiegeln, z. B. wenn bekannte Experimentiertechniken wie das Mikroskopieren berücksichtigt werden.

Z.06 *Stufenübergreifender Vergleich*: Es sollten Aufgaben entwickelt werden, die in beiden Schulstufen (6. und 9. Schuljahr) eingesetzt werden können.

Z.07 *Vorwissen*: Die erfolgreiche Bearbeitung der Aufgaben sollte möglichst wenig schulisches Vorwissen erfordern.

Z.08 *Innovation und Offenheit*: Die Aufgaben sollten für die Praxis fachdidaktisch innovative Inputs liefern und Vorbildfunktion haben. Aus diesem Grund sollten u. a. vor allem offene Aufgabenstellungen entwickelt werden.

Z.09 *Bedeutsamkeit*: Die Aufgaben sollten für die Schülerinnen und Schüler interessante und bedeutsame Inhalte thematisieren.

Z.10 *Ethik*: Die Aufgaben sollten ethisch vertretbar sein.

Z.11 *Praktikabilität*: Das Experimentiermaterial sollte möglichst einfach und handlich sein.

Z.12 *Sicherheit* und *Funktionstüchtigkeit*: Die Experimente sollten unabhängig von Ort, Zeit und Testsituation ungefährlich sein und zuverlässig funktionieren.

Z.13 *Finanzieller Aufwand*: Die Experimente sollte kostengünstig angefertigt werden können.

Die Zielkriterien hatten in verschiedenen Phasen der Aufgabenentwicklung unterschiedlich viel Gewicht. Während man sich bei der Erstentwicklung massgeblich an den Zielkriterien Z.03 bis Z.13 orientierte, wurden bei der Überarbeitung der Aufgaben die ersten zwei Kriterien enorm wichtig. Die Aufgabenentwürfe erwiesen sich in der Pilotierung als zu anspruchsvoll und mussten überarbeitet werden. Unter der Vorgabe der Testzeit wurden die Aufgaben auf zweierlei Weise vereinfacht. Einerseits wurden kompetenzrelevante Schwierigkeiten reduziert, indem individuell das ‹Problem lösen› mit diversen *Korrekturmassnahmen* (kompetenzrelevante Korrekturen: rK.01 - rK.11) vereinfacht wurde.

rK.01 *Aufgabenumfang*: Aufgaben wurden gekürzt, wobei individuell Teilaufgaben zu bestimmten Teilkompetenzen weggelassen wurden (e. g. besonders Aufgaben zur reflexiven Teilkompetenz «Über Ergebnisse und Untersuchungsmethoden nachdenken» erwies sich in der Pilotierung allgemein als zu schwierig).

rK.02 *Komplexität*: Problemstellungen wurden vereinfacht, indem die Anzahl der Variablen und Verknüpfungen reduziert wurden.

rK.03 *Abstraktheit*: Statt abstrakte Konzepte bzw. wissenschaftliche Begriffe wurden alltägliche, konkrete Konzepte bzw. Begriffe verwendet.

rK.04 *Vorwissen*: Erforderliches Vorwissen wurde reduziert oder es wurde im Aufgabenstamm vorgegeben.

rK.05 *Qualitative statt quantitative Messungen*: Anstelle quantitativer Messungen wurden qualitative Vergleiche gewählt.

rK.06 *Mathematisierung*: Mathematische Auswertungen der Messungen wurden vermieden.

rK.07 *Offenheit*: Aufgaben wurden weniger offen formuliert.

rK.08 *Strukturiertheit*: Aufgaben wurden stärker strukturiert.

rK.09 *Zielklarheit*: Die Ziele (gesuchten Antworten) wurden klarer umschrieben, indem mehr konkrete Vorgaben gemacht wurden, i. e. die Antwortinhalte wurden vorstrukturiert.

rK.10 *Experimentiermaterial*: Das notwendige Experimentiermaterial wurde reduziert.

rK.11 *Manipulationsstrategien*: Experimentieraufträge wurden derart umformuliert, dass komplizierte Manipulationsstrategien vermieden werden konnten.

Es wurden auch kompetenzirrelevante Schwierigkeiten reduziert mit Korrekturen (kompetenzirrelevante Korrekturen: iK.01 - iK.05), die jeweils die Arbeitsschritte ‹Aufgabe erfassen› und ‹Antwort geben› betrafen.

iK.01 *Begrifflichkeit*: In der Aufgabenstellung wurden Fachbegriffe und Fremdwörter durch Alltagsbegriffe ersetzt (e. g. ?, ?).

iK.02 *Grammatik*: Aufgabentexte wurden gekürzt und grammatikalisch vereinfacht (e. g. Anzahl Nebensätze und Satzkonstruktionen wie Konditionalsätze).

iK.03 *Inhaltsformate*: Inhaltliche Informationen im Aufgabenstamm wurden in einfacheren Formaten präsentiert.

iK.04 *Antwortformate*: Es wurden geschlossene Lückenformate gewählt, die weniger anspruchsvolle Füllformate zulassen (e. g. Markieren, Stichwort-Antworten, Zeichnen).

iK.05 *Kombination von Antwortformaten*: Mit Hilfe von kombinierten Antwortformaten wurde versucht, die Validität der Eigenrapporte zu erhöhen (Labudde, Metzger & Gut, 2009, 315f).

Die schwierigkeitsreduzierende Wirkung der bei jeder Aufgabe individuell gemachten Korrekturen wurde durch wiederholte Pilotierung in der von Solano-Flores und Shavelson (1997) beschriebenen zyklischen Weise überprüft und in diesem Sinne letztlich auch "verifiziert". Unklar bleibt dabei, welche dieser Korrekturen tatsächlich ursächlich für die erzielten Vereinfachungen der Aufgaben sind.

Schwierigkeitsrelevante Aufgabenmerkmale. Die Erfahrung der Aufgabenentwicklung bei HarmoS gibt Anlass zur begründeten Hypothese, dass gewisse der oben genannten Korrekturen für Experimentieraufgaben allgemein schwierigkeitsreduzierend wirken. D. h. es gibt ein Set von allgemeinen schwierigkeitsrelevanten Aufgabenmerkmalen ξ_z (z= 1, ... m), die durch diese Korrekturen verändert werden, und eine Funktion f, welche die Itemschwierigkeit σ_i eines beliebigen Items $i \in I$ funktional mit dessen Merkmalsausprägungen $\xi_z(i)$ (z= 1, ... m) verknüpft.

$$\forall i \in I \; \exists \; \xi_z \text{ mit } (z \in Z) \text{ und eine Funktion } f : \sigma_i = f(\xi_1(i), ...\xi_m(i)). \qquad \text{(H.1)}$$

Im Hinblick auf eine statistische Überprüfung der Hypothese (H.1) wird der vermutete funktionale Zusammenhang linearisiert, wobei für jedes Merkmal ξ_z ein dazugehöriges Gewicht x_z angenommen wird, das die Stärke der schwierigkeitsinduzierenden Wirkung beschreibt.

$$\sigma_i = f(\xi_1(i), ...\xi_m(i)) = C(i) + \sum_{z=1}^{m} x_z \xi_z(i) + O(\xi_1(i)^2) + ... + O(\xi_m(i)^2) \qquad \text{(H.2)}$$

Der folgende Abschnitt ist der Entwicklung eines Systems von Aufgabenmerkmalen $\{\xi_z\}_{z \in Z}$ gewidmet, das zwei Zwecke erfüllen soll: Erstens soll mit dem Merkmalsystem die Itemschwierigkeit der deutschsprachigen Experimentieraufgaben modelliert werden. Zweitens soll mit dem Merkmalsystem die schwierigkeitsreduzierende Wirkung der Korrekturen (rK.01-rK.11, iK.01-iK.05) erklärt werden. Da in den Tests jeweils nur die Endversion einer Aufgabe getestet wurde, kann der isolierte Effekt einer Korrektur anhand der Testdaten nicht untersucht werden. Wegen der geringen Anzahl eingesetzter Items ist zudem die Anzahl der Merkmale, die statistisch untersucht werden können, eingeschränkt. Es ist daher grundsätzlich nicht möglich, anhand der vorliegenden Testdaten die Wirkung aller Korrekturen zu analysieren.

Fragen. Zusammenfassend sollen folgende vier Forschungsfragen beantwortet werden:

2.1. Lässt sich die Itemschwierigkeit des HarmoS-Experimentiertests anhand eines Systems von Itemmerkmalen modellieren?

2.2. Können schwierigkeitsinduzierende Anforderungen eruiert werden, die für die experimentelle Kompetenz relevant sind? Inwieweit lassen sich Einflüsse der Korrekturen erkennen, die bei der Aufgabenentwicklung verwendet wurden?

2.3. Können darüber hinaus schwierigkeitsinduzierende Anforderungen eruiert werden, die für die experimentelle Kompetenz nicht relevant sind? Inwieweit lassen sich Einflüsse der Korrekturen erkennen, die bei der Aufgabenentwicklung verwendet wurden?

2.4. Inwieweit eignet sich der HarmoS-Experimentiertest als Messinstrument experimenteller Kompetenz? Inwieweit misst er kompetenzrelevante und -irrelevante Fähigkeiten?

7.2 HarmoS-Experimentiertest ⟨E08|69d⟩: Modellierung der Itemschwierigkeit

7.2.1 Entwicklung eines Systems schwierigkeitsrelevanter Itemmerkmale

Beim HarmoS-Experimentiertest wird mit verschiedenen Aufgabentypen und unterschiedlichen Teilprozessen ein breites Themenfeld abgedeckt. Die Items sind sowohl in Bezug auf die Aufgabenstellungen als auch in Bezug auf die Kodierschemen sehr heterogen, dies u.a. als Folge der erwähnten Korrekturmassnahmen. Neben manipulativ-experimentellen

Items, die praktische Handlungen mit Experimentiermaterial erfordern, und rein kognitiven "Experimentier"-Items, die im Rahmen eines Papier-und-Bleistift-Tests verwendet werden könnten, gibt es auch Anwendungs- und Theorie-Items, die nicht zur Messung experimenteller Kompetenz benutzt werden. Eine Zusammenstellung aller kodierten Items findet sich im Appendix B.

Die für unsere Zwecke benutzten Aufgabenmerkmale beziehen sich bis auf vier Merkmale auf einzelne Items. Das Merkmal {Aufgabentyp} wird nur den Testlet-Items zugeordnet. Die drei Merkmale zu den Manipulationsstrategien werden nur manipulativen Items zugeordnet, bei denen praktische Handlungen geplant, getätigt oder reflektiert werden. Da sich die betrachteten Merkmale auf drei verschiedene Itemgruppen beziehen, werden wir schliesslich auch drei Analysen vornehmen (cf. Kapitel 7.3).

Entwicklung des Merkmalsystems. Bei der Entwicklung des Merkmalsystems haben wir folgende Punkte berücksichtigt. Man vergleiche auch die schematische Darstellung der Struktur des entwickelten Merkmalkatalogs in Abbildung 2.1 auf S. 7.

i. *Aufgabenelemente*: Eine Experimentieraufgabe umfasst den Testbogen, das Experimentiermaterial und das Kodierschema. Die analysierten Aufgabenmerkmale beziehen sich daher jeweils auf eines der drei Aufgabenelemente.

ii. *Arbeitsschritte*: Bei der Entwicklung des Merkmalsystems wurde darauf geachtet, dass alle vier Arbeitsschritte ‹Aufgabe erfassen›, ‹Problem lösen›, ‹Antwort geben› und ‹Lösung kodieren› berücksichtigt sind.

iii. *Progressionsdimensionen*: Ebenso wurde darauf geachtet, dass die drei Progressionsdimensionen $[A,P,Q]$ berücksichtigt werden.

iv. *Korrekturmassnahmen*: Angestrebt wurde ein Merkmalsystem, das zu jeder Korrekturmassnahme (rK.01-rK.11,iK.01-iK.05) auf Seite 124 ein korrespondierendes Aufgabenmerkmal enthält.

v. *Literatur*: Weiter wurden in der Literatur beschriebene Aufgabenmerkmale auf den HarmoS-Experimentiertest adaptiert und übernommen.

vi. *Aufgabenstrukturmerkmale versus Problemstrukturmerkmale*: Die Kodierung von Aufgaben erfolgt analog zur Unterscheidung von Prenzel et al. (2002, 125) zwischen "formalen Aufgabenmerkmalen" und "kognitiven Anforderungen": Einerseits werden "Oberflächenmerkmale" der Aufgabenstellung und des Experimentiermaterials beschrieben, andererseits wird die "Tiefenstruktur" eines Problems analysiert. Wir unterscheiden daher zwischen *Merkmalen der Aufgabenstruktur* (e. g. Textlänge, grammatische Satzstruktur, Darstellungsformaten von Informationen, Antwortformaten

etc.), welche die inneren und äusseren Merkmale der Aufgabenstellung beschreiben, und *Merkmale der Problemstruktur* (e. g. Problemoffenheit, Zielklarheit, Strukturiertheit), welche die innere Problemstruktur betreffen. In aller Regel werden Aufgabenstrukturmerkmale den Arbeitsschritten ‹Aufgabe erfassen› und ‹Antwort geben› zugeordnet, während die Charakterisierung der Arbeitsschritte ‹Problem lösen› und das ‹Lösung kodieren› eine Analyse der Problemstruktur benötigt (cf. Abs. 4.4).

vii. *Dichotome oder metrische Variablen*: Geplant ist eine multilineare Regressionsanalyse der Itemschwierigkeit mit einem Satz von Aufgabenmerkmalen wie er u. a. bei Prenzel et al. (2002, 129ff) beschrieben wird. Hierfür werden entweder rein metrische oder – im Falle von nicht metrischen Skalen – dichotome Variablen (Dummy-Variablen) benötigt. Wir werden ausschliesslich dichotome Variablen benutzen. Metrische Masse, die sich meist auf äussere Aufgabenmerkmale beziehen, werden wir hierfür vorgängig dichotomisieren.

viii. *Monotoniebedingung*: Die Skalen der Aufgabenmerkmale werden so gewählt, dass durch das Zusammenfassen von zwei Items zu einem Summenitem die Merkmalswerte nicht kleiner werden. Mit dieser Monotonieannahme werden wir einerseits der Vermutung gerecht, dass mit zunehmendem Aufgabenumfang eine Aufgabe nicht leichter wird. Wichtig wird diese Prämisse jedoch bei der späteren Testlet-Bildung von abhängigen Items (cf. Kap. 7.3.5).

ix. *Zirkuläre Entwicklung*: Ob sich ein Merkmal eignet, um Testaufgaben zu klassifizieren, entscheidet sich letztlich erst bei der Kodierung. Die Experimentieraufgaben wurden daher bereits während der Entwicklung des Merkmalsystems mehrmals pilotkodiert, wobei nach jeder Kodierung das Merkmalsystem angepasst wurde.

x. *Reduktion des Merkmalsystems*: Aufgrund der beschränkten Anzahl von 96 Experimentieritems darf das Merkmalsystem einen gewissen Umfang nicht überschreiten. Eine Reduktion möglicher Aufgabenmerkmale musste daher vorgenommen werden. Hierfür wurden in einem ersten Schritt Teilsysteme von Merkmalen, die sich auf jeweils ein Arbeitsschritt ‹Aufgabe erfassen›, ‹Problem lösen›, ‹Antwort geben› und ‹Lösung kodieren› beziehen, separat analysiert und reduziert. In einem zweiten Schritt wurden die verbleibenden Merkmale zu einem System zusammengezogen und nochmals analysiert.

7.2.2 System kompetenzirrelevanter Itemmerkmale

Analyse der Aufgabenstruktur

Informationsformate. Die Analyse der inneren und äusseren Aufgabenstruktur bauen wir auf einer Analyse der *Informationsformate der Aufgabenstellung* auf. Dabei gehen wir vom Prinzip aus, dass jedes Element der Aufgabenstellung (Satz, Tabelle, Abbildung etc.) Informationen enthält, die einem bestimmten Hauptzweck dienen. Wir unterscheiden sieben Zwecke, die entsprechend sieben Typen von Informationsformaten definieren. Es gilt zu beachten, dass die Analyse der Aufgabenstellung mit Hilfe von Informationsformaten nicht nur äussere Aufgabenmerkmale betrifft, sondern auch Bewertungen über die innere Problemstruktur einschliesst. Wir werden daher später Informationsformate auch benutzen, um Merkmale der inneren Problemstruktur zu definieren.

C/D: *Inhaltsformate C* enthalten für die Aufgabe relevante naturwissenschaftliche Inhalte wie Fakten, Zusammenhänge, Definitionen, Bezeichnungen. *Disktrationsformate D* sind Inhaltsformate, die für die Aufgabe irrelevante Inhalte enthalten.

A: *Antwortformate A* sind freie Stellen in der gedruckten Aufgabenstellung (freie Flächen zwischen den Texten, leere Zeilen, Lücken im Text, leere Zellen in einer Tabelle, leere Diagramme, Abbildungen oder leere Felder zum Ankreuzen etc.), die von der Testperson mit ihren Antworten gefüllt werden. Diese Stellen werden von uns daher auch als *Lückenformate L* bezeichnet. Ein Lückenformat strukturiert die Antwort mehr oder weniger vor und gibt die Darstellungsart der Antwort vor (Texte, Zeichnungen, Markierungen, Graphen etc.). Die Darstellungsart der Antworten bezeichnen wir im Weiteren als *Füllformat F*.

P/T: Bei Testaufgaben, deren Kodierung ausschliesslich auf dem Eigenrapport der Testperson beruht, gilt es zwischen der *intendierten Aufgabe* und der *kodierten Aufgabe* zu unterscheiden. Die intendierte Aufgabe enthält das von den Testkonstrukteuren intendierte, zu lösende naturwissenschaftliche Problem. Aus der Beschreibung dieses Problems, die mit Hilfe von *Problemformaten P* erfolgt und die auch motivationale Aspekte enthalten kann, wird meist nicht klar, was denn letztlich der Inhalt der gesuchten Antwort ist. Die zweite Aufgabe für die Testperson besteht also darin, herauszufinden, mit welchen Inhalten und Füllformaten die Lückenformate zu füllen sind. Informationseinheiten, die den Inhalt und das Füllformat der gesuchten Antwort beschreiben, bezeichnen wir als *Aufgabenformat T*.

S: Im Gegensatz zu Papier-und-Bleistift-Aufgaben enthalten Hands-on-Tests häufig Anweisungen oder Tipps zum Vorgehen beim Experimentieren und zur Verwen-

dung und Handhabung des Experimentiermaterials. Mit diesen Vorgaben wird der Lösungsweg vorstrukturiert, weshalb wir sie als *Strukturierungsformate S* bezeichnen. Nicht jede Strukturierung ist jedoch für die Problemlösung gleich relevant. Hinweise wie "Lese zuerst den Text in der Informationsbox, bevor du mit Experimentieren beginnst" werden als wenig relevant und daher als nur schwache Strukturierung gewertet. Die Anweisung "Führe zwei Experimente durch" wird hingegen als relevant und demzufolge als starke Strukturierung erachtet.

O: *Orientierungsformate O* enthalten Angaben, die den Zweck anderer Informationsformate erklären (e. g. Überschriften wie "Deine Aufgaben").

M: *Motivationsformate M* stellen die Aufgabe in einen übergeordneten, für die Lösung der Aufgabe aber unerheblichen Bedeutungskontext.

Abgrenzung der Informationsformate. Bei der Kodierung der Formate ist man mit einem doppelten Abgrenzungsproblem konfrontiert. Vor der Kodierung stellt sich das Problem, die *Länge der Informationsformate* festzulegen. Je nachdem, ob man einen grammatikalisch abgeschlossenen Satz oder jeden Teilsatz als Einheit nimmt, ergibt die Formatanalyse ein anderes Ergebnis. Man muss sich die Frage stellen, ob man Abbildungen, Tabellen oder Diagramme als ein Format betrachtet oder in mehrere sinnvolle Teilformate zerlegt. Die Festlegung der Formatlänge erfolgt a priori aufgrund von Praktikabilitätsüberlegungen. Für unsere Zwecke wurden bei Texten grammatikalische Sätze als Informationseinheiten kodiert. Tabellen, Abbildungen und Diagramme wurden ebenfalls je als eine Informationseinheit behandelt.

Das zweite Problem ergibt sich bei der Kodierung. Nicht immer kann einem Informationsformat ein eindeutiger Zweck zugeordnet werden. Oft erfüllen Formate mehrere Zwecke (e. g. Aufgabenformate sind immer auch zu einem bestimmten Grad Problemformate). In diesen Fällen wurden bei unserer Kodierung einer Informationseinheit zwei Formate zugeordnet. Mehr als zwei Formate wurden in keinem Fall kodiert.

Formatanalyse der Aufgabenstellung. Die Formatanalyse erfolgt in mehreren Schritten, die am Beispiel des Aufgabenstamms ⟨N1E23i00⟩ und der ersten drei Items ⟨N1E23i01-03⟩ der Aufgabe ⟨Balkenwaage⟩ erläutert werden. Die Kodierung dieser Items ist in der Abbildung 7.1 als Auszug einer EXCEL-Tabelle dargestellt. Im Folgenden werden die Codes Spalte für Spalte erklärt.

⟨Balkenwaage⟩	alle Formate							C/D		P	T	S	A (L/F)		
Informationseinheiten der Aufgabenstellung	{I-Typ}	{I-Anz}	{I-Sep}	{I-Sät}	{I-Tei}	{IDTei}	{I-Wör}	{C-Art}	{C-Wör}	{P-For}	{T-For}	{S-Art}	{SrFor}	{L-Art}	{F-Art}
⟨N1E23i00⟩ Strukturcode: **CPCP**	**2**	**2**	**9**	**13**	**4**	**118**	**TEX/ABB**	**75**	**2**	**0**			**0**		
Balkenwaage	**C** 1					1	TEX	1							
Finde ...	**P** 1		1	2	1	14			1						
Bei diesem ...	**C**	1	1	1	0	12	TEX	12							
Informationen	O					1									
Es gibt ...	C		1	2	1	16	TEX	16							
Nur in ...	C		1	1	0	9	TEX	9							
Wenn die ...	C		1	2	1	17	TEX	17							
Zwei Bilder ...	O		1	1	0	6									
[Abbildung]	C					7	ABB	6							
[Abbildung]	C					6	ABB	6							
Deine Aufgaben	O					2									
Auf den ...	O		1	1	0	9									
Unter diesen ...	C		1	1	0	8	TEX	8							
Überprüfe ...	**P**	1	1	2	1	10			1						
⟨N1E23i01⟩ Strukturcode: **CSTA**	**4**	**0**	**6**	**8**	**2**	**52**	**TEX**	**9**	**0**	**2**			**1**	**Abb/MuC**	**zei/mar**
Behauptung 1	O					2									
Eine symmetrisch ...	**C** 1		1	1	0	9	TEX	9							
Führe zwei ...	**S** 1		1	2	1	9						rel.	1		
Zeichne sie ...	**T** 1		1	1	0	6					1				
Kreuze ...	T		1	2	1	10					1				
1. Experiment	O					2									
[Abbildung]	**A** 1					3								Abb	zei
[Multiple choice]	A		1	1	0	5								MuC	mar
(dito)			1	1	0	6								MuC	mar
⟨N1E23i02⟩ Strukturcode: **A**	**1**	**0**	**2**	**2**	**0**	**16**		**0**	**0**	**0**			**0**	**Abb/MuC**	**zei/mar**
2. Experiment	O					2									
[Abbildung]	**A** 1					3								Abb	zei
[Multiple choice]	A		1	1	0	5								MuC	mar
(dito)			1	1	0	6								MuC	mar
⟨N1E23i03⟩ Strukturcode: **TA**	**2**	**0**	**4**	**4**	**0**	**15**		**0**	**0**	**2**			**0**	**MuC**	**mar**
Was ...	**T** 1		1	1	0	5				1					
Kreuze an!	T		1	1	0	2				1					
[Multiple choice]	**A** 1		1	1	0	4								MuC	mar
(dito)			1	1	0	4								MuC	mar

Abbildung 7.1 – Formatanalyse am Beispiel des Aufgabenstamms ⟨N1E23i00⟩ und der ersten drei Items ⟨N1E23i01-03⟩ der Aufgabe ⟨Balkenwaage⟩

- Zuallererst wird die gedruckte Aufgabenstellung in zusammenhängende informationstragende Elemente, so genannte Formate (e. g. Sätze, Titel, Abbildungen, Tabellen, Diagramme, Antwortlücken in der gedruckten Aufgabenstellung etc.) zerlegt und die Formate werden dem natürlichen Lesefluss nach geordnet (Spalte "Inhaltsformate gedruckte Aufgabenstellung").

- Dann wird jedem Format ein Zweck zugeordnet, d. h. es wird der Formattyp bestimmt (Spalte {I-Typ}).

- Anhand der Reihenfolge und der Anzahl der Formattypen wird zuerst der Strukturcode, dann die Itemmerkmale {I-Anz} und {I-Sep} bestimmt (cf. Ausführungen auf S. 134).

- Zusätzlich werden für jedes Item die Anzahl Sätze {I-Sät}, Anzahl Teilsätze {I-Tei}, die Differenz von Anzahl Teilsätzen und Sätzen {IDTei} sowie die Anzahl Wörter {I-Wör} berechnet.

- Im Weiteren werden die einzelnen Formattypen C/D, P, T, A(L/F) qualifiziert. Bei den Inhaltsformaten wird die Darstellungsart {C-Art} und die Anzahl Wörter {C-Wör} erhoben.

- Bei den P- und T-Formaten wird die Häufigkeit (Anzahl) dieser Formate bestimmt: {P-For}, {T-For}.

- Bei den Strukturierungsformaten wird festgehalten, wie viele relevante Strukturierungen in einem Item vorkommen: {SrFor}.

- Zuletzt werden bei jedem Aufgabenformat A das Lücken- und das Füllformat bestimmt: {L-Art}, {F-Art}.

Sämtliche äusseren Aufgabenmerkmale, die aus der Formatanalyse resultieren, sind in der Tabelle 7.1 aufgelistet.

Begründung der Formatanalyse. Grundsätzlich lassen sich mit einer Formatanalyse beliebig viele Itemmerkmale konstruieren, für die es keinen Zusammenhang mit der Itemschwierigkeit gibt. Die Verwendung eines Itemmerkmals muss daher mit einer plausiblen Hypothese begründet werden, die diesen Zusammenhang erklärt. Es besteht also der Bedarf einer allgemeinen Begründung der Formatanalyse.

Format	Merkmal	Beschreibung
alle	{I-Typ}	Formattyp: M, O, C/D, P/T, S, A
	{I-Anz}	Anzahl unterschiedliche Formattypen
	{I-Sep}	Anzahl Separationen zwischen zwei Formaten des gleichen Typs C/D, S, P/T, A
	{I-Sät}	Anzahl Sätze
	{I-Tei}	Anzahl Teilsätze
	{IDTei}	Teilsatzdifferenz: Anzahl Teilsätze abzüglich die Anzahl Sätze
	{I-Wör}	Anzahl Wörter (Textlänge)
C/D	{C-Art}	Darstellungsart: Text (TEX), Abbildung (ABB), Tabelle (TAB), Diagramm (DIA), Formeln (FOR)
	{C-Wör}	Anzahl Wörter (Länge des Formats)
P	{P-For}	Anzahl Problemformate
T	{T-For}	Anzahl Aufgabenformate
S	{S-Art}*	Art des Strukturierungsformats: relevant, irrelevant
	{SrFor}*	Anzahl relevanter S-Formate
A	{L-Art}	Typ des Lückenformats: Multiple choice (MuC), Mulitple select (MuS), zu ergänzende Abbildungen (Abb), zu ergänzende Tabellen (Tab), zu ergänzende Diagramme (Dia), Lückentexte (LuT), leere Zeilen (LeZ) und leere Flächen (LeF)
	{F-Art}	Typ des Füllformats: Markieren (mar), Zeichnen (zei), in einem Diagramm Darstellen (dar), Benennen mit Stichworten (ben), Beschreiben mit Text (bes)

Tabelle 7.1 – Formatanalyse: Beschreibung der Aufgabenstrukturmerkmale: Die mit * gekennzeichneten Merkmale beschreiben kompetenzrelevante Anforderungsaspekte.

- *Arten von C-Formaten*: Wie im Abschnitt 7.1 dargelegt wurde, gründen kompetenzirrelevante Itemschwierigkeiten in Anforderungen an die Lese- und Mitteilungskompetenz einer Testperson, die anderes Wissen und andere Fähigkeiten erfordert als beim Experimentieren und Problemlösen benötigt werden. Mit Blick auf die Lesekompetenz zählen dazu das Verständnis unterschiedlicher Informationsarten (Texte, Schemata, Formeln, Diagramme, Abbildungen etc.) und notwendige terminologische Kenntnisse (cf. S. 100ff). Das Vorhandensein einer bestimmten Art eines Inhaltsformats {C-Art} gibt Aufschluss darüber, welches Verständnis und welche Wissensbasis das Erfassen der Aufgabe bedingt.

- *Arten von F-Formaten*: Wer ein Sprachsystem nicht versteht, kann sich auch nicht in diesem Sprachsystem ausdrücken. Insofern hängt die Mitteilungskompetenz von der

Lesekompetenz ab. Ob Lese- und Mitteilungskompetenz sich parallel entwickelnde Fähigkeiten im Sinne von Einhaus (2007, 172) sind, die psychometrisch nicht zu unterscheiden sind, kann hier nicht beantwortet werden. Es ist jedoch plausibel, Lesen und Schreiben eines Sprachsystems als vorerst unterscheidbare Fähigkeiten anzunehmen, auch wenn beide Kompetenzen auf teils dieselbe Wissensbasis zurückgreifen. Wir gehen also davon aus, dass das Verstehen einer Abbildung einer Experimentiersituation und das Zeichnen derselben Situation unterschiedliche Anforderungen an eine Testperson stellen. Die Unterscheidung des Füllformats {F-Art} ist daher begründet.

- *Arten von L-Formaten*: Es ist empirisch belegt, dass das ‹Antwort geben› signifikant durch die Offenheit des Antwortformats (Sind inhaltliche Vorgaben gemacht?) beeinflusst wird (cf. Abs. 4.4.2). Die Offenheit wird durch das Lückenformat {L-Art} erfasst. Es gilt zu beachten, dass die F- und L-Formate nicht unabhängig voneinander sind: Z. B. verlangt das Lückenformat {leere Zeilen} unbedingt das Füllformat {Beschreiben}.

- *P- und T-Formate*: Experimentieraufgaben sind keine Routineaufgaben, die durch eine einfache Frage der Art "Wie gross ist ... ?" – i. e. also durch ein einziges T-Format – hinreichend erklärt wird. Bei Experimentieraufgaben müssen unvorbereitete praktische Handlungen der Testperson initiiert werden. Dies gelingt nur mittels Versuchsbeschreibungen, Handlungsanweisungen und Erläuterungen, die über den Sinn und Zweck des Experiments aufklären. P-Formate, die das zu lösende experimentelle Problem um- und beschreiben, ohne die letztlich gesuchte Antwort anzusprechen, sind daher typisch für Experimentieraufgaben. Zuweilen besteht sogar eine grosse Diskrepanz zwischen dem Lösen des experimentellen Problems und dem Geben der richtigen Antwort. Die Unterscheidung der zwei Formate ist deshalb begründet.

- *S-Formate*: S-Formate strukturieren die Problemlöseprozesse, indem sie Vorgaben zum Lösungsweg machen (meist für das praktische Vorgehen beim Experimentieren). Sie schränken den Experimentiersuchraum für die Testperson ein, was die Hypothese begründet, dass der Einsatz von S-Formaten eine Aufgabe leichter macht. Im Extremfall einer Kochrezept-Aufgabe mit verschwindendem Experimentiersuchraum enthält die Aufgabenstellung hauptsächlich S-Formate. Dies gilt jedoch nicht für alle S-Formate. Eine Anweisung der Art "Read ALL directions carefully"[1] strukturiert zwar die Handlungen der Testperson, ist aber letztlich inhaltlich irrelevant für die Lösung. Wir unterscheiden daher zwischen irrelevanten und relevanten S-Formaten, wobei wir das Vorhandensein von relevanten S-Formaten {SrFor} kodieren.

- *Textlänge*: Die Länge der Aufgabenstellung hat sich bei Papier-und-Bleistift-Tests als

[1] Anweisung aus der Plastiline-Aufgabe des TIMSS-Experimentiertests (Labudde & Stebler, 1999, 27)

nicht relevant für die Schwierigkeit der Aufgabe erwiesen (cf. Abs. 4.4.1). Um einen Vergleich mit Experimentieraufgaben herzustellen, erfassen wir die Textlänge der Aufgabenstellungen in Form der Anzahl Wörter {I-Wör}. Die Textlänge des inhaltlichen Inputs erfassen wir ebenfalls als Anzahl der Wörter in C-Formaten {C-Wör}.

- *Satzbau*: Um das Lesen von Texten zu erleichtern, darf der Text weder zu hypotaktisch (hoher Anteil an untergeordneten Nebensätzen) noch zu parataktisch (niedriger Anteil an untergeordneten Nebensätzen) sein. Empfohlen wird "eine Balance, also weder ein zu parataktischer noch ein zu hypotaktischer Satzbau" (Kulgemeyer, 2009, 40). Angestrebt werden soll ein mittlerer Nebensatzquotient (= Anzahl Sätze / Anzahl Nebensätze). Der Nebensatzquotient ist jedoch kein extensives Itemmerkmal, das bei der Bildung von Summenitem die Monotoniebedingung erfüllt. Als Ersatz wählen wir die Teilsatzdifferenz, d. h. die Differenz aus Anzahl Teilsätze und Anzahl Sätze:
  {IDTei} = {I-Sät}-{I-Tei}.

- *"Struktur" der Aufgabenstellung*: Schreibt man die Abfolge der Formattypen innerhalb einer Aufgabenstellung auf, reduziert Blöcke von Mehrfachnennungen desselben Typs und streicht unwesentliche Orientierungsformate, erhält man den *Strukturcode* der Aufgabe, also so etwas wie die Struktur der Aufgabenstellung. Ein Vergleich der Strukturcodes der ersten drei Items der HarmoS-Aufgabe ⟨Balkenwaage⟩ (Abb. 7.1), der HarmoS-Aufgabe ⟨Solarzellen⟩ (cf. App. A.2) und der Tabletten-Aufgabe des TIMSS-Experimentiertests (Labudde & Stebler, 1999, 30) zeigen Ähnlichkeiten und Unterschiede der Aufgabenstellungen. Als gemeinsame Regelmässigkeit fällt z. B. auf, dass jede Teilaufgabe mit einem Antwortformat abgeschlossen wird.

  ⟨HarmoS:Balkenwaage⟩: CPCP,CSTA,A,TA, ...

  ⟨HarmoS:Solarzellen⟩: CPCS,SCPSTA,PTA, ...

  ⟨TIMSS:Tabletten⟩: CSP,STA,STA,TA, ...

  Um die "Homogenität" der Aufgabenstellung zu erfassen, kodieren wir zu jeder *Teilantwort* (= Sequenz im Strukturcode zwischen zwei A-Formaten) die Anzahl der verschiedenen Formattypen {I-Anz} und die Anzahl der Wiederholungen desselben Typs {I-Sep} (cf. Tab. 7.2) . Eine Wiederholung eines Formattyps innerhalb einer Teilantwort zeigt sich daran, dass dasselbe Format separiert vorkommt. Wir gehen von der Vermutung aus, dass *kohärentere Aufgabenstellungen* – also solche mit weniger Formattypen – einfacher zu verstehen sind.

Aufgabe		Stamm	1. Item	2. Item	3. Item	...
HarmoS: ⟨**Balkenwaage**⟩	Strukturcode	CPCP	CSTA	A	TA	...
	{I-Anz}	2	4	1	2	
	{I-Sep}	2	0	0	0	
HarmoS: ⟨**Solarzellen**⟩	Strukturcode	CPCS	SCPSTA	PTA	...	...
	{I-Anz}	3	5	3		
	{I-Sep}	1	1	0		
TIMSS: ⟨**Tabletten**⟩	Strukturcode	CSP	STA	STA	TA	...
	{I-Anz}	3	3	3	2	
	{I-Sep}	0	0	0	0	

Tabelle 7.2 – Struktur der Aufgabenstellungen: Harmos-Aufgaben ⟨Balkenwaage⟩ und ⟨Solarzellen⟩ und TIMSS-Aufgabe ⟨Tabletten⟩ im Vergleich

Homogenität von Aufgabenstellungen. Die Formatanalyse eröffnet nebst der Konstruktion von Aufgabenmerkmalen auch die Möglichkeit, Testaufgaben zu vergleichen. Die Tabelle 7.2 deutet an, wie aus den Strukturcodes die Homogenität von Aufgabenstellungen in einem Test beurteilt werden kann. Der Vergleich der Strukturcodes zeigt, dass der HarmoS-Experimentiertest im Vergleich zum TIMSS-Experimentiertest sehr heterogen ist.

Itemmerkmale zu den Arbeitsschritten ‹Aufgabe erfassen› und ‹Antwort geben›

Reduktion und Dichotomisierung der Itemmerkmale. Mit der Formatanalyse können prinzipiell unzählig viele metrische Merkmale konstruiert werden, die in irgendeiner Weise für die Beschreibung von Aufgabenstellungen sinnvoll sein können. Dies birgt für die Analyse die Gefahr von exotischen Merkmalen, die nur Einzelfälle beschreiben, oder unbegründeten Metriken. Für die Analyse von kompetenzirrelevanten Itemschwierigkeiten musste die Fülle der vorgestellten Merkmale daher reduziert und die einzelnen Merkmale im Hinblick auf die Regressionsanalyse dichotomisiert werden. Hierbei wurde wie folgt vorgegangen.

- *Reduktion von Einzelfällen*: Merkmale, die bei weniger als 8% der Items (i. e. acht Items) vorkommen, wurden entweder in der Analyse nicht berücksichtigt (e. g. Merkmal {Lückentext} oder {tabellarischer Input}) oder mit anderen Merkmalen zusammengelegt (e. g. die Merkmale {MuC} und {MuS} wurden aufgrund inhaltlicher Überlegungen zum Merkmal {MuCS} zusammengefasst).
- Dichotomisierung der metrischen Merkmale: Die metrischen Merkmale wurden mit

Hilfe eines Mediansplits dichotomisiert. In wenigen Fällen wurde aufgrund inhaltlicher Überlegungen von diesem Split abgewichen und entweder nach dem Schema "vorhanden / nicht vorhanden" (e. g. {P-For}) oder nach einem vom Median abweichenden Grenzwert gesplittet (e. g. {T-For}). Die dichotomisierten Variablen wurden mit dem Suffix "-dic" gekennzeichnet.

- *Reduktion korrellierender Merkmale*: Die Hauptkomponenten-Faktorenanalyse mit den verbliebenen Merkmalen ergab, dass sieben Merkmale {I-Anz-dic}, {I-Sät-dic}, {I-Tei-dic}, {IDTei-dic}, {I-Wör-dic}, {C-Wör-dic} und {C-Art-TEX} zusammen auf einem Faktor laden, der im weitesten Sinne den "Umfang" der Aufgabenstellung beschreibt. Deshalb wurde auf die Verwendung der Merkmale {I-Sät-dic}, {I-Tei-dic}, {C-Wör-dic} verzichtet.

Aus der Merkmalsreduktion resultiert ein Set von 17 dichotomen Variablen (Dummy-Variablen), die sechs Merkmale des Arbeitsschritts ‹Aufgabe erfassen› und drei Merkmale des Arbeitsschritts ‹Antwort geben› erfassen (cf. Tab. 7.3).

7.2.3 System kompetenzrelevanter Itemmerkmale

Itemmerkmale zum Arbeitsschritt ‹Problem lösen›

{Teilprozesse} und {Aufgabenumfang}. Mit dem dichotomen Itemmerkmal {Auf-Umf-dic} wird erfasst, ob ein Item mehrere Teilprozesse umfasst. Die Unterscheidung der Teilprozesse erfolgt in Anlehnung an das idealisierte Aufgabenmodell (siehe Tab. 3.2) anhand des in der Tabelle 7.4 dargestellten Rasters.

{Aufgabentypen}. In Analogie zu den von Ruiz-Primo und Shavelson (1996) entwickelten Aufgabentypen (cf. Tab. 4.8) haben wir die in sich abgeschlossenen Testlet-Aufgaben gemäss dem in der Tabelle 7.5 dargestellten Raster in fünf Typen eingeteilt. Die Typologie gibt das Resultat der in der Pilotierungsphase des Experimentiertests vorgenommenen Vereinfachungen der Aufgaben wider (cf. Abs. 7.1): Im Test nicht vertreten sind z. B. Aufgaben, wo der kausale Zusammenhang dreier oder mehrerer Variablen untersucht und ein Kontrollansatz erforderlich wird. Dafür gibt es viele Testlet-Aufgaben, bei denen es um die Messung einer einzelnen Variablen geht. Damit solche Aufgaben nicht allzu rezeptartig werden und der experimentelle Anspruch gewährleistet bleibt, wurde entweder der Messprozess offen gestaltet – i. e. die Testpersonen müssen zwischen alternativen Messmethoden auswählen oder selber ein Messverfahren entwickeln – oder die Messung beinhaltet die Herstellung eines bestimmten Effekts, der handwerklich-technische Präzision erfordert.

Das Merkmal {Aufgabentyp} bezieht sich sinnvollerweise auf ganze Aufgaben. Wir

Formatanalyse			Regressionsanalyse	
Merkmal	**Skala**	**Wertebereich**	**Dummy-Variable**	**Split**
‹Aufgabe erfassen›				
{I-Anz}	metrisch	0–6	{I-Anz-dic}	0–3|4–6
{I-Sep}	metrisch	0–5	{I-Sep-dic}	0|1–5
{I-Sät}	metrisch	1–12	—	—
{I-Tei}	metrisch	1–18	—	—
{IDTei}	metrisch	0–6	{IDTei-dic}	0–2|3–6
{I-Wör}	metrisch	9–206	{I-Wör-dic}	0–50|51–206
{C-Art}	nominal	TEX, ABB,	{C-Art-TEX}	—
		TAB	{C-Art-ABB}	—
{C-Wör}	metrisch	0–73	—	—
{P-For}	metrisch	0–2	{P-For-dic}	0|1–2
‹Antwort geben›				
{T-For}	metrisch	0–5	{T-For-dic}	0–2|3–5
{L-Art}	nominal	MuC, MuS,	{L-Art-MuCS}	(= MuC∨MuS)
		Abb, Tab, LuT,	{L-Art-Abb}	
		LeZ, LeF	{L-Art-Tab}	
			{L-Art-LeZ}	
			{L-Art-LeF}	
{F-Art}	nominal	mar, zei, dar	{F-Art-mar}	
		ben, bes,	{F-Art-zei}	
			{F-Art-ben}	
			{F-Art-bes}	

Tabelle 7.3 – Kompetenzirrelevante Itemmerkmale der Arbeitsschritte ‹Aufgabe erfassen› und ‹Antwort geben›

ordnen dieses Merkmal daher nicht einzelnen Items zu, sondern Itemgruppen, so genannten Testlet-Items, die eine abgeschlossene und gegenüber anderen Aufgaben unabhängige Experimentiereinheit bilden. Für die Analyse dieses Merkmals werden wir daher eine weitere Itemstichprobe, die Testlet-Stichprobe ⟨E08|d69|T⟩ in Betracht ziehen (cf. App. B).

Anhand der Verteilung der Items auf die verschiedenen Aufgabentypen und Teilprozesse kann die inhaltliche Struktur des HarmoS-Test abgelesen werden. In der Tabelle 7.6 sind zu den einzelnen Aufgabentypen alle Items aufsummiert, die einen bestimmten Teilprozess enthalten. Während einem Item jeweils nur ein Aufgabentyp zugeordnet wird (Dies entspricht dem Aufgabentyp des übergeordneten Testlet-Items), gibt es im HarmoS-Experimentiertests auch Items mit zwei Teilprozessen (Zahlen in Klammern in der Tab.

Teilprozess	Beschreibung
{Auf-Prz-Hyp} = 1	*Fragen / Hypothesen*: Naturwissenschaftlich überprüfbare Fragen bzw. Hypothesen werden formuliert.
{Auf-Prz-Pla} = 1	*Planung*: Untersuchungen, Experimente oder einzelne Messungen und Beobachtungen werden geplant.
{Auf-Prz-Dur} = 1	*Durchführung*: Untersuchungen, Experimente oder einzelne Messungen und Beobachtungen werden durchgeführt.
{Auf-Prz-Aus} = 1	*Auswertung*: Daten aus Messungen und Beobachtungen werden (anhand von vorgegebenen Fragestellungen oder Hypothesen) ausgewertet.
{Auf-Prz-Ref} = 1	*Reflexion*: Ergebnisse und Methoden von Untersuchungen, Experimenten, Messungen oder Beobachtungen werden reflektiert.
{Auf-Umf-dic} = 1	*Aufgabenumfang*: Die Kodierung eines Items bezieht sich auf mehr als nur einen Teilprozess.

Tabelle 7.4 – Itemmerkmale {Teilprozesse} und {Aufgabenumfang}: Differenzierung von fünf Teilprozessen. Die Merkmalvariable erhält den Wert 1, wenn die Beschreibung auf das Item zutrifft, ansonsten ist der Wert 0.

7.6). Und zwar betrifft dies jeweils eine der zwei Prozesskombinationen: «Planung, Durchführung» und «Durchführung, Auswertung» (cf. App. B).

{Aufgabenkontext}. Im Hinblick auf den Transferumfang spielt der fachliche Kontext eine wichtige Rolle für die Beherrschung einer Kompetenz. Mangelndes fachliches Vor- und Hintergrundwissen können die experimentellen Fähigkeiten beeinträchtigen (cf. Abs. 4.3.3). Mit dem Itemmerkmal {Aufgabenkontext} wird daher festgehalten, zu welcher Experimentieraufgabe ein Item gehört. Die Kontexte des HarmoS-Experimentiertests sind in der Tabelle 7.7 zusammengestellt.

{Problemoffenheit}, {Zielklarheit} und {Strukturiertheit}. Mit den zwei Merkmalen Zielklarheit {Pro-Zie} und Problemoffenheit {Pro-Off} wird kodiert, ob es zu einer Aufgabe mehrere Lösungen gibt und ob die Aufgabe auf mehr als eine Art gelöst werden kann (cf. Abs. 4.3.1, S. 82). Mit dem Merkmal der Strukturiertheit {SrFor-dic} wird festgehalten, ob der Lösungsweg in der Aufgabenstellung vorstrukturiert wird (vgl. hierzu die Codes in Tab. 7.8). Die zwei Merkmale Problemoffenheit und Zielklarheit sind zwar miteinander verknüpft, sie bedingen sich aber nicht gegenseitig. Das erste Testlet-Item der Balkenwaage-Aufgabe bestehend aus den ersten drei Items ⟨N9E23i01-03⟩ (cf. App. A.1) besitzt nur eine Lösung (Die Behauptung 1 ist richtig). Es führen aber sehr viele Wege zum Ziel (Die Behauptung lässt sich mit Hilfe unendlich vieler Belastungen der Waage

Teilprozess	Beschreibung
{Auf-Typ-Mes} = 1	*Messung*: Eine isolierte Variable wird mit Hilfe einer gegebenen Skala (Messinstrument) gemessen.
{Auf-Typ-Beo} = 1	*Beobachtung*: Ein Gegenstand (e. g. Blatt eines Baumes, Baum und Umgebung) wird beobachtet und beschrieben.
{Auf-Typ-Ver} = 1	*Vergleich*: Verschiedene Gegenstände werden anhand einer Variablen (Merkmal, Eigenschaft) verglichen. Der Vergleich kann durch die separate Messung der Variablen oder durch den direkten Vergleich in einer integrierten Versuchsanordnung erfolgen.
{Auf-Typ-Unt} = 1	*Untersuchung*: Der funktionale Zusammenhang zweier Variablen wird untersucht.
{Auf-Typ-Her} = 1	*Herstellung*: Messungen bzw. Beobachtungen werden durchgeführt, die auf der Herstellung eines Effekts beruhen.

Tabelle 7.5 – Itemmerkmal {Aufgabentyp}: Differenzierung von fünf Aufgabentypen: Die Merkmalvariable erhält den Wert 1, wenn die Beschreibung auf das Item zutrifft, ansonsten ist der Wert 0.

bestätigen). Zum Teilitem ⟨N9E23i01⟩ hingegen gibt es mehrere Lösungen. Im Gegensatz dazu besitzt das Testlet-Item ⟨N9E84i08T⟩ (cf. App. A.2) mehrere Lösungswege und dementsprechend viele Antworten. Die Frage, welche Solarzellenschaltung (parallel oder seriell) weniger Licht benötigt, um den Motor anzutreiben, kann unterschiedlich interpretiert und experimentell umgesetzt werden (Entweder verringert man bei gleicher Zellenfläche die Stärke des Lichtflusses oder man verkleinert die Zellenfläche bei konstantem Lichtfluss).

	Teilprozesse							
	Frage/Hyp.	**Plan.**		**Durchf.**		**Ausw.**	**Refl.**	Σ
Aufgabentyp								
Beobachtung	0	0	(0)	12	(6)	1	1	20
Messung	0	2	(2)	5	(3)	2	1	15
Vergleich	1	2	(0)	3	(0)	3	2	11
Untersuchung	2	6	(2)	5	(1)	6	2	24
Herstellung	0	0	(12)	5	(2)	7	0	26
Σ	3	10	(16)	30	(12)	19	6	96

Tabelle 7.6 – ⟨E08 |69d⟩: Itemstruktur der vollständigen Itemstichprobe; Frage/Hyp.: Frage/Hypothese; Plan.: Planung; Durchf.: Durchführung; Ausw.: Auswertung; Refl.: Reflexion

Itemmerkmal	Beschreibung
{N1E13} = 1	*Kontext "Gleiten auf schiefer Ebene"*: Das Item gehört zur Aufgabe ⟨Steine⟩.
{N9E21} = 1	*Kontext "Durchschnittsgeschwindigkeit"*: Das Item gehört zur Aufgabe ⟨Murmel⟩.
{N1E22} = 1	*Kontext "Batterien"*: Das Item gehört zur Aufgabe ⟨Taschenlampe⟩.
{N1E23} = 1	*Kontext "Verallgemeinertes Hebelgesetz"*: Das Item gehört zur Aufgabe ⟨Balkenwaage⟩.
{N9E41} = 1	*Kontext "Wasserhärte"*: Das Item gehört zur Aufgabe ⟨Seife⟩.
{N9E42} = 1	*Kontext "Ölfleckversuch"*: Das Item gehört zur Aufgabe ⟨Öl⟩.
{N1E43} = 1	*Kontext "Auflösen von Brausetabletten"*: Das Item gehört zur Aufgabe ⟨Tabletten⟩.
{N6E44} = 1	*Kontext "Schwimmen und Sinken in Wasser"*: Das Item gehört zur Aufgabe ⟨Schwimmen & Sinken⟩.
{N9E53} = 1	*Kontext "Zellen"*: Das Item gehört zur Aufgabe ⟨Wasserpest⟩.
{N1E54} = 1	*Kontext "Verhalten von Tieren"*: Das Item gehört zur Aufgabe ⟨Asseln⟩.
{N6E56} = 1	*Kontext "Blumen"*: Das Item gehört zur Aufgabe ⟨Gänseblümchen⟩.
{N6E61} = 1	*Kontext "Bäume und Blätter"*: Das Item gehört zur Aufgabe ⟨Laubbäume⟩.
{N9E83} = 1	*Kontext "Spar- und Glühlampen"*: Das Item gehört zur Aufgabe ⟨Sparlampe⟩.
{N9E84} = 1	*Kontext "Solarzellen"*: Das Item gehört zur Aufgabe ⟨Solarzellen⟩.

Tabelle 7.7 – Itemmerkmal {Aufgabenkontext}: Die Merkmalvariable erhält den Wert 1, wenn die Beschreibung auf das Item zutrifft, ansonsten ist der Wert 0.

Itemmerkmal	Beschreibung
{Pro-Off} = 1	*Problemoffenheit*: Ein Item kann auf zwei oder mehr als zwei qualitativ unterschiedlichen Wegen gelöst bzw. erfolgreich bearbeitet werden.
{Pro-Zie} = 1	*Zielklarheit*: Zu einem Item gibt es zwei oder mehr als zwei als korrekt kodierte Lösungen.
{SrFor-dic} = 1	*Strukturiertheit*: Der Lösungsweg eines Items wird in der Aufgabenstellung vorstrukturiert. Das Item enthält mindestens ein relevantes Strukturierungsformat (cf. Tab. 7.1 auf S. 132).

Tabelle 7.8 – Itemmerkmale Problemoffenheit, Zielklarheit und Strukturiertheit: Die Merkmalvariable erhält den Wert 1, wenn die Beschreibung auf das Item zutrifft, ansonsten ist der Wert 0.

Itemmerkmal	Beschreibung
{Man-I-bek} = 1	*Bekanntheit der relevanten Variablen*: Die zu manipulierenden Variablen werden in der Aufgabenstellung benannt.
{Man-R-bek} = 1	*Bekanntheit der Realisationsmethode*: Die Methode, wie ein bestimmter Variablenwert realisiert werden soll, ist bekannt.
{Man-R-dir} = 1	*Direktheit der Realisationsmethode*: Die Realisation der Messsituation erfordert weder Messinstrumente noch spezielles theoretisches Vorwissen.
{Man-O-bek} = 1	*Bekanntheit der Observationsmethode*: Die Methode, wie die gesuchten Variablen gemessen werden, ist bekannt.
{Man-O-dir} = 1	*Direktheit der Observationsmethode*: Die Messung / Beobachtung erfolgt ohne Hilfsmittel und Messinstrumente (z. B. mit blossem Auge), ansonsten ist der Wert 0.

Tabelle 7.9 – Itemmerkmale zu Manipulationen: Die Merkmalvariable erhält den Wert 1, wenn die Beschreibung auf das Item zutrifft.

{Manipulationsstrategien}. Auf der Ebene der Manipulationen, die bei praktischen Items erforderlich sind, wurden drei Aufgabenaspekte kodiert. Ausgangspunkt der Betrachtung ist die Idee, dass beim praktischen Experimentieren stets bestimmte Variablen eines physikalisch-chemischen oder biologischen Systems manipuliert werden, indem sie entweder gezielt verändert oder konstant gehalten werden, während andere Variablen gemessen bzw. beobachtet werden. Die Analyse eines experimentellen Problems bezieht somit folgende vier Phasen der Manipulation ein.

Identifikation: Die für die Messung relevanten Variablen werden identifiziert.

Realisation: Die Messsituation wird vorbereitet, indem die notwendigen Variablen eingestellt werden. Dabei werden bei diesen Variablen bestimmte Werte realisiert. Dies kann auch die Messung dieser Variablen beinhalten.

Observation: Ist die Messsituation vorbereitet, werden bestimmte Variablen gemessen.

Derivation: Aus den Messergebnissen werden gesuchte Variablen berechnet bzw. abgeleitet. Dies kann eine mathematische Rechnung oder einen logischen Schluss beinhalten.

Die Schwierigkeit einer experimentellen Problemlöseaufgabe kann nun darin bestehen, dass erstens nicht bekannt ist, welche Variablen relevant sind (Problem der Identifikation), zweitens nicht klar ist, wie und mit welchen Mittel eine bestimmte Messsituation realisiert (Problem der Realisation), die Messung gemacht (Problem der Observation)

und der Schluss daraus gezogen wird (Problem der Derivation). Daraus ergibt sich eine Vielzahl an Faktoren, welche die Schwierigkeit einer Experimentieraufgabe beeinflussen können. Aus dieser Vielzahl konnten wegen der kleinen Itemzahl nur wenige Merkmale sinnvoll kodiert werden (u. a. wurde auf das Erfassen der Derivationsphase verzichtet, cf. Tab. 7.9).
Die vorgestellten Merkmale können nur Items sinnvoll zugeordnet werden, die manipulative Prozesse beinhalten oder thematisieren. Um den Einfluss von Manipulationsmerkmalen auf die Itemschwierigkeit zu untersuchen, wurde eine separate Analyse mit einer auf Planungs- und/oder Durchführungsitems reduzierte Teilstichprobe $\langle$E08|d69|man$\rangle$ (= $\langle$E08|d69|{Auf-Typ-Pla}=1 $\vee$ {Auf-Typ-Dur}=1$\rangle$) gemacht.

Itemmerkmal	**Beschreibung**
{Kod-T} = 1	*Theoretische Korrektheit*: Das Kodiersystem enthält einen oder mehrere Massstäbe, die die Korrektheit von verwandtem Theoriewissen bewerten.
{Kod-E} = 1	*Evidentielle Korrektheit*: Das Kodiersystem enthält einen oder mehrere Massstäbe, die die Korrektheit von Messergebnissen oder Beobachtungen bewerten.
{Kod-H} = 1	*Heuristische Korrektheit*: Das Kodiersystem enthält einen oder mehrere Massstäbe, die den korrekten Umgang mit Variablen bewerten.
{Kod-M} = 1	*Messtechnisch-praktische Korrektheit*: Das Kodiersystem enthält einen oder mehrere Massstäbe, die die korrekte Anwendung von technischem und praktischen Wissen beim Experimentieren und Messen bewerten.
{Kod-S} = 1	*Schlusslogische Korrektheit*: Das Kodiersystem enthält einen oder mehrere Massstäbe, die die Korrektheit von einer schlusslogischen Folgerung bewerten.
{Kod-PE} = 1	*Präzision von Messergebnissen*: Das Kodiersystem enthält einen oder mehrere Massstäbe, die die Genauigkeit von Messergebnissen bewerten.
{Kod-PB} = 1	*Präzision von Beobachtungen*: Das Kodiersystem enthält einen oder mehrere Massstäbe, welche die Genauigkeit von Beobachtungen bewerten.
{Kod-VL} = 1	*Vollständigkeit bzgl. der Lückenformate*: Das Kodiersystem bewertet direkt oder indirekt, ob alle Antwortformate bearbeitet wurden.
{Kod-VI} = 1	*Vollständigkeit bzgl. der verlangten Inhalte*: Das Kodiersystem bewertet, ob die Antwort alle inhaltlichen Vorgaben in der Aufgabenstellung vollständig erfüllt.

Tabelle 7.10 – Itemmerkmale {Kodiermassstäbe}: Die Merkmalvariable erhält den Wert 1, wenn die Beschreibung auf das Kodiersystem des Items zutrifft, ansonsten ist der Wert 0.

Itemmerkmale zum Arbeitsschritt ‹Lösung kodieren›

Mit den Itemmerkmalen zum Arbeitsschritt ‹Lösung kodieren› wird festgehalten, welche Massstäbe bei der Kodierung verwendet werden. Im Fall des HarmoS-Experimentiertests betrifft dies die korrekte Anwendung von Wissensarten (Korrektheit), die präzise Durchführung von bestimmten Teilprozessen (Präzision) sowie die vollständige Bearbeitung einer Aufgabe (Vollständigkeit). Die kodierten Massstäbe sind in der Tabelle 7.10 beschrieben.

7.3 Analyse der Itemschwierigkeit

7.3.1 Merkmalkatalog

Für alle im letzten Abschnitt eingeführten und in der Tabelle 7.11 zusammengestellten Itemmerkmale wird ein Einfluss auf die Itemschwierigkeit vermutet. Im Einzelfall kann eine Wirkung durch die theoretische Erörterung der Modellierung von Itemschwierigkeit (cf. Abs. 4.3 und 4.4) oder durch die Ausführungen zur Entwicklung des HarmoS-Experimentiertests zu Beginn dieses Kapitels (cf. Abs. 7.2) begründet werden.

Zur Übersicht vergleiche man die Tabelle 7.11 mit der Abbildung 2.1 auf S. 7. Während die Abbildung 2.1 schematisch zeigt, welche Kategorien einer Experimentieraufgabe mit welchen Instrumenten erfasst werden, enthält die Tabelle 7.11 die konkreten dichotomisierten Variablen, mit welchen die Kategorien letztlich erfasst wurden.

7.3.2 Analysen und Methoden

Die eingangs des Kapitels wiederholten Forschungsfragen (cf. S. 125) wollen wir mit drei Analysen der Itemschwierigkeit beantworten. Die Analysen erfolgen jeweils nach derselben Methode, jedoch mit verschiedenen Itemstichproben. Die erste *vollständige Stichprobe* ⟨E08|69d|V⟩ umfasst alle Items, die in der Deutschschweiz im 6. und/oder 9. Schuljahr verwendet wurden. Zur zweiten *reduzierten Stichprobe* ⟨E08|69d|red⟩ gehören alle Items der vollständigen Stichprobe, die einen manipulativen Teilprozess beinhalten (Planung oder Durchführung von Manipulationen). Bei der dritten, so genannten *Testlet-Stichprobe* ⟨E08|69d|T⟩ wurden alle Items der vollständigen Stichprobe, die untereinander Itemabhängigkeiten aufweisen, zu Summenitems (Testlets) zusammengefasst. Je nach Stichprobe variieren die Anzahl der Items (cf. Tab. 7.12) und die Menge der analysierten Merkmale. Merkmale, die Manipulationen oder den Aufgabentyp beschreiben, können mit der vollständigen Stichprobe nicht analysiert werden.

kompetenzirrelevante Itemmerkmale		**kompetenzrelevante Itemmerkmale**	
‹Aufgabe erfassen›	**‹Antwort geben›**	**‹Problem lösen›**	**‹Lösung kodieren›**
SPRACHE	LÜCKENFORMATE	PROBLEM	KORREKTHEIT
Textlänge:	{L-Art-LeF}	*Teilprozesse*:	{Kod-E}
{I-Wör-dic}	{L-Art-LeZ}	{Auf-Prz-Hyp}	{Kod-H}
	{L-Art-Tab}	{Auf-Prz-Pla}	{Kod-M}
Textkohärenz:	{L-Art-Abb}	{Auf-Prz-Dur}	{Kod-S}
{I-Anz-dic}	{L-Art-MuCS}	{Auf-Prz-Aus}	{Kod-T}
{I-Sep-dic}		{Auf-Prz-Ref}	
Satzstruktur:		*Aufgabenumfang*:	
{IDTei-dic}		{Auf-Umf-dic}	
		Aufgabentyp:	
		{Auf-Typ-Mes}	
		{Auf-Typ-Beo}	
		{Auf-Typ-Ver}	
		{Auf-Typ-Unt}	
		{Auf-Typ-Her}	
		Aufgabenkontext:	
		{N1E13}, {N9E21}	
		{N1E22}, {N1E23}	
		{N9E41}, {N9E42}	
		{N1E43}, {N6E44}	
		{N1E53}, {N9E54}	
		{N6E56}, {N6E61}	
		{N9E83}, {N9E84}	
INPUT INHALTE	FÜLLFORMATE	LÖSUNGSWEG	PRÄZISION
{C-Art-TEX}	{F-Art-mar}	{Pro-Off}	{Kod-PE}
{C-Art-ABB}	{F-Art-zei}	{Pro-Zie}	{Kod-PB}
	{F-Art-ben}	{SrFor-dic}	
	{F-Art-bes}		
PROBLEM-BESCHREIBUNG	AUFGABEN-BESCHREIBUNG	MANIPULA-TIONEN	VOLLSTÄNDIG-KEIT
{P-For-dic}	{T-For-dic}	{Man-I-bek}	{Kod-VL}
		{Man-R-bek}	{Kod-VI}
		{Man-R-dir}	
		{Man-O-bek}	
		{Man-O-dir}	

Tabelle 7.11 – Katalog analysierter Itemmerkmale

Jede Analyse erfolgt nach derselben Methode und beinhaltet folgende vier Schritte.

1. Auswahl der Itemstichprobe. Ausgangspunkt ist der Experimentiertest ⟨E08|69d⟩ (cf. Abs. 5.3). Für die Analysen beschränken wir uns auf 15 Testaufgaben an insgesamt 127 Items, die in der Deutschschweiz mit insgesamt 768 Schülerinnen und Schülern des 6. und 9. Schuljahres (6. Schuljahr: 363, 9. Schuljahrs: 405) "getestet" wurden (vgl. die Itemübersicht im App. B). Von den 127 Items wurden 96 Items ausgeschieden, die sich konkret auf eine der fünf Teilprozesse experimenteller Kompetenz «Fragen/Hypothesen», «Planung», «Durchführung», «Auswertung» sowie «Reflexion» beziehen. Die 96 Experimentieritems bilden die vollständige Itemstichprobe ⟨E08|69d|V⟩. Um Manipulationsprozesse zu analysieren, wurde die vollständige Stichprobe reduziert auf eine Teilstichprobe ⟨E08|69d|red⟩ mit 67 Items, die Manipulationen mit dem Experimentiermaterial beinhalten oder thematisieren. Um die starken Itemabhängigkeiten in der vollständigen Stichprobe zu vermeiden, wurden aus den abhängigen Items Summenitems (Testlet-Items) gebildet. Daraus geht eine Stichprobe ⟨E08|69d|T⟩ von 52 Testlet-Items hervor (cf. Tab. 7.12 und Abs. 7.2.3).

	vollständige Stichprobe ⟨E08|69d|V⟩	**reduzierte Stichprobe ⟨E08|69d|red⟩**	**Testlet-Stichprobe ⟨E08|69d|T⟩**
nur 6. Schuljahr	19	17	15
6. und 9. Schuljahr	46	27	18
nur 9. Schuljahr	31	23	19
	96	**67**	**52**

Tabelle 7.12 – Itemstichproben für die Analyse von Itemschwierigkeiten

2. Bestimmung der Itemschwierigkeit. Die Itemschwierigkeit wurde für jede Stichprobe separat mit Hilfe einer eindimensionalen Rasch-Analyse mit dem Programm ConQuest 2.0 gemäss Wu et al. (2007) berechnet. Auf eine zusätzliche Itemselektion wurde verzichtet, da die Items allgemein akzeptable Fit-Werte aufweisen.

3. Auswahl der Itemmerkmale. Aus der Fülle des vollständigen Merkmalkatalogs (cf. Tab. 7.11) wurde für jede Itemstichprobe eine separate Auswahl getroffen. Die Selektion der Merkmale erfolgte aufgrund dreier Bedingungen.

- *Zuordnung*: Ein Merkmal sollte in der jeweiligen Stichprobe allen Items sinnvoll zugeordnet werden können, d. h. die in der jeweiligen Beschreibung einer Merkmalsausprä-

gung enthaltenen Präsuppositionen im Sinn von Existenzbehauptungen gemäss Frege (1892) und Russell (1905) müssen für die ganze Itemstichprobe erfüllt sein.[2]

- *Häufigkeit*: Ein Merkmal sollte bei mindestens 8% der Items vorkommen.
- *Korrelation*: Ein Merkmal sollte nicht einer Merkmalkategorie angehören, die mit einer anderen Merkmalkategorie stark korreliert. Für die Überprüfung dieser Bedingung wurden jeweils explorative Faktorenanalysen gerechnet.

4. Erklärung der Itemschwierigkeit. Der Zusammenhang der Itemschwierigkeit mit den ausgewählten Itemmerkmalen wurde mit Hilfe einer multiplen Regressionsanalyse berechnet. Für die Itemschwierigkeit werden die mit der Rasch-Analyse berechneten Itemparameter verwendet. Diese stellen die Kriteriumsvariable dar, die ausgewählten Aufgabenmerkmale (kodiert als dichotome Variablen) die Prädikatoren (cf. Tab. 7.11). Die Reduktion der teilweise sehr umfangreichen Prädikatoren erfolgte in Anlehnung an die Analyse von Prenzel et al. (2002, 131) nach einem zweistufigen, iterativen Verfahren.

1. Reduktion: Zuerst wurde die Regression mit allen ausgewählten Itemmerkmalen gerechnet. Für weitere Rechnungen wurden alle Merkmale eliminiert, deren Standardfehler r den Regressionskoeffizienten B übersteigt ($r \geq B$). Dieser Schritt wurde solange wiederholt, bis alle verbleibenden Merkmale die Bedingung $r < B$ erfüllen.

2. Reduktion: In der zweiten Phasen wurden alle Merkmale eliminiert, die keinen signifikanten Beitrag an der Varianz liefern ($p > 0.05$).

Das Ergebnis einer multiplen Regressionsanalyse kann stark von den Ausgangsmerkmalen und vom Reduktionsverfahren abhängen. Um die Stabilität der Ergebnisse zu kontrollieren, werden daher jeweils mehrere Regressionsanalysen gerechnet, wobei sowohl die Menge der Merkmale als auch das Reduktionsverfahren variiert werden.

7.3.3 Vollständige Itemstichprobe ⟨E08|69d|V⟩

Berechnung der Itemschwierigkeiten

Von der Rasch-Analyse ausgeschlossen wurde nur das Item ⟨N9E21i06⟩, das in der Deutschschweizer Stichprobe von allen Schülerinnen und Schülern vollständig gelöst wurde. Für

[2] Zum Beispiel impliziert die Beschreibung des Merkmals {Man-O-bek} "Die Methode, wie die gesuchten Variablenwerte gemessen werden, ist bekannt", dass überhaupt eine Messung stattfindet. Dieses Merkmal kann daher Items, die keine experimentellen Handlungen erfordern, nicht sinnvoll zugeordnet werden.

die restlichen 95 Items ergab die Analyse akzeptable, teilweise kritische Itemfits: 0.78 $<$ wMNSQ $<$ 1.25; T $<$ 1.7. Bei 12 Items liegt der Fitparameter wMNSQ ausserhalb der empfohlenen Bandbreite von 0.87 $<$ $\text{wMNSQ}_{expected}$ $<$ 1.13. Auf eine Itemselektion aufgrund der Fitparameter wurde verzichtet (cf. App. C.1).

Auswahl der Itemmerkmale

Aus dem vollständigen Itemkatalog mit 54 Merkmalen (cf. Tab. 7.11) wurden 29 Merkmale in den Regressionsanalysen berücksichtigt. 25 Merkmale wurden aufgrund folgender Überlegungen ausgeschieden.

- *Zuordnung*: Die Merkmale der Kategorie *Manipulationen* konnten nicht allen Items zugeordnet werden (nur 67 von 96 Items). Dies betrifft Items zu den Teilprozessen «Frage, Hypothese», «Auswertung» und «Reflexion». Die Zuordnung des Merkmals {Aufgabentyp} ergibt zudem nur Sinn für ganze Aufgaben: Teilitems sind meist zu wenig spezifisch für die differenzierten Aufgabentypen.

- *Häufigkeit*: Wegen geringer Häufigkeit (weniger als 8 Items) wurde bei allen Analysen auf den Kodiermassstab {Kod-PB}, die Teilprozesse {Auf-Prz-Hyp} und {Auf-Prz-Ref} sowie auf 9 von 14 Aufgabenkontexten verzichtet. Wegen ungenügender Häufigkeit wurden die Aufgabenkontexte in der Hauptanalyse nicht verwendet.

- *Korrelation*: Eine explorative Faktorenanalyse mit Varimax-Rotation zeigt eine Korrespondenz zwischen bestimmten Kodiermassstäben und der Merkmalgruppe der Teilprozesse. Merkmale der einen Gruppe laden mit Merkmalen der anderen Gruppe auf dieselben Faktoren. Eine ähnliche Korrespondenz besteht zwischen den Lücken- und Füllformaten. Der inhaltliche Zusammenhang ist jedoch komplexer und weniger offensichtlich, weshalb beide Merkmalkategorien in der Hauptanalyse berücksichtigt wurden.

Erklärung der Itemschwierigkeit

Haupt- und Nebenanalysen. Die Hauptanalyse wurde gemäss dem auf Seite 146 beschriebenen Verfahren durchgeführt. Zur Kontrolle der Stabilität der Ergebnisse wurden drei Nebenanalysen gerechnet. Bei der ersten und zweiten Nebenanalyse wurde der Merkmalkatalog erweitert, und zwar einmal um die fünf Item der Merkmalgruppe {Aufgabenkontext}, welche die minimale Häufigkeit erreichen. Die dritte Nebenanalyse wurde mit dem ursprünglichen Merkmalkatalog gerechnet, wobei die schrittweise Merkmalreduktion gemäss SPSS (Backhausen et al., 2006, 105ff, Bühl, 2010, 408ff) gewählt wurde.

Erweiterung der Hauptanalyse. Im Ergebnis der Hauptanalyse ist die in der Faktorenanalyse auf den Hauptfaktor ladende Merkmalgruppe (Ł-Anz-dic}, {L-Art-Abb}, {F-Art-zei}, {Pro-Off} und {Kod-H} nicht vertreten. Aus zwei der drei Nebenanalysen geht jedoch das Merkmal {L-Art-Abb} als hoch signifikanter Faktor für die Itemschwierigkeit hervor. Die Lösung der Hauptanalyse wurde daher mit dem Merkmal {L-Art-Abb} nachträglich erweitert, wodurch der Anteil der erklärten Varianz bei gleichbleibenden Signifikanzen erhöht werden konnte. Folgend soll dieses "erweiterte" Resultat besprochen werden.

Stabilität der relevanten Merkmale. Die Nebenanalysen geben Aufschluss über die Stabilität der Ergebnisse. Die stabilsten Merkmale {Kod-VI} und {Kod-VL} gehen aus jeder der vier Analysen als signifikanter Faktor hervor. Die weniger stabilen Merkmale erscheinen nicht in jeder Nebenanalyse als signifikanter Faktor. Die Stabilität der Merkmale wird in der Tabelle 7.13 durch die Graustufung dargestellt: Alle Merkmale resultieren in der Hauptanalyse und zusätzlich in einer Nebenanalyse, in zwei Nebenanalysen bzw. in allen drei Nebenanalysen als signifikante Faktoren.

	vollständige Itemstichprobe				**Testlet-Stichprobe**	
	Regression	Fehler	stand. Reg.		Korrelation	
$p < .000$, $R^2 = .444$	B	r	β	**Sig.**	ρ	**Sig.**
Konstante	.763	.314		.017		
{I-Anz-dic}	-.902	.227	-.396	.000	-.419	.002
{C-Art-ABB}	.691	.311	.216	.029	-.036	.800
{T-For-dic}	1.115	.316	.326	.001	.013	.927
{L-Art-LeF}	-1.087	.288	-.449	.000	.064	.650
{L-Art-LeZ}	-.989	.395	-.255	.014	-.131	.356
{L-Art-Abb}	-.748	.316	-.269	.020	-.226	.107
{F-Art-mar}	-.813	.339	-.345	.019	-.049	.732
{F-Art-ben}	-.965	.275	-.388	.001	.034	.809
{Kod-M}	.844	.316	.238	.009	.201	.153
{Kod-PE}	.640	.312	.194	.043	.134	.316
{Kod-VI}	1.121	.292	.369	.000	.142	.316
{Kod-VL}	.786	.233	.332	.001	.226	.108

Tabelle 7.13 – Erweiterte Hauptanalyse der vollständigen Itemstichprobe: Multiple Regressionsanalyse der Itemschwierigkeit gemäss Hauptanalyse mit Korrelation der Testlet-Items. Die Graustufung gibt die Stabilität des Merkmals wieder.

Diskussion

Kausale Interpretationen. Bei der Interpretation der Ergebnisse muss berücksichtigt werden, dass der zugrundeliegende Experimentiertest nicht für den Zweck dieser Analyse entwickelt wurde. Statt den Test gezielt auf die Analyse von Itemmerkmalen auszurichten, wurde versucht, die Items mit gezielten Korrekturen möglichst einfach zu machen. Items, die sich in der Pilotierung bereits als einfach erwiesen, wurden nicht weiter verändert. Die analysierten Itemmerkmale sind auch nicht systematisch über die Itemstichprobe verteilt. Es stellt sich daher die Frage, inwieweit die vorgestellten Ergebnisse Effekte der speziellen Aufgabenentwicklung sind oder allgemeingültige schwierigkeitsrelevante Wirkungen wiedergeben. Um diese Frage zu entscheiden, müssen die Ergebnisse für jedes Merkmal hinsichtlich vier möglicher kausaler Interpretationen untersucht werden:

- *Direkte Interpretation für schwierigkeitserzeugende Merkmale*: Die Items sind schwierig, weil sie ein schwierigkeitserzeugendes Merkmal besitzen.
- *Direkte Interpretation für schwierigkeitsreduzierende Merkmale*: Die Items sind einfach, weil sie ein schwierigkeitsreduzierendes Merkmal besitzen.
- *Indirekte Interpretation für schwierigkeitserzeugende Merkmale*: Die Items besitzen ein schwierigkeitserzeugendes Merkmal, weil ihre übergeordneten Aufgaben leicht sind und bei der Testentwicklung nicht vereinfacht wurden, indem bei den Items dieses Merkmal beseitigt wurde.
- *Indirekte Interpretation für schwierigkeitsreduzierende Merkmale*: Die Items besitzen ein schwierigkeitsreduzierendes Merkmal, weil ihre übergeordneten Aufgaben schwierig sind und bei der Testentwicklung vereinfacht wurden, indem die Items mit diesem Merkmal ausgestattet wurden.

Welche der vier Interpretationen für ein bestimmtes Merkmal plausibel ist, hängt davon ab, ob das Merkmal grundsätzlich als schwierigkeitserzeugend oder schwierigkeitsreduzierend gehalten wird. Dementsprechend wollen wir von E-Merkmalen und von R-Merkmalen sprechen. Die Plausibilität der direkten Interpretation kann anhand des Regressionskoeffizienten β festgestellt werden. Für ein E-Merkmal ist ein positiver Koeffizient plausibel, für ein R-Merkmal ein negativer Koeffizient. Die Plausibilität der indirekten Interpretation versuchen wir anhand der Korrelation ρ des Merkmals mit den Schwierigkeiten der Testlet-Items einzuschätzen: Korreliert ein E-Merkmal negativ mit der Testlet-Schwierigkeit, ist die Interpretation plausibel, dass das Merkmal bei der Testentwicklung nicht beseitigt wurde, weil die Aufgaben mit diesem Merkmal allgemein einfach sind. Korreliert ein R-Merkmal hingegen positiv mit der Testlet-Schwierigkeit, ist die Interpretation plausibel,

	plausible direkte Interpretation	**plausible indirekte Interpretation**
plausibles E-Merkmal	Die Items sind schwierig, weil sie ein E-Merkmal besitzen: $\beta > 0$.	Die Items besitzen ein E-Merkmal, weil die Aufgabenkontexte leicht sind und bei der Testentwicklung nicht vereinfacht wurden, indem bei den Items dieses Merkmal beseitigt wurde: $\rho < 0$.
plausibles R-Merkmal	Die Items sind einfach, weil sie ein R-Merkmal besitzen: $\beta < 0$.	Die Items besitzen ein R-Merkmal, weil die Aufgabenkontexte schwierig sind und bei der Testentwicklung vereinfacht wurden, indem die Items mit diesem Merkmal ausgestattet wurden: $\rho > 0$.

Tabelle 7.14 – Direkte und indirekte Interpretation der Itemschwierigkeit

dass das Merkmal bei der Testentwicklung gezielt eingesetzt wurde, weil die Aufgaben mit diesem Merkmal eher schwierig sind. Es gibt somit für jedes Merkmal zwei Interpretationen, die in der Tabelle 7.14 zusammengestellt sind.

E- und R-Merkmale. Merkmalen lassen sich nicht immer auf eindeutige und klare Weise eine schwierigkeitserzeugende oder -reduzierende Wirkung zuordnen. So kann zum Beispiel der inhaltliche Input mit einer Abbildung ({C-Art-ABB} = 1) verständlicher sein, als wenn er in Form eines Textes ({C-Art-TEX} = 1) gegeben ist. Hingegen können Aufgaben, die gar keinen inhaltlichen Input erfordern ({C-Art-ABB} = 0 und ({C-Art-TEX} = 0)) allgemein einfacher sein als Aufgaben, die ohne zusätzliche Informationen nicht gelöst werden können ({C-Art-ABB} = 1 oder {C-Art-TEX} = 1). Das Merkmal {C-Art-ABB} lässt sich somit sowohl als E-Merkmal als auch als R-Merkmal plausibel begründen.

Keine sinnvolle Zuordnung ist bei Kodiermassstäben möglich. Massstäbe, mit Ausnahme der Vollständigkeitsmassstäbe {Kod-VI} und {Kod-VL}, referieren direkt auf die Prozesse, für deren Beurteilung sie verwendet werden. Die Wahl von Kodiermassstäben ist daher keine Massnahme, um eine Aufgabe einfacher oder schwieriger zu gestalten. Die indirekte Interpretation ist daher für prozessorientierte Kodiermassstäbe auf jeden Fall keine plausible Alternative. Für die Vollständigkeitsmassstäbe {Kod-VI} und {Kod-VL} hingegen ist die Zuordnung wiederum zweideutig. Beide Massstäbe können gleichermassen als E-Merkmale und als R-Merkmale begründet werden. Die Verwendung dieser Massstäbe können auch mit schwierigkeitsreduzierenden Korrekturen der Testaufgaben bei der

Entwicklung in Verbindung gebracht werden. Spezifizierungen der geforderten Antwort, wie sie mit dem Merkmal {Kod-VI} erfasst werden, wurden eingesetzt, um eine Aufgabe verständlicher zu machen. Insofern ist man zumindest bei der Testentwicklung von einer Vereinfachung der Aufgabe ausgegangen. Testpersonen können jedoch auch an solchen zusätzlichen Spezifikationen scheitern. In ähnlicher Weise trifft dies auch auf das Merkmal {Kod-VL} zu. Die Verwendung mehrerer Antwortformate strukturiert das ‹Antwort geben› vor und kann eine Vereinfachung bewirken. Es kann aber auch die zusätzliche Schwierigkeit erzeugt werden, alle Antwortformate konsistent füllen zu müssen. Bei den beiden Vollständigkeitsmassstäben lassen wir daher sowohl die Interpretation als E-Merkmal als auch die Interpretation als R-Merkmal gelten. Alle begründbaren Interpretationen sind in der Tabelle 7.15 zusammengefasst.

Merkmal	**plausible A priori-Interpretationen**		
	Merkmaltyp	**direkt kausal**	**indirekt kausal**
{I-Anz-dic}	E-Merkmal oder R-Merkmal	möglich: $\beta < 0$	möglich: $\rho \leq -0.1$
{C-Art-ABB}	R-Merkmal	möglich: $\beta > 0$	ausgeschlossen: $\mid \rho \mid < 0.1$
{T-For-dic}	E-Merkmal	möglich: $\beta > 0$	ausgeschlossen: $\mid \rho \mid < 0.1$
{L-Art-LeF}	E-Merkmal	möglich: $\beta < 0$	ausgeschlossen: $\mid \rho \mid < 0.1$
{L-Art-LeZ}	E-Merkmal	möglich: $\beta < 0$	möglich: $\rho \leq -0.1$
{L-Art-Abb}	R-Merkmal	möglich: $\beta < 0$	möglich: $\rho \leq -0.1$
{F-Art-mar}	R-Merkmal	möglich: $\beta < 0$	ausgeschlossen: $\mid \rho \mid < 0.1$
{F-Art-ben}	R-Merkmal	möglich: $\beta < 0$	ausgeschlossen: $\mid \rho \mid < 0.1$
{Kod-M}	keine sinnvolle Zuordnung	möglich: $\beta > 0$	nicht plausibel
{Kod-PE}	keine sinnvolle Zuordnung	möglich: $\beta > 0$	nicht plausibel
{Kod-VI}	R-Merkmal oder E-Merkmal	möglich: $\beta > 0$	möglich: $\rho \geq 0.1$
{Kod-VL}	R-Merkmal oder E-Merkmal	möglich: $\beta > 0$	möglich: $\rho \geq 0.1$

Tabelle 7.15 – A priori-Interpretation von R- und E-Merkmalen

Plausibilität der Ergebnisse. Um die Plausibiltät der Ergebnisse zu beurteilen, untersuchen wir, ob sich das Resultat jedes einzelnen Merkmals kausal interpretieren lässt. Dies geschieht anhand der Interpretation des Merkmals als E- oder R-Merkmal, dem Vorzeichen des entsprechenden Regressionskoeffizienten β und dem Vorzeichen der entsprechenden

Korrelation ρ. Die daraus resultierenden acht Fälle

$$\text{(E-Merkmal, R-Merkmal)} \times (\beta < 0, \beta > 0) \times (\rho < 0, \rho > 0)$$

sind in der Tabelle 7.16 durch die grauen und weissen Felder dargestellt. Die weissen Felder zeigen Fälle an, die eine plausibel Interpretation zulassen. Die grauen Felder markieren Fälle, die nicht sinnvoll interpretiert werden können. Die schwarzen Felder entsprechen unmöglichen Kombinationen.

Wie bereits erwähnt, ist die indirekte Interpretation nicht für alle Merkmale eine plausible Alternative. Dies trifft einerseits aus inhaltlichen Überlegungen für prozessorientierte Kodiermassstäbe zu. Für alle anderen Merkmale lässt sich die indirekte Interpretation nur dann begründen, wenn die Korrelation mit den Testlet-Items einen namhaften Wert übersteigt. Als Grenzwert haben wir ad hoc den Betrag von 0.1 festgesetzt. Korrelationen mit Betrag kleiner als 0.1 werden kein eindeutiges Vorzeichen zugeordnet. Dementsprechend wird die indirekte Interpretation für die betroffenen Merkmale als plausible Alternative ausgeschlossen. Die in der Tabelle 7.16 zusammengefassten Interpretationen zeigen ein uneinheitliches Bild. Die Kodiermassstäbe {Kod-M}, {Kod-PE}, {Kod-VI} und {Kod-VL} lassen sich am besten direkt als E-Merkmale interpretieren, wobei bei den Vollständigkeitsmassstäben auch die indirekte Interpretation als R-Merkmal möglich ist. Bei den

		indirekte Interpretation:			
		$\rho > 0$		$\rho < 0$	
direkte Interpretation:		**plausibel für R-Merkmal**	**nicht plausibel für E-Merkmal**	**plausibel für E-Merkmal**	**nicht plausibel für R-Merkmal**
$\beta > 0$	**plausibel für E-Merkmal**		**{Kod-M} *{Kod-PE} ***{Kod-VI} **{Kod-VL}	**{T-For-dic}	
	nicht plausibel für R-Merkmal	***{Kod-VI} **{Kod-VL}			*{C-Art-ABB}
$\beta < 0$	**plausibel für R-Merkmal**	**{F-Art-ben}			***{I-Anz-dic}** *{L-Art-Abb} *{F-Art-mar}
	nicht plausibel für E-Merkmal		***{L-Art-LeF}	***{I-Anz-dic}** *{L-Art-LeZ}	

Tabelle 7.16 – Erweiterte Hauptanalyse mit vollständiger Itemstichprobe: Plausibilität direkter und indirekter Interpretationen. Signifikanzangaben β: **{...}, ρ: {...}*

Lückenformaten fällt auf, dass für die offenen Formate {L-Art-LeF} und {L-Art-LeZ} die Interpretation als E-Merkmal durch die Ergebnisse nicht unterstützt wird. Demgegenüber werden die Füllformate {F-Art-mar} und {F-Art-ben}, die eher auf geschlossene Antwortformate referieren, als R-Merkmale bestätigt. Die zwei extensiven Merkmale {I-Anz-dic} und {T-For-dic}, die mit der Anzahl der in der schriftlichen Aufgabenstellung verwandten Formate anwachsen, erweisen sich aufgrund der Ergebnisse als klare E-Merkmale. Das Merkmal {C-Art-ABB}, das einen inhaltlichen Input in Form einer Abbildung anzeigt, findet entgegen der Intuition als R-Merkmal keine plausible Erklärung.

Nicht alle Ergebnisse lassen sich also plausibel durch eine direkte Wirkungsweise auf die Itemschwierigkeit erklären. Dies legt einerseits die Vermutung nahe, dass die Ergebnisse neben allgemeinen kausalen Zusammenhängen, auch eher zufällige testspezifische Abhängigkeiten sowie Effekte der Itemanalyse enthalten. Letztere Annahme wird durch die im Vergleich zur Grösse der Itemstichprobe beachtliche Anzahl analysierter Itemmerkmale verständlich. Es ist im Rahmen einer Post hoc-Analyse auch zu einem gewissen Grad zufällig, welche Merkmale sich als signifikant herausstellen.

Was misst der HarmoS-Experimentiertest? Die 12 signifikanten Merkmale erklären rund 44% der Varianz, wobei nur vier Merkmale kompetenzrelevante Aspekte der Items erfassen. Zwar musste bei der Analyse mit der vollständigen Stichrobe aus verschiedenen Gründen auf wesentliche kompetenzrelevante Merkmale verzichtet werden, trotzdem ist der Schluss begründet, dass der HarmoS-Experimentiertest zu einem wesentlichen Teil kompetenzirrelevante Fähigkeiten misst. Dies entspricht den generellen Erfahrungen mit Papier-und-Bleistift-Tests, mit welchen eben immer auch Kompetenzen gemessen werden, welche für das, was eigentlich gemessen werden soll, irrelevant sind. Siehe zum Beispiel die Ergebnisse der Analyse des PISA-Ergänzungstests in Naturwissenschaften aus dem Jahre 2000 von Prenzel et al. (2002).

7.3.4 Reduzierte Itemstichprobe ⟨E08|69d|red⟩

Berechnung der Itemschwierigkeit

Die Itemschwierigkeiten wurden für die reduzierte Itemstichprobe nicht neu berechnet, sondern der Rasch-Analyse mit der vollständigen Itemstichprobe entnommen (cf. Abs. 7.3.3).

Auswahl der Itemmerkmale

Von den 54 Merkmalen des vollständigen Katalogs wurden 32 Merkmale in den Regressionsanalysen berücksichtigt. 22 Merkmale wurden aufgrund folgender Überlegungen vor

der Analyse ausgeschieden.

- *Zuordnung*: Die Einschränkung der Itemstichprobe wurde vorgenommen, um die Merkmalgruppe der *Manipulationen* allen Items zuordnen und in der Analyse verwenden zu können. Die Zuordnung des Merkmals {Aufgabentyp} bleibt jedoch auch für die reduzierte Stichprobe zu wenig spezifisch für das einzelne Item. Auf die Verwendung der Aufgabentypen wird daher verzichtet.

- *Häufigkeit*: Aufgrund der Häufigkeit (minimal 7 Items) wurden folgende Merkmale ausgeschieden: Sämtliche {Aufgabenkontexte} (nur drei von 14 Aufgaben weisen mehr als 7 Items auf), {L-Art-LeZ}, {Auf-Prz-Hyp}, {Auf-Prz-Ref}, {Kod-PB}, {Kod-S}.

- *Korrelation*: Die Korrespondenz zwischen der Merkmalgruppe {Kodiermassstäbe} und der Merkmalgruppe {Teilprozesse} ist in der reduzierten Stichprobe nicht ausgeprägt. Die Teilprozesse {Auf-Prz-Pla}, {Auf-Prz-Dur} und {Auf-Prz-Aus} werden daher in die Analyse von Beginn weg eingebunden.

Erklärung der Itemschwierigkeit

Die Modellierung der Itemschwierigkeit mit der reduzierten Stichprobe liefert recht stabile Resultate.

Haupt- und Nebenanalysen. Aus der Hauptanalyse mit 32 Merkmalen gehen acht Merkmale als signifikant hervor. Die erste Nebenanalyse wurde mit drei der häufigsten Aufgabenkontexte {N1E23}, {N6E61}, {N9E84} erweitert. In der zweiten Nebenanalyse wurden die drei Teilprozesse {Auf-Prz-Pla}, {Auf-Prz-Dur}, {Auf-Prz-Aus} mit berücksichtigt. Die dritte Nebenanalyse wurde mit dem ursprünglichen Merkmalkatalog mit schrittweiser Reduktion gemäss SPSS gerechnet.

Erweiterung der Hauptanalyse. Von den 12 Faktoren, die aus einer Faktorenanalyse aller Merkmale hervorgehen, werden im Resultat der Hauptanalyse nur sieben berücksichtigt. Der Versuch, durch Hinzunahme weiterer Merkmale mehr Faktoren abzudecken, ergibt nur für das Merkmal {F-Art-bes} einen signifikanten Beitrag. Die Lösung der Hauptanalyse wurde daher um dieses Merkmal erweitert. Im Folgenden soll dieses "erweiterte" Resultat besprochen werden.

Stabilität der relevanten Merkmale. Die Analyse der reduzierten Itemstichprobe ist weniger stabil als die Analyse mit der vollständigen Stichprobe. Die Stabilität der Merkmale wird in der Tabelle 7.17 durch die Graustufung dargestellt: Alle Merkmale erscheinen in

	reduzierte Itemstichprobe				Testlet-Stichprobe	
	Regression	Fehler	stand. Reg.		Korrelation	
$p < .000$, $R^2 = .470$	B	r	β	**Sig.**	ρ	**Sig.**
Konstante	-.611	.327		.067		
{I-Anz-dic}	-.665	.295	-.253	.028	-.419	.002
{P-For-dic}	-.972	.285	-.396	.001	-.011	.973
{T-For-dic}	1.882	.457	.474	.000	.013	.927
{L-Art-Tab}	.732	.327	.245	.029	.150	.288
{F-Art-bes}	.682	.339	.253	.049	.098	.488
{Kod-E}	.803	.295	.326	.009	-.002	.986
{Kod-M}	.889	.406	.260	.033	.201	.153
{Kod-VI}	.665	.318	.228	.041	.142	.316
{Kod-VL}	.823	.302	.324	.008	.226	.108

Tabelle 7.17 – Erweiterte Hauptanalyse der reduzierten Itemstichprobe: Multiple Regressionsanalyse der Itemschwierigkeit mit Korrelation der Testlet-Items. Die Graustufung gibt die Stabilität des Merkmals wieder.

der Hauptanalyse und zusätzlich in einer Nebenanalyse, in zwei Nebenanalysen bzw. in allen drei Nebenanalysen als signifikante Faktoren.

Diskussion

Die Diskussion der Ergebnisse erfolgt analog zur letzten Diskussion.

R- und E-Merkmale. Entscheidend für die Einschätzung der Plausibilität der Ergebnisse ist die A priori-Interpretation der Merkmale als E- und R-Merkmale. Viele Merkmale lassen bei isolierter Betrachtung beide Interpretationen zu. Im Zusammenhang mit allen Merkmalen erscheint dann meist eine Interpretation plausibler als die andere. Ausgehend von den in der Tabelle 7.18 aufgelisteten A priori-Interpretationen der Merkmale ergeben sich folgende Interpretationen der Resultate (cf. Tab. 7.19). Die Analyse der reduzierten Itemstichprobe bestätigt die Ergebnisse der vollständigen Stichprobe, insofern wir für die fünf Merkmale {I-Anz-dic} (*Anzahl unterschiedliche Formattypen*), {T-For-dic} (*Anzahl Aufgabenformate*), {Kod-M} (*messtechnisch-praktische Korrektheit*), {Kod-VI} (*Vollständigkeit bzgl. der verlangten Inhalte*) und {Kod-VL} (*Vollständigkeit bzgl. der Lückenformate*), die aus beiden Analysen als signifikant hervorgehen, vergleichbare Resultate erhalten und auf dieselben plausiblen Interpretationen schliessen können. Im Gegensatz zur Analyse mit der vollständigen Stichprobe erlauben alle Ergebnisse u. a. die

Merkmal	plausible A priori-Interpretationen		
	Merkmaltyp	**direkt kausal**	**indirekt kausal**
{I-Anz-dic}	E-Merkmal oder R-Merkmal	möglich: ($\beta < 0$)	möglich: ($\rho \leq -0.1$)
{P-For-dic}	R-Merkmal	möglich: ($\beta < 0$)	ausgeschlossen: ($\mid \rho \mid < 0.1$)
{T-For-dic}	E-Merkmal	möglich: ($\beta > 0$)	ausgeschlossen: ($\mid \rho \mid < 0.1$)
{L-Art-Tab}	R-Merkmal	möglich: ($\beta > 0$)	möglich: ($\rho \geq 0.1$)
{F-Art-bes}	E-Merkmal	möglich: ($\beta > 0$)	ausgeschlossen: $\mid \rho \mid < 0.1$
{Kod-E}	keine sinnvolle Zuordnung	möglich: ($\beta > 0$)	nicht plausibel
{Kod-M}	keine sinnvolle Zuordnung	möglich: ($\beta > 0$)	nicht plausibel
{Kod-VI}	R-Merkmal oder E-Merkmal	möglich: ($\beta > 0$)	möglich: $\rho \geq 0.1$
{Kod-VL}	R-Merkmal oder E-Merkmal	möglich: ($\beta > 0$)	möglich: ($\rho \geq 0.1$)

Tabelle 7.18 – A priori-Interpretation von R- und E-Merkmalen

plausible Erklärung durch eine direkte "kausale" Wirkung auf die Itemschwierigkeit. Die Hinzunahme der Manipulations-Merkmale ergibt jedoch keine bessere Modellierung des Aufgabenschritts ‹Problem lösen›. Dieser für die Kompetenzmessung entscheidende Teil wird durch die Analyse nicht erfasst. Die Schwierigkeit der zu lösenden experimentellen Probleme kann post hoc nicht erklärt werden.

Die Ursache dieses auf den ersten Blick ernüchternden Resultats mag in der Heterogenität des Tests liegen. Anscheinend funktionieren die untersuchten Problemmerkmale wie die Offenheit, Vielzieligkeit und Strukturiertheit eines Problems, aber auch die verschiedenen Manipulationsmerkmale in verschiedenen Aufgabentypen unterschiedlich. Problemmerkmale müssen demnach bei jedem Aufgabentyp separat untersucht werden. Die Mischung unterschiedlichster Aufgabentypen beim HarmoS-Experimentiertest mag somit die Post hoc-Analyse von Itemschwierigkeit verunmöglichen.

7.3.5 Testlet-Itemstichprobe ⟨E08|69d|T⟩

Problem der Itemabängigkeit: Testlet-Rekodierung

Der beim HarmoS-Experimentiertest gewählte Ansatz, anhand von integralen Experimentieraufgaben verschiedene Teilprozesse zu evaluieren, verursacht erhebliche Itemabhängigkeiten, welche die Ergebnisse von Rasch-Analysen verfälschen. Um dem Problem der Itemabhängigkeiten zu begegnen, wurden aus abhängigen Items Summenitems, so genannte Testlet-Items gebildet (Wainer, Bradlow & Wang, 2007). Dabei wurden die Summencodes

direkte Interpretation:		indirekte Interpretation: $\rho > 0$: plausibel für R-Merkmal	indirekte Interpretation: $\rho > 0$: nicht plausibel für E-Merkmal	indirekte Interpretation: $\rho < 0$: plausibel für E-Merkmal	indirekte Interpretation: $\rho < 0$: nicht plausibel für R-Merkmal
$\beta > 0$	**plausibel für E-Merkmal**		*{F-Art-bes} **{Kod-E} *{Kod-M} *{Kod-VI} *{Kod-VL}	***{T-For-dic}	
	nicht plausibel für R-Merkmal	*{L-Art-Tab} *{Kod-VI} *{Kod-VL}			
$\beta < 0$	**plausibel für R-Merkmal**				*{I-Anz-dic}** **{P-For-dic}
	nicht plausibel für E-Merkmal			*{I-Anz-dic}**	

Tabelle 7.19 – Regressionsanalyse mit reduzierter Itemstichprobe: Plausibilität direkter und indirekter Interpretationen. Signifikanzangaben β: *{...}, ρ: {...}**

auf maximal vier Codes (Code=0,1,2,3) rekodiert. Die Details der Rekodierung sind im Appendix D zusammengefasst.

Berechnung der Itemschwierigkeiten

Die eindimensionale Rasch-Analyse mit den 52 Testlet-Items generiert gute Itemfit-Werte: $0.81 <$ wMNSQ < 1.20; T < 1.1. Lediglich bei vier Items liegt der Fitparameter ausserhalb der empfohlenen Bandbreite von $0.87 <$ wMNSQ$_{expected} < 1.13$. Aufgrund der geringen Itemzahl wurde auf die Eliminierung des einzigen kritischen Items ⟨N6E44i05T⟩ (wMNSQ = 0.81, T = -1.6) verzichtet (cf. App. D.2).

Auswahl der Itemmerkmale

Ausgehend vom vollständigen Merkmalkatalog mit 54 Merkmalen (cf. Tab. 7.11) wurden 48 Merkmale in den Regressionsanalysen berücksichtigt. Sechs Merkmale sind aufgrund folgender Überlegungen ausgeschieden.

- *Zuordnung*: Merkmale der Kategorie *Manipulationen* können bis auf zwei Items allen Testlet-Items zugeordnet werden. Diese Merkmalgruppe wird in die Analyse aufgenommen.

- *Häufigkeit*: Aufgrund der geringen Häufigkeit wurde auf die Teilprozesse {Auf-Prz-Hyp} und {Auf-Prz-Ref} und den Aufgabentyp {Auf-Typ-Ver} verzichtet. Aus dem gleichen Grund fallen die Merkmale des Aufgabenkontextes weg. Die meisten Experimentieraufgaben (Kontexte) enthalten nur zwei bis drei Testlet-Items. Die Differenzierung einzelner Kontexte ergibt keinen Sinn.

- *Korrelation*: Aus der explorativen Faktorenanalyse mit Varimax-Rotation geht wie bereits bei der vollständigen Itemstichprobe eine eins-zu eins-Korrespondenz zwischen den drei übrigbleibenden Prozessen {Auf-Prz-Pla}, {Auf-Prz-Dur} und {Auf-Prz-Aus} mit den Kodiermassstäben {Kod-H}, {Kod-S} und {Kod-M} hervor. Auf die Verwendung der Prozessmerkmale wird daher zugunsten der Kodiermassstäbe verzichtet.

Erklärung der Itemschwierigkeit

Die Hauptanalyse mit dem Grundkatalog an Merkmalen ergibt eine Lösung mit 23 signifikanten Merkmalen. Das sind nahezu halb so viele Merkmale wie Items, die mit den Merkmalen modelliert werden. Aufgrund dieses Verhältnisses sind die Ergebnisse statistisch nicht aussagekräftig. Weshalb wir diese Analyse nicht werten. Aus der Analyse mit integrierten Teilprozessmerkmalen resultiert kein signifikantes Merkmal! Dasselbe Ergebnis liefern die beiden letzten Nebenanalysen mit dem Einschluss der Kontextmerkmale und der schrittweisen Reduktionsmethode. Bei der Modellierung der Itemschwierigkeit für die Testlet-Stichprobe resultieren weniger stabile Faktoren als bei der Modellierung der vollständigen und der reduzierten Stichprobe.

	Testlet-Stichprobe			
	Regression	Fehler	stand. Reg.	
$p < .000$, $R^2 = .318$	B	r	β	**Sig.**
Konstante	1.323	.442		.004
{I-Anz-dic}	-1.930	.468	-.505	.000
{Kod-VI}	.582	.274	.259	.039
{Kod-VL}	.629	.247	.308	.014

Tabelle 7.20 – Nebenanalyse der Testlet-Stichprobe: Multiple Regressionsanalyse der Itemschwierigkeit. Die Graustufung gibt die Stabilität des Merkmals wieder.

Diskussion

Die Analyse der Testlet-Items bestätigt für die signifikanten Merkmale {I-Anz-dic} (*Anzahl unterschiedliche Formattypen*), {Kod-VI} (*Vollständigkeit bzgl. der verlangten Inhal-*

te) und {Kod-VL} (V*ollständigkeit bzgl. der Lückenformate*) die beiden vorangegangenen Analysen. Alle drei Merkmale lassen sich a priori sowohl als R- wie auch als E-Merkmale plausibel begründen. Die Interpretation der Ergebnisse als direkte Wirkungen der Merkmale auf die Itemschwierigkeit erscheint durch diese Analyse bestätigt, die indirekte Wirkungsweise lässt sich auf der Ebene der Testlet-Items weder gut plausibel begründen noch statistisch überprüfen. Wir interpretieren die Ergebnisse dahingehend, dass das Merkmal {I-Anz-dic} schwierigkeitsvermindernd und die Vollständigkeitsmassstäbe {Kod-VI} und {Kod-VL} schwierigkeitserzeugend wirken. Das Resultat für {I-Anz-dic} steht im Widerspruch zur ursprünglichen Interpretation als Kohärenzindikator, der, wie auf der Seite 134 begründet wurde, als schwierigkeitserzeugend angenommen wurde. Möglich ist, dass dieses Merkmal ebenfalls als Vollständigkeitsindikator wirkt im Sinne der Vollständigkeit der Aufgabenstellung. Das Merkmal {I-Anz-dic} nimmt dann hohe Werte an, wenn alle relevanten Formattypen in der Aufgabenstellung vorkommen. Das ist eher bei längeren Testlet-Aufgaben als bei kurzen Einzelitems der Fall. Es gilt abschliessend zu beachten, dass alle drei signifikanten Merkmale zusammen bereits rund 32% der Varianz erklären!

7.3.6 Zusammenfassung

Signifikante Itemmerkmale. Mit den Analysen konnten verschiedene Itemmerkmale eruiert werden, welche die Itemschwierigkeit signifikant beeinflussen. Aufgrund der speziellen Testkonstruktion ist nicht in jedem Fall klar, wie die Resultate zu interpretieren sind, i. e. ob sie in Verbindung gebracht werden können mit Korrekturen, mit welchen die Aufgaben in der Pilotierungsphase vereinfacht wurden. Für zwei Ergebnisse ({L-Art-LeF}, {C-Art-ABB}) fehlen sogar plausible Interpretationen. Zumindest bei drei Merkmalen ({I-Anz-dic}, {Kod-VI}, {Kod-VL}) weisen die Resultate deutlich auf eine direkte Wirkungsweise des Merkmals auf die Itemschwierigkeit hin. Der Bezug zu bestimmten Korrekturmassnahmen kann jedoch nur vage hergestellt werden.

Modellierung der Itemschwierigkeit. Der gewählte Modellierungsansatz mit 17 Merkmalkategorien, verteilt auf die vier Arbeitsschritte ‹Aufgabe erfassen›, ‹Problem lösen›, ‹Antwort geben› und ‹Lösung kodieren› erweist sich mit Einschränkung als erfolgreich. In 9 Merkmalkategorien konnten relevante Itemmerkmale ausfindig gemacht werden. Und es wurde jeweils zwischen 32% und 47% der Varianz erklärt. Alle signifikanten Merkmale sind in der Tabelle 7.21 zusammengefasst. Die Grauschattierung gibt an, in wie vielen Analysen ein Merkmal als signifikant resultiert: Ein Merkmal ist demnach signifikant in einer Analyse, in zwei Analysen bzw. in drei Analysen.

kompetenzirrelevante Itemmerkmale		**kompetenzrelevante Itemmerkmale**	
‹Aufgabe erfassen›	**‹Antwort geben›**	**‹Problem lösen›**	**‹Lösung kodieren›**
SPRACHE	LÜCKENFORMATE	PROBLEM	KORREKTHEIT
Textlänge:	{L-Art-LeF}	*Teilprozesse*:	{Kod-E}
{I-Wör-dic}	{L-Art-LeZ}	{Auf-Prz-Hyp}	{Kod-H}
	{L-Art-Tab}	{Auf-Prz-Pla}	{Kod-M}
Textkohärenz:	{L-Art-Abb}	{Auf-Prz-Dur}	{Kod-S}
{I-Anz-dic}	{L-Art-MuCS}	{Auf-Prz-Aus}	{Kod-T}
{I-Sep-dic}		{Auf-Prz-Ref}	
Satzstruktur:		*Aufgabenumfang*:	
{IDTei-dic}		{Auf-Umf-dic}	
		Aufgabentyp:	
		{Auf-Typ-Mes}	
		{Auf-Typ-Beo}	
		{Auf-Typ-Ver}	
		{Auf-Typ-Unt}	
		{Auf-Typ-Her}	
		Aufgabenkontext:	
		{N1E13}, {N9E21}	
		{N1E22}, {N1E23}	
		{N9E41}, {N9E42}	
		{N1E43}, {N6E44}	
		{N1E53}, {N9E54}	
		{N6E56}, {N6E61}	
		{N9E83}, {N9E84}	
INPUT INHALTE	FÜLLFORMATE	LÖSUNGSWEG	PRÄZISION
{C-Art-TEX}	{F-Art-mar}	{Pro-Off}	{Kod-PE}
{C-Art-ABB}	{F-Art-zei}	{Pro-Zie}	{Kod-PB}
	{F-Art-ben}	{SrFor-dic}	
	{F-Art-bes}		
PROBLEM-BESCHREIBUNG	AUFGABEN-BESCHREIBUNG	MANIPULA-TIONEN	VOLLSTÄNDIG-KEIT
{P-For-dic}	{T-For-dic}	{Man-I-bek}	{Kod-VL}
		{Man-R-bek}	{Kod-VI}
		{Man-R-dir}	
		{Man-O-bek}	
		{Man-O-dir}	

Tabelle 7.21 – Zusammenzug aller signifikanten schwierigkeitsrelevanten Itemmerkmale.

Sensitivität des HarmoS-Experimentiertests. Die drei gerechneten Ansätze erklären zwischen 32% bis 47% der Varianz. Dieser Anteil wird jeweils zur guten Hälfte mit kompetenzirrelevanten Itemmerkmalen erreicht. Der HarmoS-Experimentiertest misst somit neben der experimentellen Kompetenz auch zu einem wesentlichen Teil kompetenzirrelevante Fähigkeiten, die unter den groben Titeln der Lese- und Schreibkompetenz subsummiert werden können. Zum kompetenzrelevanten Bereich ‹Problem lösen› liefern die Analysen zudem kein signifikantes Ergebnis. Zwar muss berücksichtigt werden, dass die Merkmale dieses Arbeitsschritts nur eingeschränkt angewandt werden konnten und teilweise auch durch die Kodiermassstäbe erfasst wurden. Der Schluss erscheint uns jedoch begründet, dass der Experimentiertest als Messinstrument der experimentellen Kompetenz bei den kompetenzirrelevanten Arbeitsschritten ‹Aufgabe erfassen› und ‹Antwort geben› zu sensitiv ist, bei den kompetenzrelevanten Arbeitsschritten ‹Problem lösen› und ‹Lösung kodieren› hingegen zu wenig sensitiv ist.

Kapitel 8

Personen-Analysen

8.1 Fragestellung

Fragestellung. Im Rahmen des Experimentiertests wurden mit dem im Appendix F abgedruckten Fragebogen ⟨Q08|69dfi⟩ persönliche Daten von den Schülerinnen und Schüler erhoben. Die Daten umfassen Angaben zur Person, zu Interessen und Einstellungen sowie zum häuslichen Umfeld und der Familie der Schülerinnen und Schüler. Die Datenerhebung erfolgte anonymisiert, wobei zumindest für die 9. Schulstufe die Ergebnisse des Experimentiertests und die Antworten des Fragebogens einander zugeordnet werden können. Bei der 6. Schulstufe wurde auf die Zuordnung bereits beim Erfassen verzichtet, da die Tests in jeder Klasse an mindestens zwei verschiedenen Halbtagen durchgeführt wurden. Die erhobenen Daten erlauben für diese Teilstichprobe nicht, Zusammenhänge zwischen individuellen Variablen und der Testleistung zu untersuchen. Konkret ist dieses Kapitel folgender allgemeinen Forschungsfrage gewidmet:

3.1. Welche Unterschiede bestehen in der Ausprägung des Interesses an Naturwissenschaften (Fach- und Sachinteressen) zwischen den Geschlechtern, Schulstufen, Sprachregionen und schulischen Anforderungsniveaus?

Repräsentativität. Die Personenstichprobe umfasst Schülerinnen und Schüler aus zehn Kantonen und drei Sprachregionen (cf. Tab. 8.1). Die Kantone und Sprachregionen sind in den beiden Schulstufen unterschiedlich stark vertreten. Weder die Auswahl der Kantone innerhalb der Sprachregionen noch die Auswahl der Kantone für die ganze Schweiz sind repräsentativ. Die Verallgemeinerung von Ergebnissen aufgrund der vorliegenden Personenstichprobe ist daher nur unter bestimmten, im Einzelfall zu diskutierenden Annahmen zulässig. Dabei können zwei Fälle auftreten.

Fall 1: Ergibt die Analyse eines Merkmals einen signifikanten Unterschied zwischen

	D-CH						F-CH			I-CH	
Kantone	**BE(d)**	**SO**	**ZH**	**SH**	**LU**	**ZG**	**GE**	**NE**	**VD**	**TI**	**CH**
6. Schuljahr	188	18	54	21	40	17	120	118	0	0	**576**
9. Schuljahr	241	17	142	0	0	0	0	0	277	128	**805**
beide Schulstufen	429	35	196	21	40	17	120	118	277	128	**1381**
			738					515			

Tabelle 8.1 – Rücklauf des Fragebogens ⟨Q08|69dfi⟩. Abkürzungen der Sprachregionen: D-CH = deutschsprachige Schweiz, F-CH = französischsprachige Schweiz, I-CH = italienischsprachige Schweiz; Abkürzungen der Kantone: BE = Bern, SO = Solothurn, ZH = Zürich, SH = Schaffhausen, LU = Luzern, ZG = Zug, GE = Genf, NE = Neuenburg, VD = Waadt, TI = Tessin.

den Sprachregionen, gilt es zu beurteilen, inwiefern das Ergebnis die wahren Verhältnisse der Sprachregionen abbildet bzw. inwiefern ein Artefakt der nicht repräsentativen Stichprobe vorliegt. Ersteres ist umso wahrscheinlicher, desto eher ein Einfluss der reinen Kantonszugehörigkeit (z. B. via Bildungssystem) auf das untersuchte Merkmal aufgrund bekannter vergleichbarer Untersuchungen ausgeschlossen werden kann. Letzteres ist umso wahrscheinlicher, je mehr die vorliegenden Daten Merkmalsunterschiede zwischen den Kantonen innerhalb einer Sprachregion aufweisen. Die Verallgemeinerung der Ergebnisse auf die Sprachregionen ist also zulässig, wenn die Verallgemeinerungsannahme

V1 Innerhalb einer Sprachregion hat die Kantonszugehörigkeit keinen direkten oder indirekten Einfluss (e. g. via das Bildungssystem) auf das untersuchte Merkmal.

haltbar ist. Dies soll für die nachfolgende Untersuchung hinreichend als gegeben gewertet werden, wenn die Annahme nicht empirischen Befunden widerspricht.[1]

Fall 2: Zeigt die Gesamtstichprobe zu einem Merkmal einen signifikanten Unterschied zwischen den Geschlechtern oder der Schulstufe, muss analog zum Fall 1 beurteilt werden, inwiefern das Ergebnis die wahren Verhältnisse zwischen den Geschlechtern oder den Schulstufen abbildet bzw. inwiefern das Ergebnis ein Artefakt der nicht repräsentativen Stichprobe ist. Ersteres ist umso wahrscheinlicher, je eher ein Einfluss der Kantonszugehörigkeit sowie der Sprachregion aufgrund bekannter empirischer Untersuchungen ausgeschlossen werden kann. Letzteres ist hingegen umso wahrscheinlicher, je mehr die vorliegenden Daten das Gegenteil beweisen. Die Verallgemeinerung der Ergebnisse auf die

[1] Dies stellt die schwächste Basis dar, um Ergebnisse einer nicht repräsentativen Stichprobe zu verallgemeinern. Im konkreten Fall bedeutet das, dass sprachregionale Unterschiede nur dann als wahrscheinlich akzeptiert werden, wenn der Einfluss der Sprachregion via kantonale Bildungssysteme mehr oder weniger ausgeschlossen werden kann.

ganze Schweiz ist demnach nur zulässig, wenn nebst der Verallgemeinerungsannahme V1 zusätzlich auch die Annahme

V2 Die Zugehörigkeit zu einer Sprachregion hat keinen direkten oder indirekten Einfluss (e. g. via sprachspezifische Denkweisen oder Werthaltungen) auf das untersuchte individuelle Merkmal bzw. den untersuchten Zusammenhang.

haltbar ist. Analog zum Fall 1 soll dies hinreichend gegeben sein, wenn die Annahmen nicht empirischen Befunden widersprechen.

Analysen. Die Auswertung des Fragebogens erfolgt zunächst mit einer Analyse der zugrundeliegenden Subskalenstruktur im Abschnitt 8.2. Im Abschnitte 8.3) werden die Zusammenhänge der Subskalen analysiert und signifikante Mittelwertunterschiede eruiert.

8.2 Personen-Variablen

Entwicklung des Fragebogens. Die im 6. und im 9. Schuljahr verwandten Items sind im wesentlichen identisch. Unterschiede gibt es lediglich bei den Fragen nach dem Interesse für bestimmte Schulfächer. In dieser Hinsicht bestehen auch Unterschiede zwischen den Sprachregionen. Während in der Deutschschweiz und im Tessin auf der Sekundarstufe I die Naturwissenschaften integriert unterrichtet werden, kennt die Waadt nur den separierten Naturwissenschaftsunterricht. Bezüglich dieser und weiterer Differenzen wurden in den Fragebogen für die verschiedenen Stufen und Schulsysteme zutreffende Items entsprechend angepasst.

Der vollständige Fragebogen umfasst 49 Items verteilt auf sieben thematische Blöcke (der deutschsprachige Fragebogen, der im 9. Schuljahr verwendet wurde, ist im Appendix F abgedruckt). Die Blöcke betreffen Fragen zur «Person und Schule» ⟨i01-i04,i11⟩[2], zum «Zuhause» ⟨i21-i28⟩[3], zur «Familie» ⟨i31-i37⟩[4], zu schulischen «Fachinteressen» ⟨i41-i47⟩, zur «Schulleistung» ⟨i51⟩[5], zu «Einstellungen» zum HarmoS-Test ⟨i61-i67⟩[6] und zu «Sachinteressen» im Bereich Naturwissenschaften ⟨i71-i77⟩[7]. Unterschiede zwischen den Schulstufen und Sprachregionen betreffen die Itemgruppe ⟨i43.1-i43.9⟩. Während auf der

[2]Dieser Frageblock wurde vom Fragebogen des HarmoS-Validierungstests 2007 unverändert übernommen.

[3]ibidem

[4]Dieser Frageblock wurde dem Dissertationsprojekt von Adamina (2008) entnommen.

[5]Diese Items wurden dem Fragebogen des Nationalen Validierungstests 2007 entnommen und für den Zweck des Experimentiertests angepasst.

[6]Die Items stammen aus dem Fragebogen des HarmoS-Validierungstest 2007 sowie aus Adamina (2008).

[7]ibidem

6. Schulstufe keines dieser Items eingesetzt wurde, wurden auf der 9. Schulstufe in der Deutschschweiz und im Tessin nur die ersten fünf Item ⟨i43.1-i43.5⟩ und in der Romandie nur die letzten vier Item ⟨i43.6-i43.9⟩ eingesetzt.

Analyse der Subskalen. Die Analyse der Subskalen des Fragebogens erfolgte analog der bei Hoffmann et al. (1998, 17f) beschriebenen Methode. Die Faktorenanalyse mit dem Gros der Items (alle Items ausschliesslich der Itemgruppe ⟨i43.1-i43.9⟩) ergab, dass die Itemblöcke statistisch unabhängige Untergruppen bilden. Für die Bildung der Subskalen wurde daher mit jedem Itemblock separat eine explorative Hauptkomponentenanalyse mit Varimax-Rotation gerechnet. Ein Item wurde jeweils nur dann einer Subskala (Faktor) zugeordnet, wenn einerseits in konsistenter Weise für die zwei Schulstufen eine hohe Faktorladung (in der Regel ≥ 0.5) resultierte und andererseits die Ladungen für die anderen Faktoren deutlich geringer ausfielen (in der Regel < 0.3). Bei den Interessenitems wurde zudem darauf geachtet, dass sich die Subskalen auch innerhalb der Geschlechter in konsistenter Weise zeigen. Items, die diese Kriterien nicht erfüllten, wurden gestrichen oder als Einerskala (Variable) übernommen. Die so entstandenen Subskalen wurden weiter optimiert, indem Items mit geringer Trennschärfe innerhalb der Subskala eliminiert wurden, wenn sich dadurch die Reliabilität (Cronbachs α) der Skala erhöhte.

Aus diesem Verfahren resultierten insgesamt sieben Subskalen, die im Weiteren als Faktoren bezeichnet werden und in den nachfolgenden Abschnitten vorgestellt werden. Darüber hinaus werden 12 zusätzliche Variablen definiert, die für spezifische Auswertungen verwendet werden. Diese Variablen sind Einzelitems oder Konstrukte, die aus einzelnen oder mehreren Items abgeleitet wurden, die als solche aber keine Faktoren bilden. Die definierten Faktoren und Variablen werden entweder für die Interessenanalysen in diesem Kapitel oder für erweiterte Kompetenzanalysen im nachfolgenden Kapitel verwendet.

8.2.1 Variablen zur Schule und Sprache

Schule. Von den fünf Items ⟨i01-i04, i11⟩ wurde nur das Item zum Alter ⟨i03⟩ nicht ausgewertet. Aus den Angaben zum Schulort ⟨i01⟩ und zur Klasse ⟨i02⟩ wurden die **Variable 1 {Sprachregion}** und für Klassen des 9. Schuljahrs zusätzlich die dreistufige, in der Tabelle 8.2 beschriebene Variable 2 {Anforderungsniveau} abgeleitet.

Sprache. Obwohl das Sprachitem ⟨i11⟩ "*Welche Sprache sprichst du zuhause am meisten?*" aufgrund der Formulierung nur die Angabe einer Sprache impliziert, wurde dieses Item häufig mit Mehrfachantworten beantwortet. Um der Antwortvielfalt Rechnung zu tragen, wurden aus diesem Item zwei dichotome Variablen extrahiert. Die Variable {Unterrichtssprache} beschreibt, ob im Alltag zuhause u. a. auch die Unterrichtssprache ge-

	BE(d)	ZH	SO	VD	TI
{AnfNiv}=3	Gymnasialer Unterricht	Langzeit-gymnasium	*keine Klasse*	voie secondaire de baccalaureat	*keine Zuordnung*
{AnfNiv}=2	Sekundar-schule	Sekundar-stufe A	*keine Klasse*	voie secondaire générale	*keine Zuordnung*
{AnfNiv}=1	Realschule	Sekundar-stufe B/C	Sekundar-schule	voie secondaire options	*keine Zuordnung*

Tabelle 8.2 – Variable 2: Anforderungsniveau auf der Sekundarstufe I {AnfNiv}

sprochen wird. Die Variable {Mehrsprachigkeit} hält fest, ob zuhause mehrere Sprachen gesprochen werden.

{UntSpra}=1	Die Schülerin/der Schüler gibt auf die Frage ⟨i11⟩ die Unterrichtssprache an.
{UntSpra}=0	Die Schülerin/der Schüler gibt auf die Frage ⟨i11⟩ die Unterrichtssprache nicht an.

Variable 3: Unterrichtssprache {UntSpra}

8.2.2 Variablen zur Familie

Bildungsbewusstsein, Lernbedingungen und materielle Ausstattung. Mit den acht Fragen des Itemblocks «Zuhause» ⟨i21-i28⟩ wird erfasst, inwieweit das Elternhaus ein Umfeld bietet, welches das Lernen für die Schule und das Verrichten von Schularbeiten positiv unterstützt. Unterschieden wird zwischen den {häuslichen Lernbedingung} ⟨i22,i23,i25,i26⟩ und dem {familiären Bildungsbewusstsein} ⟨i27-i28⟩. Nicht berücksichtigt werden die Items ⟨i21⟩ und ⟨i24⟩. Eine mit allen sieben Items für jeweils jede Schulstufe separat gerechnete Faktorenanalysn legen trotz schwacher Reliabilität die Bildung des Faktors {famliliäres Bildungsbewusstsein} nahe.

{MehrSpra}=1	Die Schülerin/der Schüler gibt auf die Frage ⟨i11⟩ mehr als eine Sprache an.
{MehrSpra}=0	Die Schülerin/der Schüler gibt auf die Frage ⟨i11⟩ eine Sprache an.

Variable 4: Mehrsprachigkeit {MehrSpra}

Faktor 1: Familiäres Bildungsbewusstsein {BildBew} ($\alpha = .467, N = 1381$)

⟨i27⟩ Hast du bei dir zu Hause klassische Literatur?

⟨i28⟩ Hast du bei dir zu Hause Kunstwerke?

Bildungsorientierung und Interesse an gesellschaftlichen Belangen. Die sieben Item des Blocks «Familie» ⟨i31-i37⟩ laden sowohl über beide Schulstufen als auch über jeweils eine Schulstufe betrachtet auf die zwei Faktoren {Elterliche Bildungsorientierung} ⟨i31-i35⟩ und {Interesse an gesellschaftlichen Belangen} ⟨i36-i37⟩. Die Items einer Subskala laden jeweils mit Faktorladungen grösser als 0.5 auf den eigenen und mit geringen Ladungen (< 0.3) auf den anderen Faktor.

Faktor 2: Elterliche Bildungsorientierung {BildOr} ($\alpha = .665, N = 1381$)

⟨i31⟩ Wir sprechen zu Hause über Themen aus der Schule.

⟨i32⟩ Wir gehen mit der Familie oft nach draussen.

⟨i33⟩ Wir besuchen mit der Familie Museen, Lehrpfade usw.

⟨i34⟩ Wenn wir in den Ferien sind, schauen wir uns Sachen an diesem Ort an und unternehmen dort Ausflüge.

⟨i35⟩ Ich gehe in Bibliotheken, Mediotheken und sehe mir Bücher und Filme zu Themen an oder leihe sie aus.

Faktor 3: Interesse an gesellschaftlichen Belangen {GesInt} ($\alpha = .528, N = 1381$)

⟨i36⟩ Ich sehe im Fernsehen Nachrichtensendungen.

⟨i37⟩ Ich lese Tageszeitungen.

8.2.3 Variablen zu Fachinteressen

Fachinteresse versus Sachinteresse. Die Schülerinnen und Schüler wurden nach zwei verschiedenen Interessen befragt: zum einem nach dem Interesse an den einzelnen Schulfächern, zum anderen, eingeschränkt auf die Naturwissenschaften, nach dem Interesse an bestimmten Inhalten. Ersteres ist das so genannte Fachinteresse, das sich auf den Unterricht und einen schulischen Kontext beschränkt. Letzteres wird als Sachinteresse bezeichnet, das sich auch in einem ausserschulischen Kontext äussern kann.[8] Im Folgenden

[8]Die Unterscheidung der zwei Interessentypen erfolgt in Anlehnung an die Interessenstudie von Hoffmann et al. (1998). Im Gegensatz zu Hoffmann et al. (1998) wird hier mit den Variablen zu den

werden beide Interessentypen separat behandelt.

Fachinteressen. Eine Faktorenanalyse über alle Items zu Fachinteressen ⟨i41-i47⟩ legt die Bildung einer Subskala bestehend aus den Items zum Interesse am Bildnerischen Gestalten/Zeichen und zum Interesse am Technischen/Textilen Gestalten nahe, die wir jedoch nicht weiter verwenden werden. Beide Items laden auf einen Faktor mit Ladungen grösser als .78 sowohl in stufenübergreifender als auch auf jeweils eine Stufe reduzierte Betrachtung. Aufgrund der geringen Antworthäufigkeit wurden die Items ⟨i24-i28⟩ nicht weiter verwandt. Damit übernimmt das Item ⟨i43⟩, als Einziges die Funktion, das Interesse an naturwissenschaftlichen Fächern zu messen. Die restlichen Items werden als Variablen behandelt. Alle Items haben eine vierstufige Likert-Antwortskala.

Um das generelle Interesse an der Schule zu modellieren, wurden die Summenscores aller Fachinteressen errechnet. Die dazugehörige Variable wird mit Schulinteresse {SchuInt} bezeichnet.

Variable 5: Interesse an naturwissenschaftlichen Fächern {IntFachNat}

⟨i43⟩ Wie gross ist dein Interesse am Fach Natur-Mensch-Mitwelt / Mensch und Umwelt?

8.2.4 Variablen zu Sachinteressen

Sachinteressen. In einer Faktorenanalyse zerfallen die sieben Items zu Sachinteressen in zwei Gruppen, eine mit sechs und eine mit einem Item. Die erste Gruppe, bestehend aus den Items ⟨i71-75,i77⟩, erfasst ein verallgemeinertes Interesse an naturwissenschaftlichen Themen, die den klassischen Wissenschaften Biologie, Chemie, Physik bzw. Technik zugeordnet werden können. Das Item ⟨i76⟩ erfasst das spezifische Interesse an Themen, die mit dem eigenen Körper, dem Körpergefühl sowie der Lebensweise zusammenhängen. Für die Auswertungen wird diese Aufteilung der Items als zwei Faktoren übernommen, wobei auch die einzelnen Sachinteressen für sich als Variablen weiter verwendet werden.

Sachinteressen nur eine einfache Facette erhoben.

Faktor 4: Allgemeines naturwissenschaftliches Sachinteresse {IntSachNat} ($\alpha = .727, N = 1381$)

⟨i71⟩ Mich interessieren Themen zur Erde (Steine, Wasser, Wetter ...).

⟨i72⟩ Mich interessieren Themen zum Universum (Sonne, Mond, Planeten, Sterne).

⟨i73⟩ Mich interessieren Themen zu Tieren, Pflanzen und Lebensräumen.

⟨i74⟩ Mich interessieren Themen zu Energie, Elektrizität, Maschinen und Geräten.

⟨i75⟩ Mich interessieren Themen zu Materialien (Holz, Steine, Metall, Stoffe usw.).

⟨i77⟩ Mich interessieren Themen zu unserer Umwelt, wie sich die Umgebung verändert.

Faktor 5: Spezifisches Sachinteresse am Thema Gesundheit {IntSachGes}

⟨i76⟩ Mich interessieren Themen zu Gesundheit, Essen und Trinken, Bewegung.

Variable 6: Interesse am Thema Erde {IntSachErd}

⟨i71⟩ Mich interessieren Themen zur Erde (Steine, Wasser, Wetter ...).

Variable 7: Interesse am Thema Weltall {IntSachAll}

⟨i72⟩ Mich interessieren Themen zum Universum (Sonne, Mond, Planeten, Sterne).

Variable 8: Interesse am Thema Lebewesen und Lebensräume {IntSachLeb}

⟨i73⟩ Mich interessieren Themen zu Tieren, Pflanzen und Lebensräumen.

Variable 9: Interesse am Thema Technik {IntSachTech}

⟨i74⟩ Mich interessieren Themen zu Energie, Elektrizität, Maschinen und Geräten.

Variable 10: Interesse am Thema Stoffe {IntSachSto}

⟨i75⟩ Mich interessieren Themen zu Materialien (Holz, Steine, Metall, Stoffe usw.).

Variable 11: Interesse am Thema Umwelt *IntSachUmw*

⟨i77⟩ Mich interessieren Themen zu unserer Umwelt, wie sich die Umgebung verändert.

8.2.5 Variablen zur persönlichen Einstellung und Einschätzung des Tests

Interesse und Einschätzungen in Bezug auf den Test. Mit dem Frageblock ⟨i61-i67⟩ werden Einstellungen erfasst, die das Interesse am Test und die Einschätzung der Testschwierigkeit betreffen. Faktorenanalysen, gerechnet sowohl über die einzelnen Schulstufen als auch über die Geschlechter, legen zwei Subskalen mit zwei bzw. fünf Items nahe. Um die Reliabilität zu optimieren, wurde die grössere Skala um das Item ⟨i64⟩ reduziert. Die reduzierte Subskala {Testinteresse} erfasst das Interesse an den Testaufgaben. Das Item ⟨i64⟩ wird als Einerskala weiterverwendet und erfasst die Relevanz, die dem Test beigemessen wird. Die Subskala {Einschätzung Testschwierigkeit} enthält zwei entgegengesetzte Items, weshalb bei der Berechnung des Subskalen-Summenscores die Differenz der beiden Scores berechnet wird.

Variable 12: Einschätzung Testrelevanz {TestRel}

⟨i64⟩ Es ist wichtig, dass ich solche Aufgaben lösen kann.

Faktor 6: Interesse an den Testinhalten {TestInt} ($\alpha = .735, N = 1381$)

⟨i61⟩ Diese Aufgaben haben mich interessiert.

⟨i65⟩ Diese Aufgaben waren für mich langweilig. (mit invertiertem Score)

⟨i66⟩ Bei diesen Aufgaben habe ich Neues gelernt.

⟨i67⟩ Diese Aufgaben finde ich spannend.

Faktor 7: Einschätzung Testschwierigkeit {TestSchw} ($\alpha = .479, N = 1381$)

⟨i62⟩ Zu diesen Aufgaben wusste ich schon viel. (mit invertiertem Score)

⟨i63⟩ Diese Aufgaben waren für mich schwierig.

8.3 Personen-Analysen: Interesse und Geschlecht

Methoden. In diesem Abschnitt werden Analysen der Fach- und Sachinteressen im Hinblick auf Geschlechterdifferenzen und Varianzen bezüglich Schulstufen und Anforderungsniveaus vorgestellt. Es gilt hier nochmals zu betonen, dass die Daten des Fragebogens nicht repräsentativ für die Schweiz bzw. die Sprachregionen sind.

Die zu untersuchenden Interessenausprägungen in den jeweiligen Teilstichproben (Kan-

tone, Sprachregionen, Schulstufen und Geschlechter) sind durchwegs nicht normalverteilt.[9] Die Signifikanzen von Mittelwertunterschieden wurden daher mit dem nichtparametrischen Kolmogorov-Smirnov-Verfahren gerechnet.[10]

8.3.1 Fach- und Sachinteressen: Kantonale und sprachregionale Unterschiede

Resultate. Bezüglich der Sprachregionen zeigt die untersuchte Stichprobe signifikante Unterschiede. Wie in den Tabellen 8.3 und 8.4 dargestellt, ist das Fachinteresse {IntFachNat} auf beiden Schulstufen in der Deutschschweiz signifikant höher als in der lateinischen Schweiz. Beim allgemeinen Sachinteresse {IntSachNat} gilt dasselbe, wobei die Unterschiede nur im 9. Schuljahr für die Deutschschweiz und die Romandie signifikant sind. Das spezifische Fachinteresse {IntSachGes} zeigt in keiner Stufe und für keine Sprachregion signifikante Mittelwertunterschiede. Innerhalb der Sprachregionen ergeben sich aufgrund des geringen Stichprobenumfangs für keine Kantone signifikante Interessenunterschiede.

	{IntFachNat}	**{IntSachNat}**	**{IntSachGes}**
D-CH	2.10 (.79)	1.71 (.52)	1.93 (.85)
Δ	-0.37 ***	-0.13	0.09
F-CH	1.73 (1.01)	1.58 (.65)	2.02 (.92)

Tabelle 8.3 – Fach- und Sachinteressen: Sprachregionale Unterschiede im 6. Schuljahr: Mittelwertvergleich mit Signifikanz nach Kolmogorov-Smirnov.

Verallgemeinerungsannahmen. Die Untersuchung kantonaler und sprachregionaler Unterschiede dient dem Zweck, die Gültigkeit der zu Beginn des Kapitels vorgestellten Verallgemeinerungsannahmen (cf. S. 164) zu diskutieren.

Aufgrund der unzureichenden Datenlage kann zur Annahme V1 empirisch keine Aussage gemacht werden. Zu keinem der drei Interessenmerkmale ({IntFachNat}, {IntSachNat} und {IntSachGes}) können signifikante kantonale Unterschiede festgestellt werden, obwohl zumindest für das Fachinteresse Naturwissenschaften {IntFachNat} Differenzen

[9] Der Kolmogorov-Smirnov-Test bestätigt für alle Interessenvariablen angewandt auf die Teilstichproben Schulstufen und Geschlecht mit hoher Signifikanz ($p < .001$) die Hypothese der Nichtnormalverteilung.

[10] Der Kolmogorov-Smirnov-Test ist gegenüber anderen nichtparametrischen Verfahren vorzuziehen, wenn die Variablen nur wenige Antwortkategorien besitzen, was bei den Fach- und Sachinteressen hier der Fall ist (Bühl, 2010, 352).

	{IntFachNat}	{IntSachNat}	{IntSachGes}
D-CH	1.73 (.88)	1.54 (.53)	2.02 (.89)
Δ	-0.33 ***	-0.25 ***	-0.11
F-CH	1.40 (.92)	1.29 (.60)	1.91 (1.15)
Δ	-0.08	0.17 ***	-0.25
I-CH	1.32 (.90)	1.46 (.91)	1.66 (1.13)
Δ	0.41 *	0.14	0.42 *
D-CH	1.73 (.88)	1.54 (.53)	2.02 (.89)

Tabelle 8.4 – Fach- und Sachinteressen: Sprachregionale Unterschiede im 9. Schuljahr: Mittelwertvergleich mit Signifikanz nach Kolmogorov-Smirnov.

zu erwarten wären. So ist aus der nationalen Erhebung zu PISA 2006 bekannt, dass im 9. Schuljahr sowohl die kantonal unterschiedliche Unterrichtszeit (Moser & Angelone, 2009, 26f) als auch die Organisation in fächerübergreifenden bzw. fachspezifischen Unterricht (Moser & Angelone, 2009, 31f) einen geringen, jedoch signifikanten Einfluss auf das Fachinteresse Naturwissenschaften haben. Für das Fachinteresse {IntFachNat} ist daher die Verallgemeinerung der Resultate aufgrund der nicht repräsentativen Stichprobe auf die Sprachregionen unwahrscheinlich oder zumindest höchst fragwürdig. Im Gegensatz dazu sind für naturwissenschaftliche Sachinteressen aus der Literatur innerhalb der Sprachregionen keine nennenswerten kantonalen Unterschiede bekannt (siehe u. a. Brühwiler et al., 2009, 52f). Die Verallgemeinerung von Merkmalen der Sachinteressen auf die Sprachregionen ist daher als möglich zu betrachten.

Zur Annahme V2 lassen die Resultate zumindest für das Fachinteresse Naturwissenschaften {IntFachNat} eine klare Aussage zu. Aufgrund der höchst signifikanten Unterschiede zwischen den Sprachregionen ist die Verallgemeinerung von weiteren Unterschieden, z. B. bezüglich des Geschlechts oder der Anforderungsstufe (cf. Tab. 8.5 und 8.8 beim Merkmal {IntFachNat}) auf die ganze Schweiz unwahrscheinlich. Für das allgemeine Sachinteresse Naturwissenschaften sowie das spezifische Sachinteresse Gesundheit kann aufgrund der Resultate die Verallgemeinerung auf die ganze Schweiz als durchaus wahrscheinlich angenommen werden.

8.3.2 Fach- und Sachinteressen: Unterschiede bezüglich Stufe, Anforderungsniveau und Geschlecht

Resultate. Aus den Tabellen 8.5 und 8.6 lässt sich herauslesen, dass sowohl das naturwissenschaftliche Fachinteresse {IntFachNat} als auch das Sachinteresse {IntSachNat} zwischen dem 6. und 9. Schuljahr höchst signifikant abnimmt. Die Jungen geben auf bei-

	Mädchen	Δ	Jungen	beide Geschlechter
6. Schulstufe	1.85 (.89)	0.20 *	2.05 (.92)	1.95 (.91)
Δ	-0.34 **		-0.46 ***	-0.40 ***
9. Schulstufe	1.51 (.89)	0.08	1.59 (.94)	1.55 (.92)
beide Schulstufen	1.65 (.91)	0.13	1.78 (.96)	1.72 (.93)

Tabelle 8.5 – Mittelwerte des naturwissenschaftlichen Fachinteresses {IntFachNat} im 6. und 9. Schuljahr: Mittelwertvergleich mit Signifikanz nach Kolmogorov-Smirnov.

	Mädchen	Δ	Jungen	beide Geschlechter
6. Schulstufe	1.58 (.55)	0.15 ***	1.73 (.60)	1.66 (.58)
Δ	-0.20 ***		-0.23 ***	-0.22 ***
9. Schulstufe	1.38 (.64)	0.12 *	1.50 (.64)	1.44 (.64)
beide Schulstufen	1.46 (.61)	0.14 ***	1.60 (.63)	1.53 (.62)

Tabelle 8.6 – Mittelwerte des Sachinteresses «Naturwissenschaften» im 6. und 9. Schuljahr: {IntSachNat}: Mittelwertvergleich mit Signifikanz nach Kolmogorov-Smirnov.

	Mädchen	Δ	Jungen	beide Geschlechter
6. Schulstufe	2.15 (.82)	-0.28 ***	1.77 (.90)	1.96 (.88)
Δ	0.06		-0.08	-0.03
9. Schulstufe	2.21 (1.03)	-0.56 ***	1.65 (.95)	1.93 (1.03)
beide Schulstufen	2.18 (.95)	-0.48 ***	1.70 (.93)	1.94 (.97)

Tabelle 8.7 – Mittelwerte des Sachinteresses «Gesundheit» im 6. und 9. Schuljahr: {IntSachGes}: Mittelwertvergleich mit Signifikanz nach Kolmogorov-Smirnov.

den Schulstufen ein signifikant höheres Interesse an naturwissenschaftlichen Fächern und Inhalten an als die Mädchen. Eine Ausnahme bildet das zu {IntSachNat} komplementäre Sachinteresse {IntSachGes} (Tab 8.7). Das Interesse an Gesundheitsfragen und Themen zum eigenen Körper bleibt auf beiden Schulstufen auf hohem Niveau. Mädchen zeigen zudem ein höheres Interesse an diesen Themen als die Jungen. Dies gilt zumindest für die Schulstufe 6 signifikant.

Geschlechterunterschiede bestehen zudem in der Rangordnung der drei Interessenvariablen {IntFachNat}, {IntSachNat} und {IntSachGes} (Tab. 8.8). Auf der 6. Schulstufe kommt bei den Jungen das Fachinteresse «Naturwissenschaften» signifikant vor den beiden Sachinteressen. Mädchen hingegen geben dem Sachinteresse «Gesundheit» die signifikant höhere Zustimmung als dem naturwissenschaftlichen Fachinteresse und den Sachthemen. Auf der 9. Schulstufe erhält das Sachinteresse Naturwissenschaften die niedrigsten Inter-

essenwerte. Bei den Mädchen gehört das Thema «Gesundheit» zu den Topthemen.

Das naturwissenschaftliche Fachinteresse {IntFachNat} sowie die Faktoren des Sachinteresses {IntSachNat} und {IntSachGes} zeigen bei den Klassen des 9. Schuljahrs keine signifikante Variation bezüglich des Anforderungsniveaus.[11]

	{IntFachNat}	Δ	**{IntSachNat}**	Δ	**{IntSachGes}**
♀/**6.Sj.**	1.85 (.89)	-0.27 *** ‡	1.58 (.55)	0.57 *** ‡	2.15 (.82)
Δ	0.20 *		0.15 ***		-0.28 ***
♂/**6.Sj.**	2.05 (.92)	-0.32 *** ‡	1.73 (.93)	0.04 ‡	1.77 (.90)
♀/**9.Sj.**	1.51 (.89)	-0.13 * ‡	1.38 (.64)	0.83 *** ‡	2.21 (1.03)
Δ	0.08		0.12 *		-0.56 ***
♂/**9.Sj.**	1.59 (.94)	-0.09 ‡	1.50 (.64)	0.15 ** ‡	1.65 (.95)

	{IntFachNat}		**{IntSachGes}**
♀/**6.Sj.**	1.85 (.89)	-0.30 *** ‡	2.15 (.82)
Δ	0.20 *		-0.28 ***
♂/**6.Sj.**	2.05 (.92)	0.28 *** ‡	1.77 (.90)
♀/**9.Sj.**	1.51 (.89)	-0.70 *** ‡	2.21 (1.03)
Δ	0.08		-0.56 ***
♂/**9.Sj.**	1.59 (.94)	-0.06 ‡	1.65 (.95)

Tabelle 8.8 – Fachinteresse Naturwissenschaften {IntFachNat}, naturwissenschaftliches Sachinteresse {IntSachNat} und Sachinteresse zu Gesundheit {IntSachGes}: Mittelwertvergleich mit Signifikanz, mit ‡ nach Wilcoxon, ansonsten nach Kolmogorov-Smirnov.

Diskussion. Wie bereits im vorangehenden Abschnitt 8.3.1 begründet, lassen sich die Ergebnisse zum Fachinteresse Naturwissenschaften aufgrund der Daten weder auf die Sprachregionen noch auf die ganze Schweiz verallgemeinern. Hingegen kann die Verallgemeinerung der Ergebnisse zu Sachinteressen als wahrscheinlich angenommen werden.

Die Daten des HarmoS-Fragebogens geben die bekannten Muster der Interessenentwicklung bei Schülerinnen und Schülern im Verlauf der Schulkarriere wieder. Das Interesse an naturwissenschaftlichen Sachthemen nimmt in der Sekundarstufe I ab und erreicht den Tiefpunkt am Ende der obligatorischen Schulzeit (Neuhaus & Vogt, 2004). Der teilweise massive Interessenschwund konnte für die Fächer Physik und Chemie in verschiedenen Untersuchungen festgestellt werden, in abgeschwächter Form auch in den Fächern Biologie und Geographie (Hoffmann et al., 1998; Finke, 1999). Dieses Phänomen gehört zum

[11] Die Niveaustufen sind untereinander varianzhomogen, weshalb der Scheffè-Test eingesetzt wurde.

allgemeinen Trend, wonach bereichsspezifische Interessen im Verlauf der Schulzeit generell abnehmen (Artelt et al., 2003, u. a.). Derselbe Trend zeigt sich auch beim Interesse an naturwissenschaftlichen Fächern, wobei er sich bei den Mädchen stärker manifestiert als bei Jungen (Hoffmann et al., 1998, 22ff). Jungen zeigen zudem ein signifikant höheres Interesse an naturwissenschaftlichen Fächern als Mädchen (Hoffmann et al., 1998; Moser & Angelone, 2009, 84ff).

8.3.3 Sachinteressen: Geschlechter- und stufenspezifische Unterschiede

Resultate. Die Rangordnungen der naturwissenschaftlichen Sachinteressen folgen bei Mädchen und Jungen demselben Grundmuster, unterscheiden sich aber in einem Merkmal signifikant (cf. Abb.8.1 und Abb. 8.2). Bei beiden Geschlechtern gehören die Themen «Erde» und «Stoffe» signifikant zu den Themen mit tiefster Interessenausprägung, während die Themen «Gesundheit», «Lebewesen» und «Umwelt» signifikant die höchsten Interessenausprägungen aufweisen. Die Rangordnungen unterscheiden sich aber markant hinsichtlich der Einordnung des Themas «Technik»: Während die «Technik» bei Mädchen die tiefsten Interessenwerte erhält, zählt sie bei den Jungen jeweils zu den Topthemen. Sowohl das gemeinsame Grundmuster, als auch die Diskrepanz bezüglich des Sachinteresses «Technik» lässt sich auf beiden Schulstufen wiederfinden.

Das gleiche gilt für die spezifischen Merkmale der Verteilungen der Sachinteressen. Bei den Mädchen lassen sich signifikant grössere Interessendifferenzen beobachten als bei den Jungen. Die Verteilung bei den Jungen ist flacher. Die Themen «Stoffe» und «Technik» stossen bei Jungen signifikant auf höheres Interesse als bei Mädchen. Umgekehrt erhalten die Themen «Gesundheit», «Lebewesen», «Umwelt» von den Mädchen signifikant mehr Zuspruch als von den Jungen.

Alle erwähnten Unterschiede gelten signifikant auf beiden Schulstufen. Die Entwicklung der Sachinteressen zwischen dem 6. und 9. Schuljahr schlägt sich demnach weniger durch Änderungen in der Rangordnung der Sachinteressen und in der Verteilung der Interessenwerte nieder, sondern zeigt sich in einer allgemeinen Abnahme der Interessen. Wie in den Abbildungen 8.3 und 8.4 dargestellt, gilt dies bei den Jungen signifikant bei fünf Themen und bei den Mädchen bei drei von sieben Sachthemen. Mädchen und Jungen differenzieren im Verlauf der Sekundarstufe I stärker zwischen den einzelnen Sachinteressen. Dies zeigt sich an der zunehmenden Spannbreite der Interessenwerte und an den höheren Signifikanzwerten der Rangordnungen. Bei Mädchen erscheint diese Tendenz stärker ausgeprägt als bei Jungen.

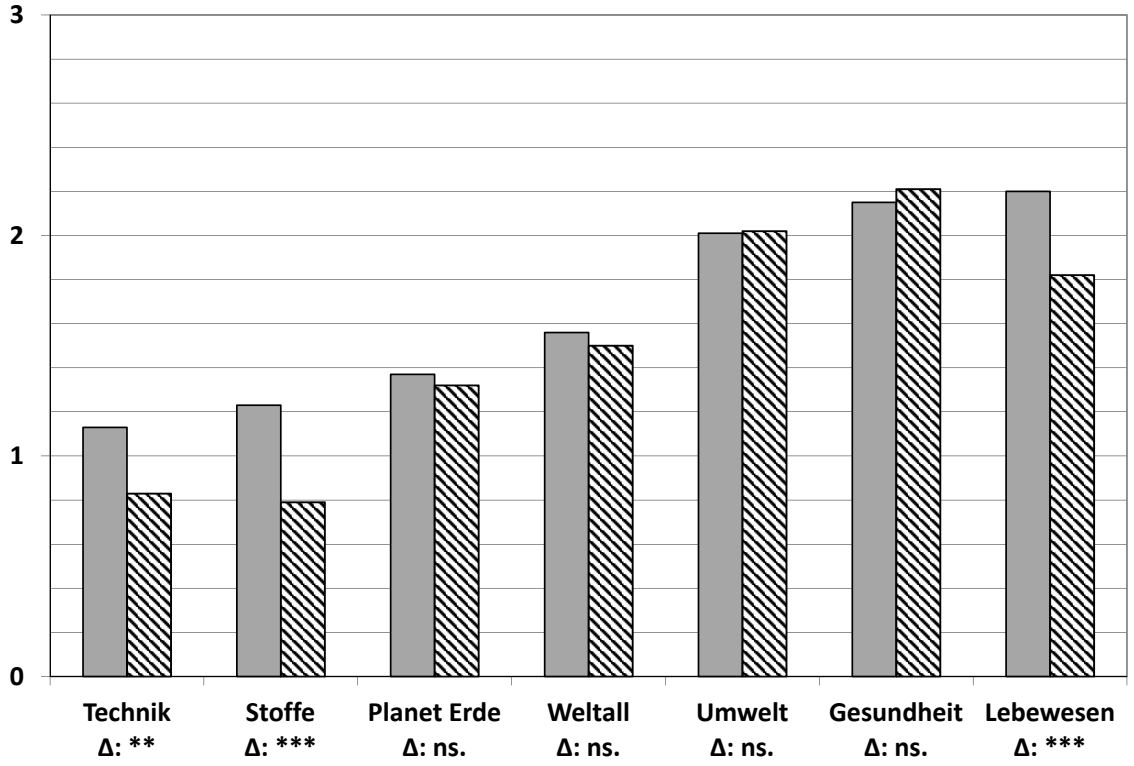


Abbildung 8.1 – Naturwissenschaftliche Sachinteressen bei Mädchen im 6. Schuljahr (grau) und 9. Schuljahr (schraffiert): Mittelwertvergleich mit Signifikanz nach Kolmogorov-Smirnov.

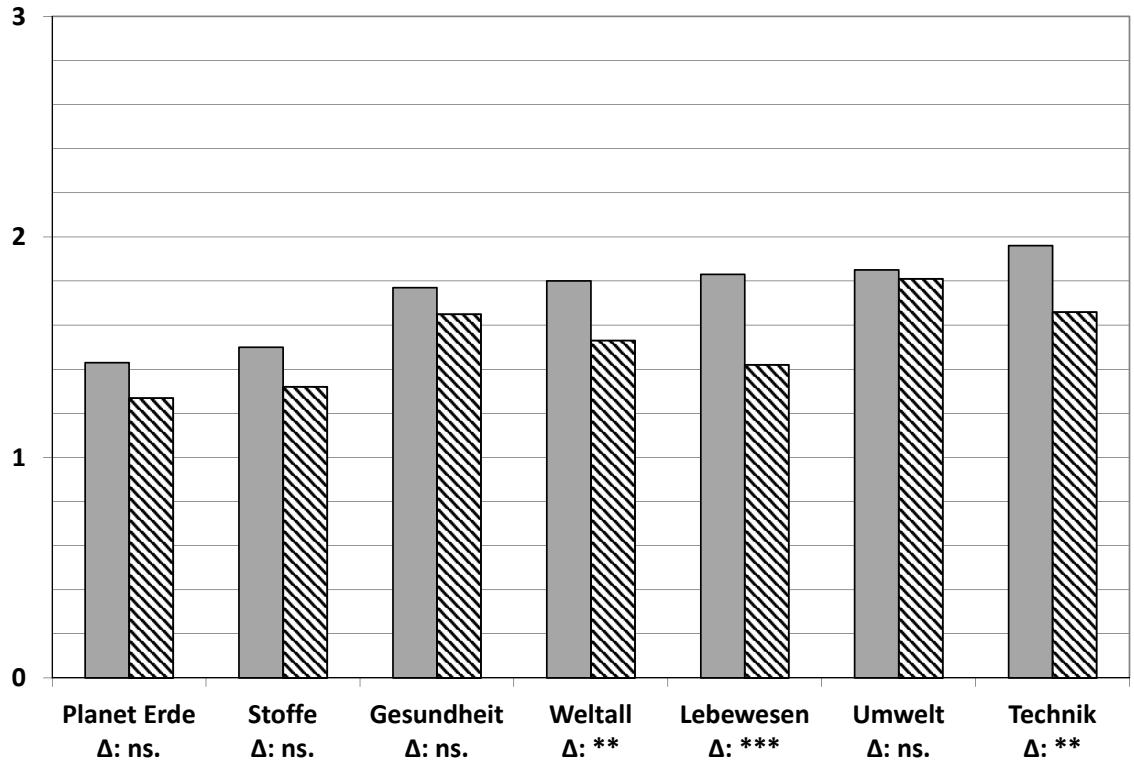


Abbildung 8.2 – Naturwissenschaftliche Sachinteressen bei Jungen im 6. Schuljahr (grau) und 9. Schuljahr (schraffiert): Mittelwertvergleich mit Signifikanz nach Kolmogorov-Smirnov.

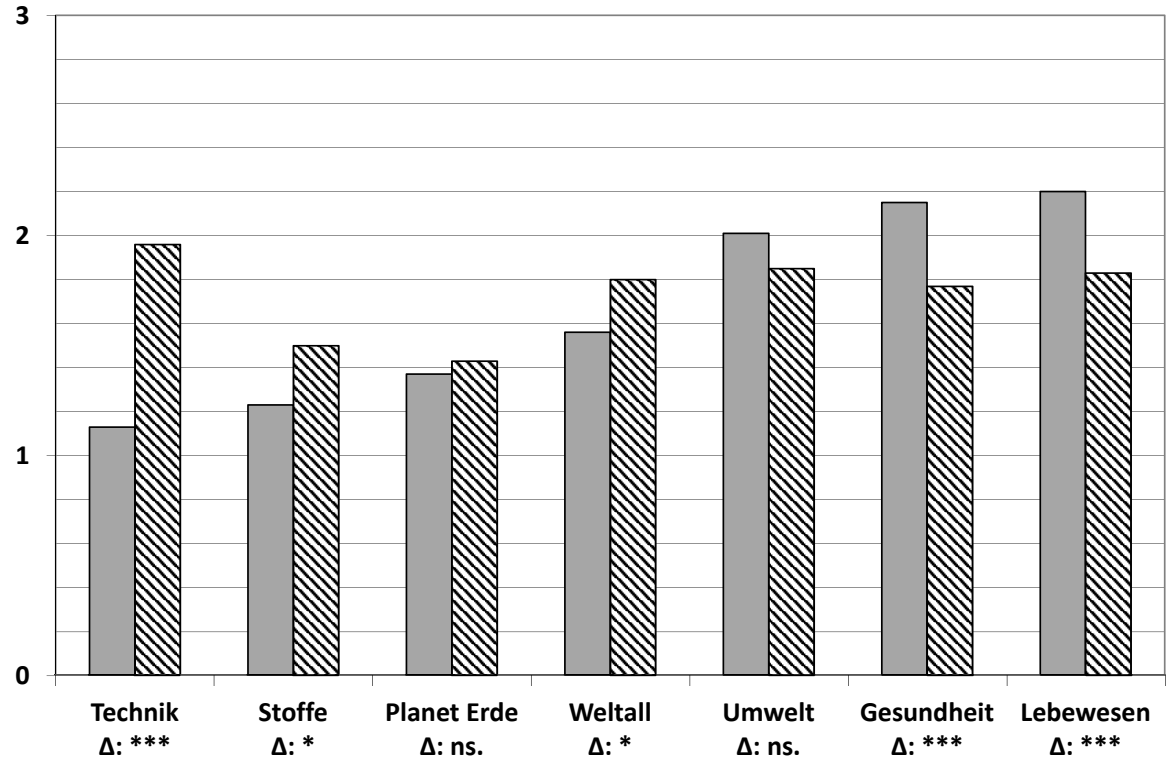


Abbildung 8.3 – Naturwissenschaftliche Sachinteressen im 6. Schuljahr bei Mädchen (grau) und Jungen (schraffiert): Mittelwertvergleich mit Signifikanz nach Kolmogorov-Smirnov.

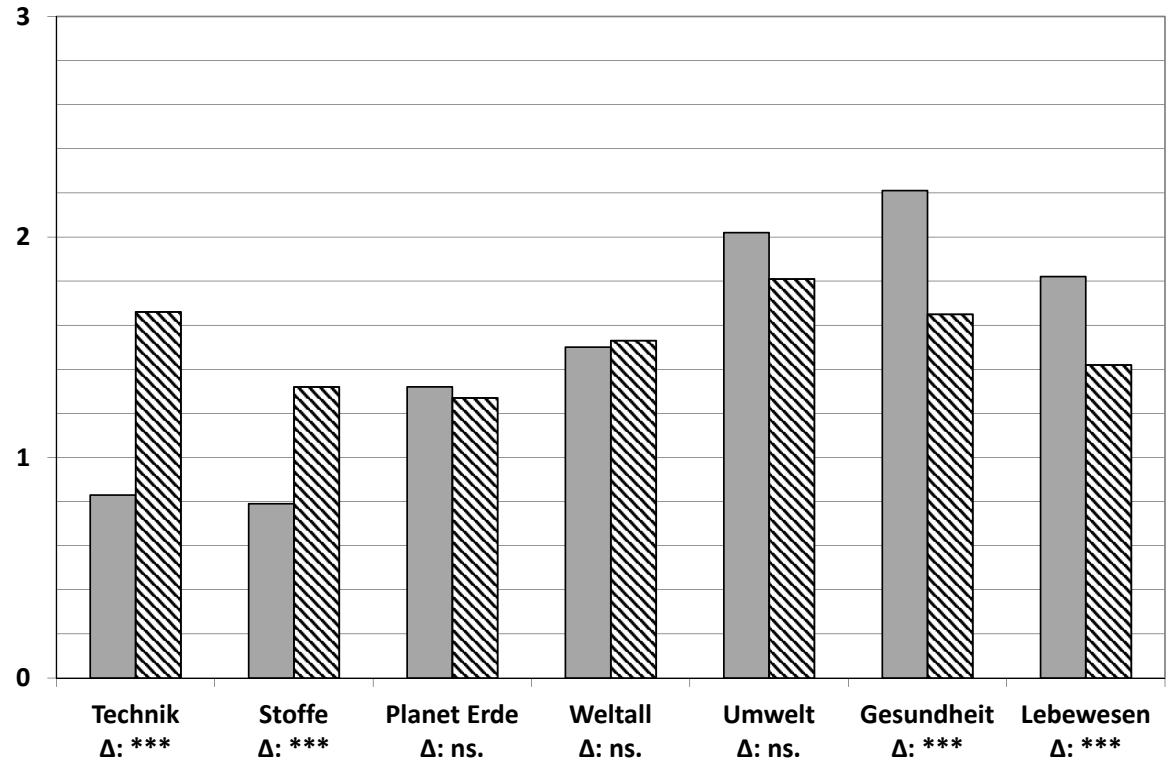


Abbildung 8.4 – Naturwissenschaftliche Sachinteressen im 9. Schuljahr bei Mädchen (grau) und Jungen (schraffiert): Mittelwertvergleich mit Signifikanz nach Kolmogorov-Smirnov.

Diskussion. Die Verallgemeinerung der Resultate zum Sachinteressen auf die ganze Schweiz darf gemäss vorangegangenen Überlegungen (vgl. S. 172) trotz nicht repräsentativer Stichprobe als möglich und wahrscheinlich betrachtet werden. Ein direkter Vergleich mit Resultaten bekannter Interessenstudien ist insofern schwierig, als dass die verwandten Kategorien sich teilweise überschneiden oder nicht übereinstimmen. Trotzdem ergibt ein Vergleich grundsätzlich die Ergebnisse früherer Untersuchungen wieder. Dies betrifft die Rangfolge der Sachinteressen, die geschlechterspezifischen Interessenunterschiede sowie die Interessenentwicklung mit dem Alter.

Die Resultate geben die bekannte Rangfolge der Sachinteressen in groben Zügen wieder, wonach z. B. Chemie- und Physikthemen bei Schülerinnen und Schülern des 9. Schuljahres auf weniger Interesse stossen als Geografiethemen (inklusive Astronomie) und Biologiethemen (Brühwiler et al., 2009; Holstermann & Bögeholz, 2007, 52).

Innerhalb eines Themenbereichs bestehen zu unterschiedlichen Kontexten grosse Interessenunterschiede. Bei beiden Geschlechtern wirkt der Bezug zur Umwelt, insbesondere in Verbindung mit Fragen zur gesellschaftlichen Bedeutung, sowie der Bezug zum menschlichen Körper im Zusammenhang mit Fragen der Gesundheit förderlich für das Interesse (Häußler & Hoffmann, 1995; Hoffmann et al., 1998; Holstermann & Bögeholz, 2007). Mädchen zeigen zudem mehr Interesse an Naturphänomenen als an technischen Objekten. (Hoffmann & Lehrke, 1986; Häußler & Hoffmann, 1995; Häußler et al., 1998; Hoffmann et al., 1998).

Ebenfalls bestätigt wird die Tendenz, dass sowohl bei Jungen als auch bei Mädchen das Interesse an physikalischen Themen zwischen dem 5. und 9. Schuljahr kontinuierlich abnimmt, wobei Jungen signifikant höhere Interessenausprägungen zeigen als Mädchen (Hoffmann & Lehrke, 1986; Hoffmann et al., 1998).

Kapitel 9

Personen-Test-Analysen

9.1 Fragestellung

Im Kapitel 7 wurde ausgeführt, dass die mit dem HarmoS-Experimentiertest gemessenen Leistungsunterschiede zu einem guten Teil als Kompetenzunterschiede interpretiert werden dürfen. Die Leistungsunterschiede können hingegen nicht mit bekannten Kompetenzprogressionen erklärt werden. In diesem Kapitel vergleichen wir die Kompetenzbeherrschung von verschiedenen Personenstichproben differenziert nach den Sprachregionen, den Schulstufen und den Geschlechtern.

4.1. Welcher Kompetenzunterschied besteht zwischen den Geschlechtern?

4.2. Welche Kompetenzprogression besteht zwischen den Schulstufen?

4.3. Welcher Kompetenzunterschied besteht zwischen den Sprachregionen?

4.4. Welche Kompetenzprogression besteht zwischen den Anforderungsniveaus auf der Sekundarstufe I?

4.5. Welcher Zusammenhang besteht zwischen der Kompetenz und der Fremdsprachigkeit und dem familiären Bildungshintergrund?

4.6. Welcher Zusammenhang besteht zwischen der Kompetenz und den Fach- und Sachinteressen?

4.7. Wie schätzen Schülerinnen und Schüler ihre Testleistung ein?

9.2 Auswertung: Rasch-Analyse

Itemstichprobe. Für die Beantwortung der Fragen wurde die Stichprobe der bereits analysierten deutschsprachigen Experimentieritems mit den entsprechenden in der französ-

sischsprachigen Schweiz eingesetzten Items erweitert: ⟨E08|69df⟩. Die untersuchte Personenstichprobe erhöht sich dadurch auf 1363 Schülerinnen und Schüler (cf. Tab. 9.1). Die korrespondierenden deutsch- und französischsprachigen Items wurden als identische Items behandelt, solange sie in ihren jeweiligen Sprachgruppen vergleichbar wirken. Dies wurde bei einem DIF-Parameter < 0.4 gegeben angenommen, analog zur Auswertung des HarmoS-Validierungstests 2007 (cf. Ramseier et al., 2011, 16). 35 Items zeigten einen zu grossen Schwierigkeitsunterschied zwischen den Sprachregionen (DIF-Parameter ≥ 0.4). Davon wurden 24 Items gesplittet, i. e. das deutsch- und das französischsprachige Item wurden als zwei separate Items behandelt (cf. App. E.1: D/F-Split in Tab. E.1). Die restlichen 11 Items wurden wegen ungenügender Trennschärfe gestrichen.

	D-CH	**F-CH**	**beide Sprachregionen**
6. Schulstufe	399	270	**669**
9. Schulstufe	408	286	**694**
beide Schulstufen	807	556	**1363**

Tabelle 9.1 – Personen-Test-Analyse: Personenstichproben; Abkürzungen der Sprachregionen: D-CH = deutschsprachige Schweiz, F-CH = französischsprachige Schweiz.

Itemselektion. Von den 96 (ungesplitteten) Experimentieritems zeigen 25 Items eine zu tiefe Trennschärfe (Korrelation Item-Gesamtscore < 0.3, analog zu Ramseier et al., 2011, 16). Davon wurden 18 gestrichen. 7 Items wurden aufgrund inhaltlicher Überlegungen weiter verwendet. Dies betrifft z. B. besonders schwierige oder besonders einfache Items (cf. Ramseier et al., 2011, 16). Im Gegensatz zur Trennschärfe ist der Fit mit dem Rasch-Modell bei allen Items ausreichend ($0.7 \leq \text{wMNSQ} \leq 1.3$). Aufgrund der Fit-Parameter wurden daher keine weiteren Items gestrichen. Alle im Rahmen der Itemselektion vorgenommenen Korrekturen der Itemstichprobe sind im Appendix E.1 zusammengefasst.

Personenparameter. Mit der korrigierten Itemstichprobe wurde eine zweite eindimensionale Rasch-Analyse gerechnet. Die relevanten Itemparameter sind im Appendix E.2 zusammengestellt. Die Personenparameter wurden als EAP-Werte geschätzt (expected a posteriori Schätzung) und auf eine der PISA-Metrik vergleichbaren Skala mit Mittelwert 500 und Standardabweichung 100 für das 9. Schuljahr transformiert.

9.3 Kompetenz, Geschlecht und Schulstufe

Die Signifikanz von Mittelwertunterschieden der Kompetenz wurde mit dem Programm SPSS gerechnet. Da die Fähigkeitsparameter in nur wenigen Teilstichproben normalverteilt vorliegen, wurden alle Mittelwertvergleiche mit dem nichtparametrischen U-Test nach Mann-Whitney berechnet.

Resultate. Die Mittelwertunterschiede zwischen der 6. und 9. Schulstufe sind hoch signifikant. Da keine vordringlichen sonstigen Gründe für den Differenzen in den Stichproben bekannt sind, wollen wir den Mittelwertunterschied als Kompetenzzuwachs interpretieren. Der diagnostizierte Kompetenzzuwachs findet bei Jungen und Mädchen im gleichen Masse statt. Es gibt jedoch weder innerhalb der Stufen noch über die Stufen hinweg signifikante geschlechtsspezifische Kompetenzunterschiede.

	Mädchen	Δ	**Jungen**	**beide Geschlechter**
6. Schuljahr	469 (90)	6	475 (90)	472 (90)
Δ	27 ***		29 **	28 ***
9. Schuljahr	498 (100)	4	502 (100)	500 (100) §
beide Schulstufen	484 (96)	5	489 (96)	486 (96)

Tabelle 9.2 – Experimentelle Kompetenz im Vergleich der Geschlechter und Schulstufen: Mittelwerte mit Standardabweichung, Mittelwertvergleich mit Signifikanz nach Mann-Whitney. § Fixpunkt der 500/100-Normierung.

Diskussion. Der ermittelte Kompetenzzuwachs zwischen dem 6. und 9. Schuljahr fällt mit knapp einem Drittel einer Standardabweichung der 6. Schulstufe sehr gering aus. Zum Vergleich beträgt der Leistungszuwachs beim Papier-und-Bleistift-Test von HarmoS zwei Drittel der Standardabweichung (cf. Ramseier et al., 2011, 22). Die TIMS-Studie weist in der Schweiz sogar bereits zwischen der Stichprobe des 6. und 7. Schuljahrs einen Zuwachs von rund einem Drittel der Standardabweichung aus (TIMSS, 1996, 29). Da alle drei Anforderungsniveaus Sekundarstufe B/C, Sekundarstufe A und gymnasiale Stufe in der Stichprobe der 9. Schulstufe praktisch gleich stark gewichtet sind und das hohe Anforderungsniveau übervertreten ist, wird zudem aufgrund der nicht repräsentativen Stichprobe der 9. Schulstufe ein eher zu grosser Kompetenzzuwachs erwartet. Eine Erklärung für den vergleichsweise geringen Zuwachs im HarmoS-Experimentiertest kann in den unterschiedlichen Testbedingungen gesucht werden: Im Gegensatz zum 9. Schuljahr wurden im 6. Schuljahr die Testaufgaben von der Testleitung vorgelesen, damit fällt mangelhafte

Lesekompetenz im 6. Schuljahr weniger stark ins Gewicht als im 9. Schuljahr. Die Bearbeitungszeit pro Item wurden ebenfalls kontrolliert. Es wurden daher im 6. Schuljahr immer alle Items bearbeitet, während im 9. Schuljahr Items am Schluss einer Aufgabe häufiger unbearbeitet blieben.

Das Resultat bezüglich des Geschlechterunterschieds korrespondiert mit den Ergebnissen anderer Experimentiertests: Beim internationalen TIMSS-Experimentiertest wurden weder im 4. Schuljahr noch im 8. Schuljahr signifikante Performanceunterschiede zwischen den Geschlechtern festgestellt (TIMSS, 1997, 107). Dies galt auch für die Schweiz, die mit Klassen des 7. Schuljahres am Test teilnahm (Labudde & Stebler, 1999, 36, Stebler et al., 1997, 21). Einzelne Aufgaben wurden jedoch signifikant besser von Mädchen gelöst, andere Aufgaben wiederum von Jungen. In einer Studie von Jovanovic, Solano-Flores und Shavelson (1994) mit Klassen der Schulstufe 5 und 6 wurden ebenfalls keine allgemeinen Leistungsunterschiede festgestellt, in Bezug auf spezifische Themen wurden hingegen signifikante Unterschiede gemessen. Jovanovic et al. (1994) führen dieses Ergebnis teilweise auf geschlechtertypische Erfahrungen mit der natürlichen und technischen Umwelt zurück: "The present results suggest that, regardless of the method of measurement, students' prior experiences play a role in their testing performance. [...] in general boys' (sic!) experiment with batteries and bulbs more often that (sic!) girls [...], whereas girls collect flowers and/or plants more often than boys [...]" (Jovanovic et al., 1994, 9). Ein ähnliches Resultat liefert eine erweiterte Studie von Klein et al. (1997). Während auf der Schulstufe 6 ein leichter Leistungsvorsprung der Mädchen gemessen wurde, konnten auf der Schulstufe 5 und 9 keine signifikanten Unterschiede festgestellt werden. Hingegen gab es Hinweise, dass Jungen bei bestimmten Fragetypen höhere Scores erreichen als Mädchen.

9.4 Teilkompetenzen, Geschlecht und Schulstufe

Die im vorangehenden Abschnitt durchgeführte Untersuchung Kompetenzunterschieden wollen wir auf Teilkompetenzen erweitern. Aufgrund der Dimensionsanalysen im Kapitel 6 und der Ergebnisse zur Interessenverteilung im Kapitel 8 bietet sich hierfür das zweidimensionale Themenbereichmodell an (cf. Abs. 6.4). Untersucht wurden die Unterschiede der zwei Teilkompetenzen (Kompetenz in physikalischen Kontexten versus Kompetenz on nicht physikalischen Kontexten) zwischen den Schulstufen und den Geschlechtern. Zum zweiten wurde untersucht, ob es zwischen den Teilkompetenzen signifikante Ausprägungsunterschiede gibt. Hier gehen wir von der Vermutung aus, dass geschlechtsspezifische Interessenunterschiede mit Unterschieden in der Kompetenzbeherrschung in den entsprechenden Themen zusammenfallen könnten.

Auswertung. Für die Analyse wurden die Personenparameter verwendet, die aus der Berechnung des zweidimensionalen Themenbereichmodells resultierten. Es gilt zu beachten, dass dieses Modell nur mit der Deutschschweizer Teilstichprobe gerechnet wurde. Die Mittelwertunterschiede und Signifikanzen wurde mit SPSS berechnet. Die Fähigkeitsparameter wurden in die PISA-Metrik umgerechnet, wobei die "physikalische" Teilkompetenz des 9. Schuljahrgangs auf die Metrik 500/100 geeicht wurde. Ein Vergleich der Skalen der Teilkompetenzen mit der Skala der eindimensionalen Kompetenz ist hingegen nicht möglich.

Resultate. Die Statistik der Tabelle 9.3 zeigt, dass der im vorangehenden Abschnitt diagnostizierte Kompetenzzuwachs zwischen den Schulstufen allein auf dem signifikanten Zuwachs der "physikalischen" Teilkompetenz beruht. Im Bereich der biologischen und chemischen Themen scheint kein Zuwachs der experimentellen Kompetenz zu erfolgen. Die unterschiedliche Entwicklung der Teilkompetenzen, die in drei wesentliche Merkmale zusammengefasst werden kann, gilt für beide Geschlechter in ähnlicher Weise (cf. Tab. 9.4):

a) Bei der "physikalischen" Teilkompetenz findet zwischen den Schulstufen ein signifikanter Zuwachs statt. Damit verbunden ist eine Zunahme der Streuung (Schereneffekt).

b) Bei der "nicht-physikalischen" Teilkompetenz findet zwischen den Schulstufen weder ein Kompetenzzuwachs noch eine Zunahme der Streuung statt.

c) Im 9. Schuljahr werden Aufgaben mit "physikalischen" Kontext signifikant besser gelöst als Aufgaben mit biologischen oder chemischen Kontexten. Im 6. Schuljahr gibt es keine solche Differenz.

	Biologie & Chemie	Δ	**Physik**
6. Schuljahr	473 (63)	2 †	475 (81)
Δ	4		25 **
9. Schuljahr	477 (66)	23 *** ‡	500 (100) §
beide Schulstufen	476 (66)	21 *** ‡	497 (98)

Tabelle 9.3 – Experimentelle Teilkompetenzen im Vergleich der Schulstufen: Mittelwerte mit Standardabweichung, Mittelwertvergleich mit Signifikanz mit † t-Test, ‡ Wilcoxon-Test, ansonsten mit Mann-Whitney-Test. § Fixpunkt der 500/100-Normierung.

	Mädchen			Δ	Jungen		
	Biologie & Chemie	**Δ**	**Physik**		**Biologie & Chemie**	**Δ**	**Physik**
	470 (63)			6 †	476 (64)		
6. Schuljahr		-4 †				9 †	
			466 (80)	19 †			485 (82)
Δ	10		30 *		-3 †		19
	480 (65)			-7	473 (67)		
9. Schuljahr		16 *** ‡				31 *** ‡	
			496 (94)	8			504 (106)

Tabelle 9.4 – Experimentelle Teilkompetenzen in Bezug auf physikalische und nicht physikalische Themenbereiche im Vergleich der Geschlechter und Schulstufen auf der Basis der Deutschschweizer Teilstichprobe: Mittelwerte mit Standardabweichung, Mittelwertvergleich mit Signifikanz, † t-Test, ‡ Wilcoxon-Test, ansonsten Mann-Whitney-Test.

Diskussion. Der Vergleich der Teilkompetenzen gibt keinen Hinweis auf einen möglichen Gendereffekt (cf. Tab. 9.4). Ein Zusammenhang zwischen der Ausprägung der Teilkompetenzen und geschlechtsspezifischen Interessenunterschieden kann somit nicht hergestellt werden. Dieses Ergebnis deckt sich mit der Untersuchung zu den Sach- und Fachinteressen, der im Abschnitt 9.8 folgt.

Als Erklärung für die oben erwähnten Differenzen a), b) und c) zwischen den Teilkompetenzen lassen sich nur Vermutungen aufstellen, die im Rahmen der vorliegenden Arbeit nicht überprüft werden können. Hierfür soll folgende zwei Fragekomplexe diskutiert werden:

1. Inwieweit hängen die Differenzen von Aufgabenmerkmalen ab, die direkt mit der Art der Kontexte – biologisch, chemisch oder physikalisch – zusammenhängen?
   Inwieweit spielen Aufgabenmerkmale eine Rolle, die überhaupt nicht vom Kontext abhängt (e. g. die Aufgaben mit bestimmten Kontext erfordern zufälligerweise mehr Lesekompetenz) oder nur indirekt einen Bezug zum Kontext haben (e. g. aus Tradition werden in den verschiedenen Fächern andere Aufgabentypen bevorzugt behandelt)?

2. Inwieweit lassen sich die unterschiedlichen Kompetenzentwicklungen zwischen den Schulstufen mit natürlich erfolgenden Entwicklungen von Schülerinnen und Schülern erklären (e. g. Ausbildung des formal logischen Denkens)?
   Inwieweit lassen sich die Differenzen auf Merkmale des naturwissenschaftlichen Unterrichts auf der Sekundarstufe I zurückführen (e. g. die Aufgaben erfordern Kon-

zeptwissen und Fähigkeiten, die erst auf der Sekundarstufe vermittelt werden bzw. nicht vermittelt werden)?

Ad 1.: Analysiert man die Itemschwierigkeit für Aufgaben mit physikalischem Kontext und mit biologischem oder chemischem Kontext separat, stellt man fest, dass aus den Modellierungen unterschiedliche schwierigkeitserzeugende Itemmerkmale reduzieren, die alle jedoch keinen Kontextbezug haben. Die Modellierung der "physikalischen" Items gelingt mit den im Kapitel 7 hergeleiteten Instrumenten zudem deutlich besser als die Modellierung der biologischen und chemischen Items.[1] Diese Ergebnisse deuten an, dass die Schwierigkeit der Aufgaben in den untersuchten Kontexten auf unterschiedliche Weise erzeugt wird und nicht nur vom Kontext abhängt.

Ad 2: Die ausbleibende Entwicklung der "biologisch-chemischen" Teilkompetenz zwischen den Schulstufen lässt die Vermutung zu, dass die Aufgaben dieser Themenbereiche nur wenig auf schulischem Vorwissen oder in der Schule vermittelte Fähigkeiten aufbauen bzw. wenig entwicklungsbedingte Fähigkeiten erfordern, dies im Gegensatz zu den "physikalischen" Aufgaben. Die These kann an dieser Stelle zwar nicht wissenschaftlich erhärtet werden. Wir wollen jedoch einige Erklärungsansätze erwähnen.

Die Phänomene, die in den Aufgaben thematisiert werden, sind mehr oder weniger artifiziell bzw. alltäglich. Zu den artifiziellen Kontexten zählen eher die technischen Aufgaben wie ⟨Solarzellen⟩ und ⟨Sparlampe⟩. Aufgaben mit eher alltäglichen Phänomene sind eher "nicht-physikalisch" wie ⟨Gänseblümchen⟩, ⟨Laubbäume⟩, ⟨Asseln⟩, ⟨Sinken & Schwimmen⟩ oder ⟨Seife⟩ (man vergleiche die Kontextbeschreibungen in der Tabelle 7.7 auf Seite 140). Es gibt jedoch auch pphysikalische" Aufgaben mit lebensweltlichem Bezug, e. g. ⟨Murmel⟩ oder ⟨Taschenlampe⟩. Bestimmte Aufgaben haben eher schulischen Charakter und referieren stark auf vorschulische Schülervorstellungen. Dies gilt u. a. für die Aufgaben ⟨Balkenwaage⟩, ⟨Taschenlampe⟩, ⟨Steine⟩. Dies gilt jedoch auch für die "chemische" Aufgabe ⟨Öl⟩ oder ⟨Wasserpest⟩. Gewisse Aufgaben erfordern zudem mathematische, kombinatorische und logische Fähigkeiten. Dies betrifft wiederum eher die "physikalischen" Aufgaben ⟨Balkenwaage⟩ (vgl. die Ausführungen zum entsprechenden Fallbeispiel auf den Seiten 66ff), ⟨Murmel⟩, ⟨Steine⟩. Auch hier gibt es jedoch nicht-physikalische Beispiele, wie z. B. die Öl-Aufgabe.

Die Erklärung der unterschiedlichen Entwicklungen der Teilkompetenzen erfordert zusätzliche Untersuchungen. Aus solchen Analysen könnten wichtige Erkenntnisse für die Modellierung von Itemschwierigkeit resultieren.

[1] Bei den "physikalischen" Items lässt sich 55% der Varianz der Schwierigkeit mit 8 signifikanten Itemmerkmalen erklären. Bei den Items mit biologischem oder chemischen Kontext können nur 36% der Varianz erklärt werden.

9.5 Kompetenz, Sprachregion und Schulstufe

Unterschiede zwischen den Sprachregionen wurden analog zur Auswertung der Geschlechterunterschiede analysiert. Die Mittelwertunterschiede wurden ebenfalls mit dem nichtparametrischen U-Test nach Mann-Whithey nach Signifikanzen getestet.

Resultate. Der HarmoS-Experimentiertest liefert auch in den Sprachregionen einen Kompetenzzuwachs von knapp einem Drittel der Standardabweichung der 6. Schulstufe. Die Testleistung der französischsprachigen Schülerinnen und Schüler ist zudem signifikant höhere als die Testleistung der Deutschschweizer Schülerinnen und Schüler. Die Performancedifferenz beträgt rund ein Sechstel der Standardabweichung.

	D-CH	**Δ**	**F-CH**	**beide Sprachregionen**
6. Schuljahr	466 (93)	16 *	482 (84)	472 (90)
Δ	28 ***		26 **	28 ***
9. Schuljahr	494 (98)	14 *	508 (102)	500 (100) §
beide Schulstufen	480 (97)	16 **	496 (95)	486 (96)

Tabelle 9.5 – Experimentelle Kompetenz im Vergleich der Sprachregionen und Schulstufen: Mittelwerte mit Standardabweichung, Mittelwertvergleich Vergleich mit Signifikanz nach Mann-Whitney. § Fixpunkt der 500/100-Normierung.

Diskussion. In Bezug auf die sprachregionalen Unterschiede liefert der Experimentiertest ein konträres Bild zum HarmoS-Validierungstest 2007 und zur PISA-Staffel 2006 mit Schwerpunkt Naturwissenschaften (Nidegger, Moreau & Gingins, 2009). Während im Papier-und-Bleistift-Test von HarmoS die Deutschschweizer Schülerinnen und Schüler (beide Schulstufen zusammengenommen) auf der 500/100-Skala eine um 17 Punkte bessere Performance zeigten als die Schülerinnen und Schüler aus der Romandie (bei PISA 2006 waren es 16 Punkte), erreicht die Westschweiz im Experimentiertest 16 Punkte mehr als die Deutschschweiz. Das Resultat ist bemerkenswert und lässt sich ohne weitere Detailanalysen nicht eindeutig erklären. (Wie im folgenden Abschnitt 9.6 gezeigt wird, fällt vor allem Stichprobe aus dem Kanton Zürich ab, während die Ergebnisse aus den Kantonen Bern und Waadt vergleichbar sind). Es könnte hier aber auch eine fehlerhafte Messung vorliegen: Nicht mehr nachprüfbar ist beispielsweise, ob die Schülerinnen und Schüler in der Westschweiz bei der Testdurchführung durchschnittlich mehr assistiert wurden als in der Deutschschweiz. Es könnten auch die deutschen Testhefte strenger kodiert worden sein als die französischen. Alle Testbogen wurden nämlich von Hilfsassistierenden deutscher Muttersprache kodiert. Die Kodierung in der Muttersprache ist nicht unbedingt

vergleichbar mit der Kodierung in einer Fremdsprache. Sollte das Resultat aber "korrekt" sein, wäre es ein starker Beleg dafür, dass mit dem Experimentiertest andere Kompetenzen gemessen werden als mit dem Papier-und-Bleistift-Test.

Die zweite Analyse bestätigt zudem die erste Analyse in Bezug auf die grösser werdende Leistungsstreuung zwischen dem 6. und 9. Schuljahr. Das Phänomen ist bereits vom Papier-und-Bleistifttest bekannt (cf. Ramseier et al., 2011, 18). Eine mögliche Erklärung ist der durch die Selektion der Primarschülerinnen und -schüler in unterschiedliche Schultypen verursachte Schereneffekt (Becker, Lüdtke, Trautwein & Baumert, 2006).

9.6 Kompetenz, Kanton und Anforderungsniveau auf der Sekundarstufe I

Anforderungsniveau. Die getesteten Klassen des 9. Schuljahrs wurden gemäss der Tabelle 8.2 auf der Seite 167 in drei Anforderungsniveaus unterteilt. Die verwandte Niveaustufung lehnt sich an die Einteilung von Moser und Angelone (2009, 20ff) an. Die Sekundarklasse, welche als einzige Klasse aus dem Kanton Solothurn am Test teilnahm, wurde dem tiefsten Anforderungsniveau zugeordnet. Die Klassen aus dem Tessin wurden keinem Niveau zugeordnet, da die Tessiner Scuola media alle Niveaus unter einem Dach integriert. Die Zuordnung der Tessiner Klassen wird für diese Auswertung nicht benötigt, spielte jedoch für die Auswertung der Fragebogen im vorangehenden Kapitel 8 eine Rolle.

Resultate. Die Kompetenzverteilung ist bei praktisch allen Stichproben nicht normal. Die Mittelwertunterschiede wurden daher mit dem U-Test von Mann-Whitney gerechnet. Die Resultate sind in der Tabelle 9.6 dargestellt.

Diskussion. Die Untersuchung gibt für die Niveauunterschiede das erwartete Ergebnis wieder. Im Vergleich zum HarmoS-Validierungstest 2007 fällt der Kompetenzzuwachs mit zunehmender Schulkarriere geringer aus. Während beim Papier-und-Bleistifttest der Leistungsunterschied zwischen dem höchsten Anforderungsniveau und dem tiefsten Niveau auf der 500/100-Skala 138 Punkte beträgt (cf. Ramseier et al., 2011, 23), stellen wir bei vergleichbarer Niveaudifferenzierung und 500/100-Normierung nur einen Zuwachs von 87 Punkten fest. Der Leistungsunterschied im Experimentiertest zwischen dem höchsten und dem tiefsten Anforderungsniveau im 9. Schuljahr entspricht etwa dem Dreifachen des Leistungszuwachses zwischen dem 6. und 9. Schuljahr. Im Validierungstest 2007 beträgt der Niveauunterschied auf der Sekundarstufe etwa das Doppelte des Kompetenzzuwachses zwischen den Schulstufen (cf. Ramseier et al., 2011, 22f). Dieses Ergebnis ist erstaunlich, denn das kompetenzorientierte Experimentieren ist in der Schulpraxis der Sekundarstu-

	9. Schuljahr						6. Schuljahr
	Niveau 1	**Δ**	**Niveau 2**	**Δ**	**Niveau 3**	**alle Niveaus**	
ZH	437 (97)	37*	474 (81)	48 *	522 (76)	473 (91)	
Δ	39 * †		6		48 **	33 **	
BE/SO	476 (89)	4 †	480 (102)	90 ***	570 (79)	506 (100)	
Δ	2 †		10 †		-16	2 †	
VD	478 (98)	12 †	490 (98)	64 ***	554 (94)	508 (102)	
Δ	-41 * †		-16		-32 *	-35 **	
ZH	437 (89)	37 *	474 (81)	48 *	522 (76)	473 (91)	
CH	467 (95)	16 †	483 (96)	71 ***	554 (87)	500 (100) §	472 (90)

Tabelle 9.6 – Experimentelle Kompetenz in den verschiedenen Anforderungsniveaus {AnfNiv} der Sekundarstufen im 9. Schuljahr im Vergleich mit dem 6. Schuljahr: Mittelwerte mit Standardabweichung, Mittelwertvergleich mit Signifikanz nach Mann-Whitney, ausser bei † mit t-Test. § Fixpunkt der 500/100-Normierung.

fe I nicht besonders ausgeprägt. Es wird zwar vergleichsweise viel experimentiert (Börlin, 2012, 73ff), jedoch meist ohne Kompetenzorientierung, u. a. werden die Prozesse beim Experimentieren im Unterricht kaum reflektiert (Börlin, 2012, 89ff). Der Leistungszuwachs geschieht im Unterricht und beim Experimentieren somit so "nebenbei", ohne dass dies durch die Lehrpersonen bewusst und didaktisch gezielt angestrebt wird. Kein signifikanter Leistungsunterschied kann beim Experimentiertest zwischen dem tiefsten und dem mittleren Niveau festgestellt werden. In beiden Niveaus wird auch praktisch kein Kompetenzzuwachs gegenüber dem 6. Schuljahr gemessen; im tiefsten Anforderungsniveau erreichen durchschnittliche Schülerinnen und Schüler im 9. Schuljahr nicht einmal den Level einer durchschnittlichen 6. Klässerin bzw. eines 6. Klässlers. Dieser klare Befund ist wenig schmeichelhaft für die Sekundarschulen und könnte als Hinweis für einen zumindest in Bezug auf das Experimentieren "schlechteren" Unterricht gewertet werden. Es gilt aber auch zu beachten, dass der Experimentiertest auch auf Lese- und Schreibschwächen reagiert. Somit offenbart der HarmoS-Experimentiertest nicht nur beim Experimentieren, sondern auch beim Lesen und Kommunizieren von naturwissenschaftlichen Inhalten ein enormes Defizit in den Sekundarschulen.

Um die Diskussion differenziert zu führen, wollen wir auch die alternative Interpretation in Betracht ziehen, dass sich im Resultat vor allem eine Eigenschaft des Experimentiertests manifestiert. Nämlich dass im Test Aufgaben mittlerer Schwierigkeit fehlen, die zwischen den beiden Sekundarstufen A und B/C zu differenzieren vermögen. Diese Erklärung widerspricht jedoch dem Ziel der Aufgabenkonstruktion, Aufgaben für tiefe und mittlere Niveaus zu entwickeln.

Im Gegensatz zu den Stufenvergleichen widersprechen die kantonalen Unterschiede diametral den Erwartungen. Bei der PISA-Staffel 2009 war die Rangfolge der Kantone Bern, Zürich und Waadt gerade umgekehrt zur Rangfolge beim HarmoS-Experimentiertest,wobei die kantonalen Unterschiede nicht signifikant sind (cf. PISA, 2011, 44). Es fällt zudem das sehr schlechte Testresultat der Niveaus 1 und 3 im Kanton Zürich auf. Wir vermuten, dass die Abweichungen ein Effekt der Stichprobenziehung ist. Im Kanton Zürich wurden pro Niveau nur etwa drei Klassen, in Bern und in der Waadt im Schnitt über acht Klassen getestet. Drei Klassen sind nicht repräsentativ, weshalb die Wahrscheinlichkeit gross ist, dass das signifikant schlechtere Ergebnis der Deutschschweiz gegenüber der Romandie auf einen Stichprobeneffekt im Kanton Zürich zurückgeführt werden kann. Auf jeden Fall zeigen die Resultate die grosse Leistungsstreuung beim Experimentieren auf dieser Schulstufe auf.

9.7 Kompetenz, Migrations- und Bildungshintergrund

Auswertung. Die untersuchten Variablen wurden mit Hilfe eines Mediansplits dichotomisiert. Die Mittelwertunterschiede wurden mit dem Programm SPSS ausgewertet.

Resultate. Sowohl die Variable zur Unterrichtssprache {UntSpra} als auch die Variable zur Mehrsprachigkeit {MehrSpra} differenziert hoch signifikant zwischen hohen und tiefen Testleistungen (cf. Tab. 9.7). Ebenfalls signifikante Mittelwertunterschiede resultieren für die dichotomisierten Variablen des familiären Bildungsbewusstseins {BildBew-dic} sowie der elterlichen Bildungsorientierung {BildOr-dic}. Nicht signifikant korreliert das persönliche Interesse an gesellschaftlichen Belangen {GesInt-dic} mit der Testleistung (cf. Tab. 9.8).

{UntSpra}		**{MehrSpra}**	
Unterrichtssprache wird zuhause nicht häufig gesprochen	462 (110)	*Zuhause werden mehr als eine Sprache häufig gesprochen*	466 (94)
Δ	45 ***	Δ	39 **
Unterrichtssprache wird zuhause häufig gesprochen	507 (97)	*Zuhause wird nur eine Sprache häufig gesprochen*	505 (101)
Mittelwert	500 (100)		500 (100)

Tabelle 9.7 – Experimentelle Kompetenz: Mittelwertvergleich bzgl. Variablen der Sprache für das 9. Schuljahr: Mittelwerte mit Standardabweichung, Mittelwertvergleich Vergleich mit Signifikanz nach Mann-Whitney.

	{BildBew-dic}	{BildOr-dic}	{GesInt-dic}
tiefe Ausprägung	493 (99)	493 (98)	503 (100)
Δ	33 **	16 *	-6
hohe Ausprägung	526 (103)	509 (103)	497 (103)
Mittelwert	500 (101)	501 (101)	500 (101)

Tabelle 9.8 – Experimentelle Kompetenz: Mittelwertvergleich bzgl. Variablen des häuslichen Umfelds für das 9. Schuljahr: Mittelwerte mit Standardabweichung, Mittelwertvergleich Vergleich mit Signifikanz nach Mann-Whitney.

Diskussion. In Leistungstests erweisen sich die drei Herkunftsmerkmale Migrationshintergrund, Kenntnis der Schulsprache und soziale Herkunft ganz allgemein als bedeutsam für die Schulleistung. U. a. wurden diese Zusammenhänge in den kantonalen Teilauswertungen der PISA-Staffel 2009 bestätigt (Moser & Angelone, 2011, 16ff). Unsere Ergebnisse lassen sich in Übereinstimmung mit diesen Zusammenhängen interpretieren. Das Item Unterrichtssprache {UntSpra-dic} referiert direkt auf die Kenntnis der Schulsprache. Für das Item Mehrsprachigkeit {MehrSpra-dic} liegt die Vermutung nahe, dass das Merkmal einen nachteiligen Migrationshintergrund indiziert. Beide Merkmale, Kenntnis der Unterrichtssprache und die Mehrsprachigkeit, bilden zusammen übrigens die PISA-Skala Migrationshintergrund (Moser & Angelone, 2009, 18). Das Bildungsbewusstsein {BildBew-dic} und die Bildungsorientierung {BildOr-dic} enthalten verschiedene Aspekte der sozialen Herkunft.

Der diagnostizierte massive Einfluss der Sprachfähigkeiten auf die Testleistung korrespondiert mit dem Ergebnis der Itemanalyse, dass der Experimentiertest zu einem wesentlichen Teil auch Lese- und Schreibfähgkeiten mitmisst.

9.8 Kompetenz, naturwissenschaftliches Fach- und Sachinteresse

Resultate. Zwischen dem naturwissenschaftlichen Sachinteresse {IntSachNat-dic} und der Testleistung besteht kein messbarer Zusammenhang. Höheres Interesse an naturwissenschaftlichen Fächern {IntFachNat-dic} korreliert hingegen mit besseren Leistungen, der Zusammenhang ist jedoch nicht signifikant (cf. Tab. 9.9).

Diskussion. Je kompetenter die Jugendlichen sind, desto stärker ausgeprägt ist das Interesse an den naturwissenschaftlichen Fächern. Derselbe Zusammenhang gilt beim naturwissenschaftlichen Sachinteresse nicht. Der Zusammenhang beim Fachinteresse ist zwar

	{IntFachNat-dic}	{IntSachNat-dic}
tiefe Ausprägung	493 (98)	501 (100)
Δ	12	-1
hohe Ausprägung	505 (103)	500 (102)
Mittelwert	499 (101)	500 (101)

Tabelle 9.9 – Experimentelle Kompetenz: Mittelwertvergleich bzgl. naturwissenschaftlichen Fach- und Sachinteressen für das 9. Schuljahr: Mittelwerte mit Standardabweichung, Mittelwertvergleich Vergleich mit Signifikanz nach Mann-Whitney.

nicht signifikant, er entspricht aber einer bei PISA international festgestellten Tendenz, u. a. auch beim deutschen Ländervergleich zur PISA-Staffel 2006 (PISA, 2008, 104). Das PISA-Konstrukt "Interesse an Naturwissenschaften" entspricht jedoch eher dem hier erfassten Sach- als dem Fachinteresse (PISA, 2008, 97).

9.9 Kompetenz und persönliche Testeinschätzungen

Resultate. Das Interesse an den Inhalten des Tests {TestInt-dic} sowie die persönliche Einschätzung der Relevanz des Tests {TestRel-dic} spielt für die erbrachte Testleistung keine signifikante Rolle. Die Einschätzung der Testschwierigkeit {TestSchw-dic} korreliert hochsignifikant mit der Testleistung: Wer den Test als schwierig einstuft, schneidet beim Test schlecht ab, wer den Test als leicht einstuft, erhält gute Testergebnisse (cf. Tab. 9.10).

Diskussion. Wie beim Fachinteresse trifft auch beim Interesse an den Testaufgaben höhere Ausprägung mit höherer Kompetenz zusammen. Der Zusammenhang ist jedoch auch in diesem Fall nicht signifikant. Markant ist die hohe Verlässlichkeit, mit der die Jugendlichen die Testschwierigkeit in Relation zu ihrer Kompetenz diagnostizieren.

	{TestInt-dic}	{TestRel-dic}	{TestSchw-dic}
tiefe Ausprägung	497 (100)	498 (100)	540 (93)
Δ	9	13 †	-50 *** †
hohe Ausprägung	506 (103)	511 (104)	490 (100)
Mittelwert	501 (101)	500 (101)	501 (101)

Tabelle 9.10 – Experimentelle Kompetenz: Mittelwertvergleich bzgl. persönlichen Einschätzungen zum Test für das 9. Schuljahr: Mittelwerte mit Standardabweichung, Mittelwertvergleich Vergleich mit Signifikanz, † t-Test, ansonsten Mann-Whitney-Test.

Teil III

Diskussion

Kapitel 10

Zusammenfassung der Ergebnisse

Abschliessend fassen wir die Antworten auf die im Kapitel 2 vorgestellten Forschungsfragen zusammen. Mit einem kritischen Rückblick auf den HarmoS-Experimentiertest werden die Schlüsse aus den Ergebnissen gezogen und Empfehlungen für zukünftige Experimentiertests gegeben. Die Diskussion schliesst mit dem Ausblick auf weitere Forschungsanliegen.

Test-Analysen: Dimensionalität des HarmoS-Experimentiertests

Die Struktur des HarmoS-Experimentiertests wird durch die starken Itemabhängigkeiten geprägt. Diese betreffen vor allem die Differenzierung der Teilprozesse, in einem geringen Umfang auch die Aufgabentypen. Nicht betroffen von den Itemabhängigkeiten sind die Themenbereiche.

1.1. Teilprozesse als Dimensionen. Der HarmoS-Experimentiertest differenziert nicht zwischen Teilprozessen. Insofern wird durch den Test eine umfassende experimentelle Kompetenz erfasst.

1.2. Aufgabentypen als Dimensionen. Der Test unterscheidet hingegen mässig zwischen Aufgabentypen, die eine aktive oder passive Rolle der Testperson erfordern.

1.3. Themenbereiche als Dimensionen. Bezüglich der Themenbereiche misst der Experimentiertest verschiedene Dimensionen. Zumindest zwischen physikalischen Kontexten und nicht physikalischen Kontexten liegen zwei Dimensionen der experimentellen Kompetenz vor, die sich mit einer Korrelation von 0.49 statistisch deutlich voneinander unterscheiden. Ungenügend hingegen sind die EAP/PV-Reliabilitäten (0.40-0.50).

Item-Test-Analysen: Erklärung der Itemschwierigkeit

2.1. Modellierung der Itemschwierigkeit. Mit Hilfe der Unterscheidung der vier Arbeitsschritte ‹Aufgabe erfassen›, ‹Problem lösen›, ‹Antwort geben› und ‹Lösung kodieren› lässt sich die Itemschwierigkeit erfolgreich modellieren. Die Formatanalyse der gedruckten Aufgabenstellung erweist sich dabei als brauchbares Instrument, kompetenzirrelevante schwierigkeitsinduzierende Itemmerkmale zu eruieren. Jedoch ist die Zuordnung der Merkmale zu den vier Arbeitsschritten nicht immer eindeutig. Demgegenüber gestaltet sich die Analyse von kompetenzrelevanten Itemmerkmalen post hoc als schwierig. Die Schwierigkeit des Arbeitsschritts ‹Problem lösen› kann durch keine der betrachteten Merkmalkategorien (Teilprozesse, Aufgabenumfang, -typ und -kontext sowie Eigenschaften des Lösungswegs und der erforderlichen experimentellen Manipulationen) signifikant beschrieben werden. Der Arbeitsschritt ‹Lösung kodieren›, der die Qualität der Lösung erfasst, wird hingegen mit Hilfe so genannter Kodiermassstäbe, die auf Korrektheits-, Präzisions- und Vollständigkeitsidealen basieren, hinreichend gut erklärt.

2.2. Kompetenzrelevante schwierigkeitsinduzierende Itemmerkmale. Von den fünf Progressionsdimensionen für die Kompetenzbeherrschung werden nur vier vom HarmoS-Experimentiertest erfasst. Obwohl individuelle Hilfestellungen während des Tests gegeben wurden, wurde die Eigenständigkeit $[E]$ nicht separat kodiert. Hingegen variieren die Items bezüglich des Aufgabenumfangs $[A]$, der Problemkomplexität $[P]$ und der Kodierung der Prozessqualität $[Q]$ derart, dass sie mit geeigneten Itemmerkmalen sinnvoll beschrieben werden können. Der Transferumfang $[T]$ fliesst zudem aufgrund des Testdesigns (jede Testperson bearbeitete zwei Experimentieraufgaben zu verschiedenen Themenbereichen) und via Rasch-Analyse in die Testergebnisse ein.

Für die vier vom Test berücksichtigten Progressionsdimensionen konnten in verschiedenen Regressionsanalysen nur für die Prozessqualität Itemmerkmale eruiert werden, die signifikant mit der Itemschwierigkeit korrelieren. Als schwierig erweisen sich diejenigen Items, bei denen praktisch-technische Fertigkeiten beim Erzeugen von Evidenz beurteilt werden und die selbst aufgrund der erzeugten Evidenz kodiert werden. Alles Items also, die nicht mit einem Papier-und-Bleistifttest möglich wären! Hoch signifikant ist zudem der Einfluss auf die Itemschwierigkeit, wenn Aufgaben unvollständige Lösungen zulassen.

Zwar blieb die Modellierung der Problemkomplexität mit Hilfe von Itemmerkmalen ergebnislos. Der Einfluss der Problemkomplexität auf die Kompetenz wird hingegen durch die Dimensionsanalyse belegt (cf. Frage 1.2. oben): Der HarmoS-Experimentiertest differenziert statistisch zwischen verschiedenen Aufgabentypen, die qualitativ mit der Problemkomplexität zusammenhängen. Dieses Resultat lässt die begründete Interpretation

zu, dass es ein Konstrukt wie eine allgemeine Problemkomplexität beim Experimentieren nicht gibt, sondern dass die Problemkomplexität für verschiedene Aufgabentypen separat konstruiert und ausgewertet werden muss.

Keine Entsprechung auf der Ebene der Itemmerkmale gibt es letztlich für den Transferumfang. Informationen dazu erhalten wir allein durch die Dimensionsanalysen, die deutlich aufzeigen, dass dem HarmoS-Experimentiertest eine mehrdimensionale, nach Themenbereichen differenzierende Struktur unterliegt. Der themenübergreifende Kompetenztransfer von einem Kontext auf einen anderen ist somit kein graduelles Merkmal einer allgemeinen Kompetenzausprägung, sondern erweist sich zumindest für Laien – die getestete Personenstichprobe enthält keine Expertinnen und Experten – vielmehr als die qualitative Übersetzung zwischen zwei verschiedenen Kompetenzen.

2.3. Kompetenzirrelevante schwierigkeitsinduzierende Itemmerkmale. Sowohl für den Arbeitsschritt ‹Aufgabe erfassen› als auch für den Arbeitsschritt ‹Antwort geben› konnten zu allen analysierten Kategorien Itemmerkmale gefunden werden, die signifikant mit der Itemschwierigkeit korrelieren. Jedoch lassen sich nicht alle Zusammenhänge direkt – der ersten Intuition folgend – oder indirekt – unter Berücksichtigung besonderer Umstände der Testkonstruktion – plausibel interpretieren. Dies betrifft die Verwendung von Abbildungen für inhaltliche Inputs und die Verwendung offener Lückenformate. Offene Lückenformate erweisen sich als schwierigkeitsreduzierend, während Abbildungen die Schwierigkeit erhöhen. Die restlichen Resultate lassen sich plausibel interpretieren und geben aus Untersuchungen von Papier-und-Bleistifttests bekannte Zusammenhänge wieder: Aufgaben werden schwieriger beurteilt, wenn sich die Kodierung grundsätzlich an den Evidenzen orientiert und dabei messtechnisch-praktische Anforderungen sowie die Genauigkeit von Messresultaten kodiert sowie wenn die Vollständigkeit von Antworten hinsichtlich der Lückenformate und der inhaltlichen Vorgaben bewertet werden. Aufgaben werden einfacher beurteilt, wenn die Antwort mit Stichworten oder durch Ankreuzen gegeben werden kann. Items mit vielen verschiedenen Informationsformaten erwiesen sich als besser lösbar, als kurze, knapp beschriebene Aufgaben.

Der Arbeitsschritt ‹Aufgabe erfassen› erweist sich stark abhängig von der Anzahl der in der Aufgabenstellung enthaltenen Formattypen. Items mit vielen Formattypen sind signifikant leichter als Items mit wenigen Formattypen. Zusätzliche Beschreibungen des allgemeinen Problems der Aufgabenstellung wirken zudem unterstützend.

Das Arbeitsschritt ‹Antwort geben› wird erleichtert, wenn das Gesuchte ausführlich beschrieben wird. Wird eine Beschreibung eines Sachverhalts gefordert, ist dies schwieriger, als wenn nur einzelne Stichworte oder gar nur das Ankreuzen der richtigen Antwort verlangt wird.

2.4. Sensitivität des HarmoS-Experimentiertests. Die Testleistung beruht nur zu einem Teil auf experimentellen Fähigkeiten. Mit dem Experimentiertest werden auch Lese- und Schreibkompetenzen mitgemessen, die mit entsprechenden schwierigkeitsrelevanten Itemmerkmalen in Verbindung gebracht werden können; ein Ergebnis übrigens, das für einen Test mit gedruckter Experimentieranleitung, dessen Auswertung auf dem Eigenrapport durch die Testpersonen beruht, nicht unerwartet ist. Die Itemanalyse wird jedoch insofern relativiert, als dass nur rund 44% der Varianz durch Itemmerkmale erklärt werden, wobei gerade diejenigen Teilprozesse, die beim Experimentiertest stark gewichtet sind, in den Resultaten ungenügend abgebildet werden. Zudem lassen sich nicht alle gemessenen schwierigkeitserzeugenden bzw. -reduzierenden Wirkungen in konsistenter Form begründen. Die aufgeklärte Varianz bewegt sich hingegen auf dem Niveau anderer large-scale Tests, weshalb wir die Validität des Test als ausreichend gut bewerten.

Personen-Analysen:
Verteilung der naturwissenschaftlichen Interessen

3.1. Ausprägung der Fach- und Sachinteressen. Die Personenstichprobe ist weder für die ganze Schweiz noch für die einzelnen Sprachregionen repräsentativ. Die Verteilung der Interessen in Bezug auf die naturwissenschaftlichen Fächer und Sachthemen liegt hingegen nicht im Widerspruch zu bekannten Interessenstudien. Insofern liegt kein Grund vor, die Ergebnisse des HarmoS-Experimentiertests nicht zu verallgemeinern.

Personen-Test-Analysen:
Verteilung der experimentellen Kompetenz

4.1. Kompetenzdifferenz der Geschlechter. Der HarmoS-Experimentiertest misst keine Leistungsunterschiede zwischen den Mädchen und Jungen. Die Resultate stehen diesbezüglich im Einklang mit den Ergebnissen der TIMS-Studien.

4.2. Kompetenzzuwachs zwischen Schulstufen. Der Kompetenzzuwachs zwischen dem 6. und 9. Schuljahr fällt mit rund einem Drittel der Standardabweichung einer Schulstufe sehr gering aus. Das Ergebnis wird durch den Umstand relativiert, dass die 6. Klässler beim Lesen der Aufgabenstellung unterstützt wurden, während die 9. Klässler von keiner solchen Hilfe profitierten. Es ist auch möglich, dass die 6. Klässler beim Experimentieren mehr Assistenz erhielten als die 9. Klässler. Das Resultat mag aber auch den Sachverhalt wiedergeben, dass der HarmoS-Experimentiertest Kompetenzen misst, die in den aktuellen Lehrplänen der Sekundarstufe I noch nicht enthalten sind und erst mit Inkrafttreten des Lehrplans 21 einmal implementiert werden.

4.3. Kompetenzdifferenz der Sprachregionen. Die Westschweiz fällt auf beiden Schulstufen mit signifikant besseren Testleistungen auf. Dies steht konträr zu den Ergebnissen des grösseren HarmoS-Validierungstests 2007. Der Kompetenzunterschied beträgt jedoch nur ein Sechstel der regionalen Standardabweichungen. Inwieweit das Resultat sich auf unterschiedliche Handhabung bei der Gewährung von Assistenz beim Experimentieren zurückführen lässt, kann im Nachhinein leider nicht mehr beurteilt werden.

4.4. Kompetenzdifferenz und Anforderungsniveaus. Zwischen dem höchsten, progymnasialen und dem mittlern Anforderungsniveau auf der Sekundarstufe wird eine signifikante Kompetenzdifferenz von rund einem Fünftel Standradabweichung gemessen. Zwischen dem tiefsten und dem mittleren Niveau wird eine Differenz von vergleichbarer Grösse gemessen. Sie erweist sich jedoch als nicht signifikant.

4.5. Kompetenz, Migrations- und Bildungshintergrund. Der HarmoS-Experimentiertest bestätigt bekannte bildungspolitische Zusammenhänge: Der Migrationshintergrund und der familiäre Bildungshintergrund stellen gute Indikatoren für die Testleistung dar. Die Ergebnisse bestätigen aus den PISA-Studien wohl bekannte Zusammenhänge.

4.6. Kompetenz, naturwissenschaftliches Fach- und Sachinteresse. Sowohl zwischen der experimentellen Kompetenz und dem Fachinteresse an Naturwissenschaften als auch zwischen der experimentellen Kompetenz und dem naturwissenschaftlichen Sachinteresse besteht eine Zusammenhang.

4.7. Kompetenz und Einschätzung der Testschwierigkeit. Die Einschätzung der Schwierigkeit des Tests durch die getesteten Schülerinnen und Schüler korreliert hoch signifikant mit der Testleistung. Die Schülerinnen und Schüler schätzen ihre Fähigkeiten sehr gut ein.

Kapitel 11

Kritischer Rückblick

Ein Forschungsprojekt beginnt man üblicherweise mit der Klärung der Fragestellung und Definition der Methodik, bevor die Testinstrumente entwickelt werden. Bei der vorliegenden Post hoc-Analyse des HarmoS-Experimentiertests war die Reihenfolge umgekehrt. Die Forschungsfragen mussten so gewählt werden, dass sie mit den gegebenen Daten sinnvoll beantwortet werden können. Und die Auswertungsmethoden mussten den Gegebenheiten des Experimentiertests angepasst werden. Wegen der Umkehrung dieser Reihenfolge gestaltete sich die Analyse der Itemschwierigkeit und letztlich die Erklärung der Testleistung schwieriger als zu Beginn der Studie angenommen. Auf die verschiedenen Schwierigkeiten wollen wir in einem kritischen Rückblick eingehen.

Inhaltsabhängigkeit der Messung. Bei der Entwicklung des HarmoS-Experimentiertests wurden verschiedene Anforderungen formuliert, die eine gute Experimentieraufgabe erfüllen sollte. U. a. sollte möglichst wenig fachliches Vorwissen erforderlich sein. Im Einzelfall wurde diese Bedingung dadurch erfüllt, dass spezifisches Hintergrund- und Vorwissen im Itemstamm oder global im Aufgabenstamm vorgegeben wurde. Das einzelne Experimentieritem sollte aber experimentelle Fähigkeiten ansprechen ohne Fachwissen abzufragen. Die Analyse der Kodierschemen zeigt nun aber ein gegenteiliges Bild vom Experimentiertest: Bei der Kodierung von mehr als einem Drittel der Items spielt die theoretische Korrektheit der Antwort eine Rolle (die relative Häufigkeit des Itemmerkmals {Kod-T} beträgt 38%). Dieser Sachverhalt korrespondiert mit dem im Kapitel 4 diskutierten Inhaltsproblem, wonach unterschiedliches fachliches Vorwissen unterschiedliche Teilprozesse erfordert. Die Beurteilung der Prozessqualität sollte deshalb individuell von den vorhandenen Präkonzepten abhängig gemacht werden. Dass aber bei HarmoS dieser Lösungsansatz nicht konsequent umgesetzt wurde, ist auch ein Grund für die grosse Häufigkeit des Itemmerkmals {Kod-T}.

Aufgabenheterogenität. Schwierigkeiten bei der Analyse der Itemschwierigkeit hängen einerseits mit der Heterogenität des HarmoS-Experimentiertests, andererseits mit der Verwendung von Aufgabenstämmen zusammen. Die Heterogenität hat zur Folge, dass die Auswahl der analysierten Itemmerkmale eingeschränkt werden musste. Bestimmte wünschbare Merkmale konnten nicht verwendet werden, da auf sie zu wenig Items passen.

Aufgabenstamm. Eine Eigenheit des HarmoS-Experimentiertests ist die Verwendung von Aufgabenstämmen. Die in den Stämmen enthaltenen Informationen beziehen sich auf sämtliche Items einer Experimentieraufgabe. Entsprechend müssten die Informationsformate des Stamms eigentlich allen Items zugeordnet werden. Dadurch würde diesen Formaten aber ein zu grosses Gewicht zugeteilt. Deshalb wurde der Stamm bei der Formatanalyse nicht berücksichtigt. Sein Einfluss auf die Itemschwierigkeit wurde aber unter das für alle Items einer Aufgabe gemeinsame Kontextmerkmal subsummiert.

Itemabhängigkeit. Als Messinstrumente werden beim HarmoS-Experimentiertest integrale Experimentieraufgaben verwendet, die im Idealfall sämtliche experimentellen Teilprozesse umfassen. Dieser durch die TIMS-Studien in der Schweiz bekannte Testansatz gewährleistet eine möglichst authentische Messung, produziert im Gegenzug aber Itemabhängigkeiten, wenn Teilprozesse separat kodiert werden. Der Ansatz eignet sich daher nicht, um die Teilprozesse zu unterscheiden, wie das negative Ergebnis der entsprechenden Dimensionsanalyse auch bestätigt.

Für die Analyse der Itemschwierigkeit stellen die Itemabhängigkeiten insofern ein Problem dar, als dass die Zusammenfassung der abhängigen Items zu Testlets die Itemzahl reduziert. Dadurch wird die Anzahl der sinnvollerweise zu analysierenden Merkmale vermindert und die Signifikanz der Ergebnisse wird beeinträchtigt. Wie durch die Itemanalysen bestätigt wird, werden mit der Testlet-Bildung nicht mehr oder andere Zusammenhänge aufgedeckt, als in der vollständigen Itemstichprobe nicht bereits vorhanden sind. Testlets können jedoch helfen, artifizielle, durch die Itemabhängigkeiten verursachte Zusammenhänge zu verhindern.

Mängel des Itemmerkmalkatalogs. Das verwandte Merkmalsystem ist zwar sehr umfangreich, es erfasst die für die Kompetenzmessung entscheidenden Progressionsdimensionen jedoch nicht in zufriedenstellendem Masse. Dies trifft im Besonderen auf die Problemkomplexität $[P]$ und die Prozessqualität $[Q]$ zu.

Die neun Merkmale ({Aufgabentyp}, {Problemoffenheit}, {Zielklarheit}, {Strukturiertheit} und fünf weitere Manipulationsmerkmale), mit welchen die Problemkomplexität erfasst werden sollte, erweisen sich statistisch als unbedeutend. Die Vermutung liegt nahe, dass diese Problemmerkmale bei verschiedenen Aufgabentypen unterschiedlich auf

die Itemschwierigkeit wirken. So wie bestimmte Teilprozesse bei verschiedenen Aufgabentypen unterschiedlich ausgeprägt sind, so spielen bestimmte Problemmerkmale bei verschiedenen Aufgabentypen unterschiedlich grosse Rollen. Die Analyse der Dimension Problemkomplexität muss wohl für jeden Aufgabentyp separat gemacht werden.

Ebenfalls unbefriedigend ist die Modellierung der Prozessqualität. Die verwandte Unterscheidung der Kodiermassstäbe bleibt oberflächlich. Auf die Analyse der den einzelnen Massstäben zugrundeliegenden konkreten Kriterien wurde wegen der starken Heterogenität der Kodierschemen im Rahmen dieser Studie verzichtet. Um das Konstrukt Prozessqualität besser zu verstehen, sind aber Analysen auf der Ebene der einzelnen Kriterien notwendig wie u. a. die konzeptuelle Analyse von "Experimental competence rubrics" von Vogt, Müller und Kuhn (2011).

Die Objektivität der Skalen ist zudem nicht geprüft worden. Das Rating der Itemmerkmal wurde von mir alleine durchgeführt.

Kapitel 12

Ausblick

Die Eidgenössische Konferenz der kantonalen Erziehungsdirektoren (EDK) plant mittelfristig ein nationales Bildungsmonitoring, mit welchem die gesetzlich verankerten Mindeststandards (Grundkompetenzen) in regelmässigen Abständen evaluiert würden. Aus fachdidaktischer Sicht ist es erstrebenswert, wenn die Politik dannzumal die finanziellen Ausgaben nicht scheute, dem HarmoS-Test weitere nationale Experimentiertests nachfolgen zu lassen. Der in dieser Arbeit analysierte Experimentiertest ist ein erster Entwurf eines authentischen, jedoch noch zu wenig aussagekräftigen Messinstruments für experimentelle Kompetenzen. Im Hinblick auf die Verwendung im Rahmen eines Bildungsmonitorings müsste dieses Messinstrument weiterentwickelt werden. Besondere Aufmerksamkeit müsste dabei folgenden Problemen und Fragen zukommen.

Problem der Übersetzung vom Standarddiskurs in den Assessmentdiskurs. Die geltenden Schweizer Standards, die so genannten Grundkompetenzen für die Naturwissenschaften (EDK, 2011), weichen gegenüber den Vorschlägen des Konsortiums HarmoS Naturwissenschaften teilweise erheblich ab. Für das Bildungsmonitoring müssten die Standards neu operationalisiert werden, wobei die in den Standards enthaltenen Progressionsdimensionen adäquat in einem Assessment abzubilden wären.

Problematisch erscheint der Umgang mit der für den Unterricht relevanten Eigenständigkeit im Rahmen eines large-scale Assessments wie dem HarmoS-Experimentiertest. Hier müssten andere Bewertungsinstrumente entwickelt werden, dies vielleicht wie beim englischen General Certificate of Secondary Education unter Einbezug der Lehrpersonen und Assessments direkt im Unterricht (R. W. Fairbrother, 1988; Millar, 2007).

Für die anderen Progressionsdimensionen ist eine verbindliche und kommunizierte Standardisierung erforderlich. Für den Transferumfang wird diese durch den Lehrplan 21 einmal abschliessend definiert sein. Die Modellierung der Dimensionen zur Problemkomplexität, Prozessqualität und zum Aufgabenumfang im Rahmen des HarmoS-Experimen-

tiertests erscheint hingegen aufgrund der vorliegenden Analyse als problematisch und bedingt im Hinblick auf das Bildungsmonitoring zusätzliche konzeptuelle Entwicklungsarbeit. Folgende Erkenntnisse aus den Analysen des HarmoS-Experimentiertests gilt es hierbei zu berücksichtigen.

- *Aufgabentyp versus Teilprozess*: Die Standards beziehen sich gleichzeitig auf unterschiedliche Aufgabentypen und Teilprozesse dieser Aufgabentypen. Es muss geklärt werden, inwieweit einzelne Teilprozesse oder ganze Aufgaben evaluiert werden sollen. Je nach Fokus werden andere Messinstrumente benötigt. Die Analysen zeigen, dass sich der aus TIMSS bekannte integrale Messansatz mit vollständigen Experimentieraufgaben wegen der Itemabhängigkeiten nicht eignet, Teilprozesse statistisch zu differenzieren. Die Abgrenzung der Aufgabentypen funktioniert hingegen trotz Itemabhängigkeiten. Die psychometrisch korrekte Lösung für Itemabhängigkeiten wäre die Bildung von Testlets oder die beim NAEP praktizierte Kodierung, bei welcher Aufgaben de integro als ein Item behandelt werden (Brown & Shavelson, 1996).

- *Spezifizierung experimenteller Aufgabentypen*: Die effiziente Aufgabenkonstruktion für zukünftige Tests wie auch für den Gebrauch im Schulalltag bedingt die Spezifizierung und Standardisierung von Aufgabentypen. Unterschiedliche Aufgabentypen wie Beobachtungen anstellen oder systematische Untersuchungen durchführen scheinen unterschiedliche Teilkompetenzen einer übergeordneten experimentellen Kompetenz anzusprechen. Sinnvollerweise umfasst die Standardisierung neben der Aufgabenstellung auch die Kodierschemen.

- *Standardisierung der Kodiermassstäbe.* Für die Validität einer Messung sind die in den Kodierschemen verwendeten Kodiermassstäbe und ihre Verknüpfung essentiell. Sie sind für die Interpretation von Testresultaten unabdingbar. Und mit ihnen werden a priori Niveaustufen festgesetzt. Die Standardisierung der Kodierschemen ist Teil der Formulierung von Kompetenzniveaus.

- *Standardisierung kompetenzirrelevanter Itemmerkmale*: Die Testleistung bei Performance Assessments von der Art des HarmoS-Experimentiertests hängt wesentlich von nicht kompetenzspezifischen Fähigkeiten wie Lese- und Schreibkompetenz ab. Die Variation bestimmter kompetenzirrelevanter Itemmerkmale bewirkt signifikante Variationen der Itemschwierigkeit. Die Konfundierung der Kompetenzmessung mit irrelevanten Fähigkeiten lässt sich zwar nicht verhindern, mit der Standardisierung kompetenzirrelevanter Itemmerkmale wird der Einfluss auf die Testleistung zumindest konstant gehalten.

Problem der Progressionsmodellierung. Es ist ein Mangel des HarmoS-Experimentiertests, dass die Ursachen für Leistungsunterschiede bislang nicht erklärt werden können. Wegen der starken Heterogenität des Tests – diese betrifft sowohl die Aufgabenstellungen als auch die Kodierschemen – wurde in dieser Arbeit der Versuch nicht unternommen, post hoc empirische Niveaustufen ausfindig zu machen. Dieses Unterfangen gilt aus Erfahrungen mit large-scale Assessments wie PISA ganz allgemein als wenig erfolgsversprechend. Bessere Resultate wurden bisher dann erreicht, wenn a priori Kompetenzstufen festgelegt wurden (cf. Neumann et al., 2007, 103). Jedoch lieferte auch die Analyse der Itemschwierigkeit keinen Anhaltspunkt für Niveaustufen.

Grundsätzlich gibt es zwei Ansätze für Stufungen: Niveaus werden entweder über die Aufgabenstellung oder über die Kodierschemen definiert. Für beide Varianten existieren erfolgreiche Beispiele (e. g. Mayer, 2007; Kauertz, 2008). Die Verwendung von integralen Experimentieraufgaben als Messinstrumente scheint prima facie kompatibler zu sein mit Stufungen, die über die Prozessqualität definiert werden. Denkbar ist jedoch auch die Kombination beider Ansätze, d. h. die Verwendung einer kombinierten Progression (cf. Abs. 3.2.3.1 , S. 50). Dies würde bedeuten, dass man zu jedem Aufgabentyp Aufgabenstellungen von unterschiedlicher Problemkomplexität oder unterschiedlichem Aufgabenumfang definiert. Zu jeder Aufgabenstellung mit einem bestimmten Hauptniveau werden über die Kodierschemen zusätzliche Unterniveaus der Prozessqualität spezifiziert. Grundsätzlich könnten die Unterniveaus auch Aspekte der Eigenständigkeit abbilden. Auf jeden Fall erhält man verschiedene Progressionen; zu jedem Hautniveau eine, die sich mehr oder weniger überlagern können. Welche Unterniveaus welchen Hauptniveaus zu derselben Stufe gehören, muss letztlich empirisch anhand der Itemschwierigkeit festgestellt werden. In diesem Konzept noch nicht berücksichtigt ist die Dimension des Transferumfangs. Sein Einbezug würde bedingen, dass zu jedem Aufgabentyp und Hauptniveau mehrere Beispiele aus verschiedenen Kontexten zur Verfügung stünden. Die Umsetzung des skizzierten Programms im Rahmen eines Bildungsmonitoring erforderte theoretische und praktische Entwicklungsarbeit, die nur langfristig zu leisten wäre.

Problem der kompetenzirrelevanten Itemschwierigkeit. Experimentiertests wie der HarmoS-Experimentiertest mit gedruckter Aufgabenstellung und Selbstrapportierung messen prinzipell immer auch Lese- und Schreibfähigkeiten. Lösungen dieses Problems sollten stärker erforscht werden, die Vor- und Nachteile verschiedener Testarten (Hands-on-, Papier-und-Bleistift- oder Simulationstests) sowie von verschiedenen Messarten (Selbst- oder Fremdrapportierung) sollten fundierter untersucht werden.

Inhaltsproblem der experimentellen Kompetenzmessung. Die Lösung des Inhaltsproblems bei der Messung von Prozesskompetenzen ist nicht trivial und bedingt noch nicht vorhandene theoretische Grundlagenarbeit zu Kodiermassstäben. Die gleichen experimentellen Teilprozesse, eingebettet in unterschiedliche Aufgabentypen, können nicht gleich behandelt werden. Um die Qualität dieser differenzierten Teilprozesse besser und adäquater beurteilen zu können ist mehr konzeptuelle Forschung zu Kodiermassstäben und -kriterien notwendig.

Kontinuierliche Aufgabenkonstruktion. Der entscheidende Punkt des Bildungsmonitorings ist, dass die Standards im Unterricht umgesetzt werden. Am ehesten gelingt die Implementation über Standardaufgaben, die Lehrpersonen als Übungs- und Prüfungsbeispiele zur Verfügung gestellt werden. Damit wird ein Teaching-to-test im positiven Sinne initiiert. Der nachhaltige Reformprozess hin zum kompetenzorientierten Unterricht bedingt hingegen, dass mit der Zeit eine umfangreiche und vielfältige Aufgabendatenbank erarbeitet wird. Zu erhoffen ist daher, dass im Rahmen eines wiederkehrenden Bildungsmonitorings von der Politik die Ressourcen für die kontinuierliche Aufgabenkonstruktion bereitgestellt werden, damit die Innovation der Reformen nicht verloren geht.

Teil IV

Appendix

Anhang A

Aufgabenbeispiele aus dem HarmoS-Experimentiertest

A.1 ⟨Balkenwaage⟩

Balkenwaage

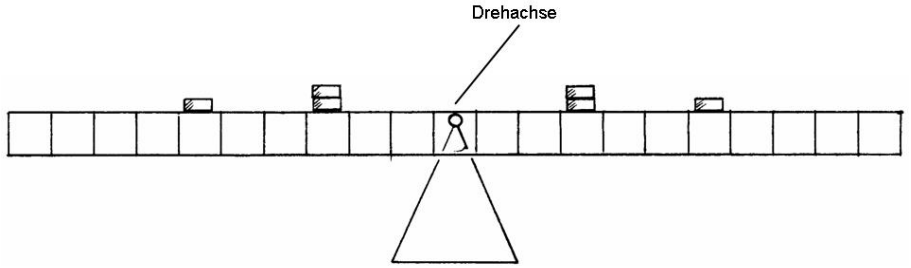
Finde durch Experimentieren und Kombinieren von Gesetzen heraus, wann eine Balkenwaage im Gleichgewicht ist.

Bei diesem Experiment findest Du folgendes Material vor:

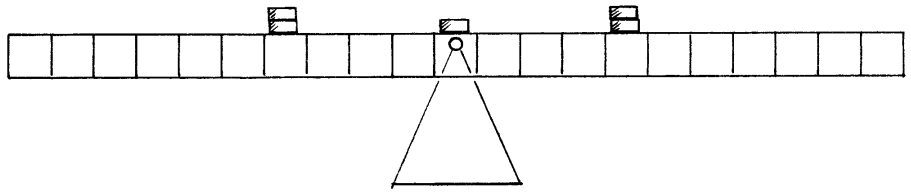
- eine Balkenwaage,
- 6 Schraubenmuttern.

Informationen

Es gibt sehr viele Arten, die sechs Gewichtsstücke (= Schraubenmuttern) auf die Felder der Balkenwaage zu verteilen. Nur in wenigen Fällen bleibt die Waage im Gleichgewicht.
Wenn die Gewichtsstücke auf beiden Seiten der Waage gleich angeordnet sind, dann ist die Waage symmetrisch belastet. Zwei Bilder helfen dir als Beispiele:



symmetrisch belastete Waage mit sechs Gewichtsstücken



symmetrisch belastete Waage mit fünf Gewichtsstücken

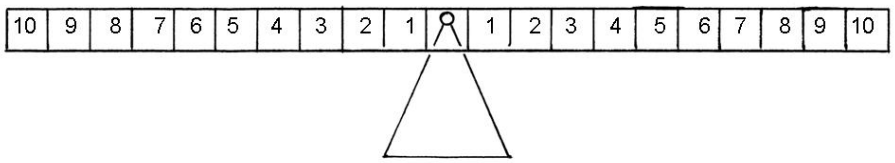
Die Aufgaben

Auf den nächsten sechs Seiten werden sechs Behauptungen aufgestellt. Unter diesen gibt es richtige und falsche Behauptungen. Überprüfe durch Experimentieren, welche Behauptungen richtig und welche falsch sind.

Behauptung 1
Eine symmetrisch belastete Waage befindet sich immer im Gleichgewicht.

Führe zwei Experimente durch, um die Behauptung zu überprüfen.
Zeichne diese in die Abbildungen ein.
Kreuze an, ob die Waage beim Experiment im Gleichgewicht ist.

1. Experiment:

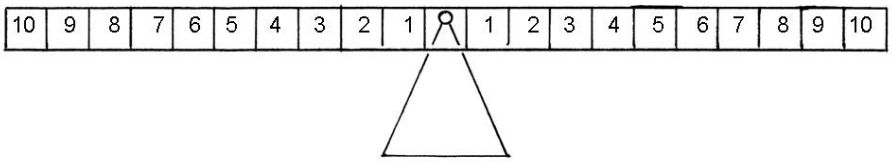


☐ Die Waage ist im Gleichgewicht.

☐ Die Waage ist nicht im Gleichgewicht.

⟨N1E23i01⟩

2. Experiment:



☐ Die Waage ist im Gleichgewicht.

☐ Die Waage ist nicht im Gleichgewicht.

⟨N1E23i02⟩

Was haben deine Experimente ergeben? Kreuze an!

☐ Die Behauptung ist richtig.

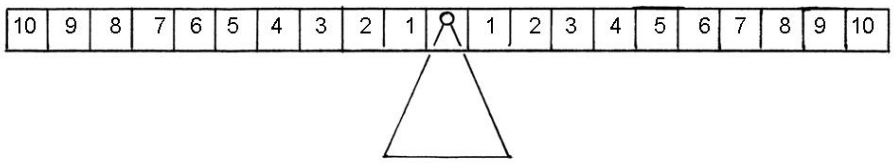
☐ Die Behauptung ist falsch.

⟨N1E23i03⟩

Behauptung 2
Wenn eine Waage im Gleichgewicht ist, dann ist sie immer symmetrisch belastet.

Führe zwei Experimente durch, um die Behauptung zu überprüfen.
Zeichne diese in die Abbildungen ein.
Kreuze an, ob die Waage beim Experiment im Gleichgewicht ist.

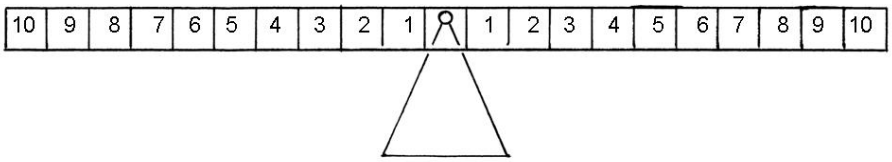
1. Experiment:



- ☐ Die Waage ist im Gleichgewicht.
- ☐ Die Waage ist nicht im Gleichgewicht.

⟨N1E23i04⟩

2. Experiment:



- ☐ Die Waage ist im Gleichgewicht.
- ☐ Die Waage ist nicht im Gleichgewicht.

⟨N1E23i05⟩

Was haben deine Experimente ergeben? Kreuze an!

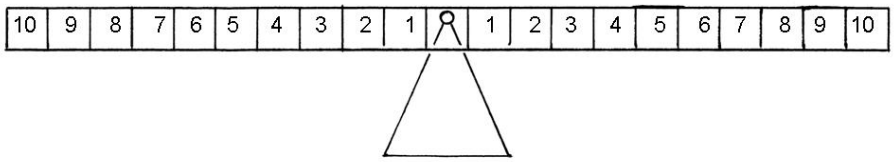
- ☐ Die Behauptung ist richtig.
- ☐ Die Behauptung ist falsch.

⟨N1E23i06⟩

Behauptung 3
Diese Veränderung stört das Gleichgewicht einer Waage nicht:
Auf beiden Seiten wird im gleichen Abstand zur Drehachse ein Gewichtsstück entfernt.

Führe zwei Experimente durch, um die Behauptung zu überprüfen.
Zeichne in die gleiche Abbildung die ***Situation vor der Veränderung mit schwarzer Farbe*** und ***nach der Veränderung mit roter Farbe***.
Kreuze für jede Situation an, wann die Waage im Gleichgewicht ist.

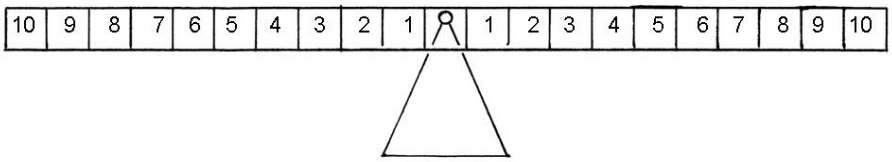
1. Experiment:



- ☐ Die Waage ist im Gleichgewicht.
- ☐ Die Waage ist nicht im Gleichgewicht.

⟨N1E23i07⟩

2. Experiment:



- ☐ Die Waage ist im Gleichgewicht.
- ☐ Die Waage ist nicht im Gleichgewicht.

⟨N1E23i08⟩

Was haben deine Experimente ergeben? Kreuze an!

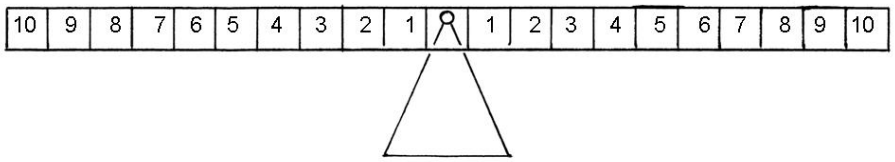
- ☐ Die Behauptung ist richtig.
- ☐ Die Behauptung ist falsch.

⟨N1E23i09⟩

Behauptung 4
Diese Veränderung stört das Gleichgewicht einer Waage nicht:
Ein Gewichtsstück wird auf der Waage nach links und ein anderes Gewichtsstück um die gleiche Strecke nach rechts verschoben.

Führe zwei Experimente durch, um die Behauptung zu überprüfen.
Zeichne in die gleiche Abbildung die ***Situation vor der Veränderung mit schwarzer Farbe*** und ***nach der Veränderung mit roter Farbe***.
Kreuze für jede Situation an, wann die Waage im Gleichgewicht ist.

1. Experiment:

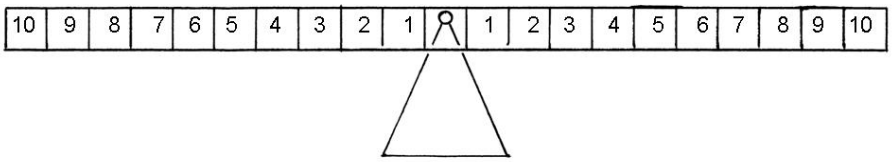


☐ Die Waage ist im Gleichgewicht.

☐ Die Waage ist nicht im Gleichgewicht.

⟨N1E23i10⟩

2. Experiment:



☐ Die Waage ist im Gleichgewicht.

☐ Die Waage ist nicht im Gleichgewicht.

⟨N1E23i11⟩

Was haben deine Experimente ergeben? Kreuze an!

☐ Die Behauptung ist richtig.

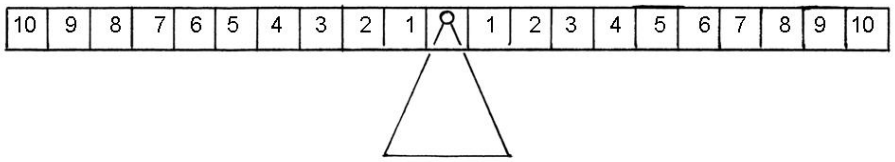
☐ Die Behauptung ist falsch.

⟨N1E23i12⟩

Behauptung 5
Belaste die Waage auf der einen Seite mit 2 Gewichtsstücken und auf der anderen Seite mit 3 Gewichtsstücken. Dann kann die Waage auf keinen Fall im Gleichgewicht sein.

Führe zwei Experimente durch, um die Behauptung zu überprüfen.
Zeichne diese in die Abbildungen ein.
Kreuze an, ob die Waage beim Experiment im Gleichgewicht ist.

1. Experiment:

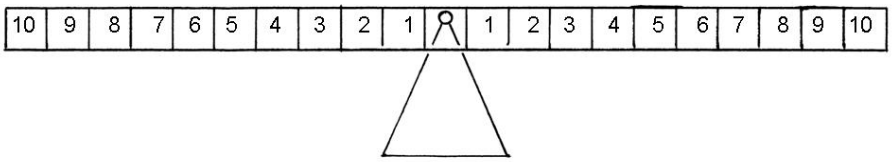


☐ Die Waage ist im Gleichgewicht.

☐ Die Waage ist nicht im Gleichgewicht.

⟨N1E23i13⟩

2. Experiment:



☐ Die Waage ist im Gleichgewicht.

☐ Die Waage ist nicht im Gleichgewicht.

⟨N1E23i14⟩

Was haben deine Experimente ergeben? Kreuze an!

☐ Die Behauptung ist richtig.

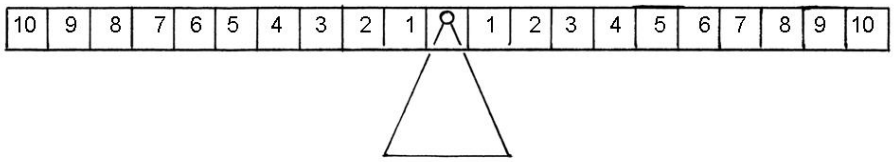
☐ Die Behauptung ist falsch.

⟨N1E23i15⟩

Behauptung 6
Belaste die Waage nur mit einem Gewichtsstück.
Dann kann die Waage auf keinen Fall im Gleichgewicht sein.

Führe zwei Experimente durch, um die Behauptung zu überprüfen.
Zeichne diese in die Abbildungen ein.
Kreuze an, ob die Waage beim Experiment im Gleichgewicht ist.

1. Experiment:

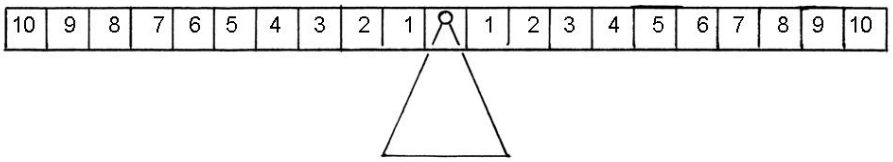


☐ Die Waage ist im Gleichgewicht.

☐ Die Waage ist nicht im Gleichgewicht.

⟨N1E23i16⟩

2. Experiment:



☐ Die Waage ist im Gleichgewicht.

☐ Die Waage ist nicht im Gleichgewicht.

⟨N1E23i17⟩

Was haben deine Experimente ergeben? Kreuze an!

☐ Die Behauptung ist richtig.

☐ Die Behauptung ist falsch.

⟨N1E23i18⟩

Balkenwaage

KODIERSCHEMA

Kodierungsaspekte

Zu jeder Behauptung werden zwei Aspekte bewertet.

a) ***Experimente**: Die Schülerinnen und Schüler sollen zu jeder Behauptung zwei Experimente durchführen, die sich eignen, um die Behauptung zu überprüfen. Unter einem Experiment wird die spezifische Verteilung der Gewichtsstücke auf der Balkenwaage verstanden.*

*Experimente, deren Ausgang in jedem Fall entweder zu einer Bestätigung oder zu einer Widerlegung der Behauptung führen, werden als **adäquat** bezeichnet. Ein adäquates Experiment kann auch zu einem falschen Schluss führen, nämlich dann, wenn es eine falsche Behauptung scheinbar bestätigt.*

*Nicht alle Experimente, die eine Behauptung bestätigen, sind adäquat. Es gibt Experimente, die auf keinen Fall die Behauptung widerlegen können. D. h., wie auch immer das Experiment ausgeht, es kann nicht als Gegenbeispiel dienen (vgl. Behauptung 2). Solche Experimente werden als **semi-adäquat** bezeichnet.*

*Experimente, die weder adäquat noch semi-adäquat sind, werden als **nicht adäquat** bezeichnet.*

Weiter soll bei der Kodierung berücksichtigt werden, ob die Qualifizierung des Gleichgewichts (Ausgang des Experiments) und die Gewichtsverteilung (Experiment) theoretisch zusammenpassen.

b) ***Gültigkeit der Behauptung**: Die Schülerinnen und Schüler sollen anhand der Experimente die Gültigkeit der Behauptung korrekt beurteilen. Hierbei besteht die Möglichkeit, dass der korrekte Schluss zu einem theoretisch falschen Ergebnis führt (Code 1), Dies ist der Fall, wenn zu einer falschen Behauptung kein Gegenbeispiel gefunden wird.*

Allgemeine Korrekturregeln

- *Als Beispiel bezeichnen wir eine Belastung der Waage mit den zur Verfügung stehenden Gewichtsstücken.*
- *Beispiele mit mehr als sechs Gewichtsstücken zählen als falsche Beispiele.*
- *Ein Beispiel kann nur bewertet werden, wenn die Anzahl Gewichtsstücke und die dazugehörigen Hebelarme eindeutig eruiert werden können und das Gleichgewicht qualifiziert wird. Unklare oder unvollständige Beispiele werden als «falsch» gewertet (Code 0).*
- *Identische Beispiele zählen nur einmal.*
- *Eine Umwandlung bezeichnet eine Veränderung der Belastung der Waage durch Wegnehmen, Hinzufügen oder Verschieben von Gewichtsstücken.*
- *Umwandlungen, deren Anfangs- oder/und Endbelastung mehr als 6 Gewichtsstücke beinhalten, werden als falsch taxiert.*
- *Unklar beschriebene Umwandlungen werden als «falsch» gewertet (Code 0).*

Behauptungen 1-6

...

Führe zwei Experimente durch, um die Behauptung zu überprüfen.
Zeichne sie in die Abbildungen ein.
Kreuze an, ob die Waage beim Experiment im Gleichgewicht ist.

1. Experiment:

⟨N1E23i01⟩, ⟨N1E23i04⟩, ⟨N1E23i07⟩, ⟨N1E23i10⟩, ⟨N1E23i13⟩, ⟨N1E23i16⟩

Matrix 1	adäquates Experiment	semi-adäquates Experiment	nicht adäquates Experiment	kein Experiment = keine Belastung angegeben
Gleichgewicht richtig qualifiziert	**Code 2**	**Code 1***	**Code 0**	**Code 9**
Gleichgewicht falsch qualifiziert	**Code 1**	**Code 0***	**Code 0**	**Code 9**
Gleichgewicht nicht qualifiziert	**Code 0**	**Code 0***	**Code 0**	**Code 9**

2. Experiment:

⟨N1E23i02⟩, ⟨N1E23i05⟩, ⟨N1E23i08⟩, ⟨N1E23i11⟩, ⟨N1E23i14⟩, ⟨N1E23i17⟩

Matrix 1	adäquates Experiment	semi-adäquates Experiment	nicht adäquates Experiment	kein Experiment = keine Belastung angegeben
Gleichgewicht richtig qualifiziert	**Code 2**	**Code 1***	**Code 0**	**Code 9**
Gleichgewicht falsch qualifiziert	**Code 1**	**Code 0***	**Code 0**	**Code 9**
Gleichgewicht nicht qualifiziert	**Code 0**	**Code 0***	**Code 0**	**Code 9**

* Dieser Fall gibt es nur für die Behauptungen 2 und 6.

Was haben deine Experimente ergeben?

☐ Die Behauptung ist richtig.
☐ Die Behauptung ist falsch.

⟨N1E23i03⟩, ⟨N1E23i06⟩, ⟨N1E23i09⟩, ⟨N1E23i12⟩, ⟨N1E23i15⟩, ⟨N1E23i18⟩

Matrix 2	korrekt gemäss Theorie	falsch gemäss Theorie
Antwort vereinbar mit den Experimenten	**Code 2**	**Code 1**
Antwort nicht vereinbar mit den Experimenten	**Code 0**	**Code 0**
Antwort bezieht sich nur auf nicht adäquate oder unvollständige Experimente	**Code 0**	**Code 0**
keine Antwort	**Code 9**	**Code 9**

A.2 ⟨Solarzellen⟩

Solarzellen

Finde heraus, wie man mit zwei Solarzellen einen Ventilator am besten betreibt.

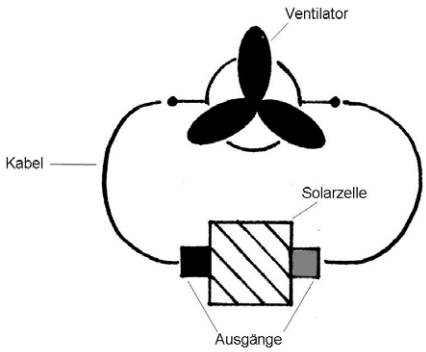
Bei diesem Experiment findest Du folgendes Material vor:

- 2 Solarzellen
- 1 Ventilator
- 4 Kabel
- eine Tischlampe
- ein Massband
- schwarzes Papier
- Schreibzeug in roter und schwarzer Farbe (selber mitbringen)

Deine Aufgaben

Benutze eine Solarzelle, um mit dem Licht der Tischlampe den Ventilator zu betreiben.

Baue die abgebildete Schaltung auf. Stelle die Tischlampe so ein, dass das Licht senkrecht auf die Solarzelle fällt.



Finde heraus, wie nahe die Glühlampe an die Solarzelle gehalten werden muss, damit der Ventilator von selbst startet. Miss die Entfernung von der Solarzelle bis zum unteren Rand des Lampenschirms. Halte das Ergebnis hier fest.

⟨N9E84i02⟩

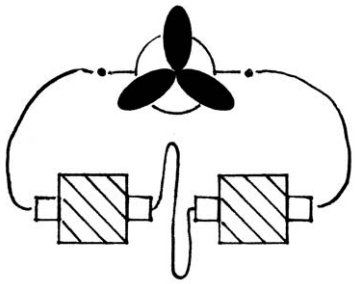
Finde heraus, wie weit die Glühlampe von der Solarzelle entfernt werden kann, ohne dass der Ventilator stoppt. Halte das Ergebnis hier fest.

⟨N9E84i03⟩

Benutze nun zwei Solarzellen, um den Ventilator zu betreiben. Es gibt zwei Arten, die Solarzellen zu schalten. Sie können entweder in Serie (*hintereinander*) oder parallel (*nebeneinander*) geschaltet werden.

Baue die unten abgebildete **Serie-Schaltung** auf.

Die Solarzellen haben einen roten und einen schwarzen Ausgang. Wie müssen die Ausgänge in der Serie-Schaltung zueinander stehen, damit der Ventilator am besten läuft?
Finde die beste Schaltung und zeichne die Farben der Ausgänge in die Abbildung ein.

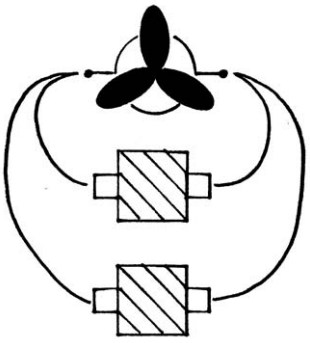


Serie-Schaltung

⟨N9E84i05⟩

Baue die unten abgebildete **Parallel-Schaltung** auf.

Wie müssen die roten und schwarzen Solarzellen-Ausgänge in der Parallel-Schaltung zueinander stehen, damit der Ventilator am besten läuft?
Finde die beste Schaltung und zeichne die Farben der Ausgänge in die Abbildung ein.



Parallel-Schaltung

⟨N9E84i07⟩

Finde heraus, bei welcher Schaltung (Serie- oder Parallel-Schaltung) der Ventilator mit weniger Licht betrieben werden kann.
Die Stärke des Lichteinfalls auf die Solarzellen veränderst du, indem du die Tischlampe unterschiedlich weit von den Solarzellen weg hältst.

Plane eine Messung, um die Frage zu entscheiden.

Schreibe deinen Plan auf. Gib an,
– wie du bei der Messung vorgehen willst und
– was du messen willst (genaue Angaben).

⟨N9E84i08⟩

Führe die Messung durch und halte die Messergebnisse fest.

⟨N9E84i09⟩

Welche Schaltung benötigt weniger Licht?

- Die Serie-Schaltung benötigt weniger Licht.
- Die Parallel-Schaltung benötigt weniger Licht.

⟨N9E84i10⟩

Solarzellen werden gelegentlich durch Gegenstände beschattet und liefern dadurch weniger Strom.
Untersuche, welche Schaltung (Serie- oder Parallel-Schaltung) durch Beschattung weniger beeinträchtigt wird.

Stelle die Tischlampe im Abstand von 20 cm zu den Solarzellen ein.
Decke mit dem schwarzen Papier schrittweise zuerst die erste Solarzelle und dann die zweite Solarzelle ab.

Halte das Ergebnis deiner Untersuchung fest.

- ☐ Die Serie-Schaltung wird durch Beschattung weniger beeinträchtigt.
- ☐ Die Parallel-Schaltung wird durch Beschattung weniger beeinträchtigt.

⟨N9E84i11⟩

Wie bist du auf dein Ergebnis gekommen?
Beschreibe, wie du beim Experimentieren vorgegangen bist und was du beobachtet hast!

⟨N9E84i12⟩

Solarzellen

KODIERSCHEMA

Benutze eine Solarzelle, um mit dem Licht der Tischlampe den Ventilator zu betreiben.

- Baue die abgebildete Schaltung auf. Stelle die Tischlampe so ein, dass das Licht senkrecht auf die Solarzelle fällt.

⟨N9E84i01⟩

Item wird nicht codiert

- Finde heraus, wie nahe die Glühlampe an die Solarzelle gehalten werden muss, damit der Ventilator von selbst startet. Miss die Entfernung von der Solarzelle bis zum unteren Rand des Lampenschirms. Halte das Ergebnis hier fest.

Der Ventilator startet, sobald die Lampe näher als 6-35cm an die Solarzelle herankommt.

⟨N9E84i02⟩

Code 1: Antwort innerhalb des Lösungsbereichs oder kleiner als Antwort von Item 3

Code 0: Antwort ausserhalb des Lösungsbereichs oder grösser als Antwort von Item 3

Code 9: keine Antwort

- Finde heraus, wie weit die Glühlampe von der Solarzelle entfernt werden kann, ohne dass der Ventilator stoppt. Halte das Ergebnis hier fest.

Der Ventilator stoppt, wenn die Lampe weiter als 25cm-50cm von der Solarzelle entfernt ist.

⟨N9E84i03⟩

Code 1: Antwort innerhalb des Lösungsbereichs

Code 0: Antwort ausserhalb des Lösungsbereichs

Code 9: keine Antwort

Benutze nun zwei Solarzellen, um den Ventilator zu betreiben. Es gibt zwei Arten, die Solarzellen zu schalten. Sie können entweder in Serie (*hintereinander*) oder parallel (*nebeneinander*) geschaltet werden.

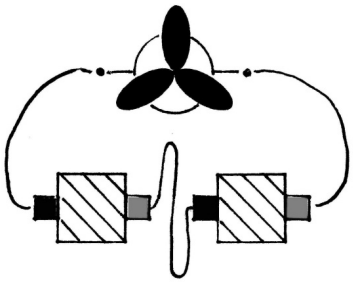
- Baue die abgebildete **Serie-Schaltung** auf.

⟨N9E84i04⟩

Item wird nicht codiert

- Die Solarzellen haben einen roten und einen schwarzen Ausgang. Wie müssen die Ausgänge zueinander stehen, damit der Ventilator am besten läuft? Finde die beste Schaltung und zeichne die Farben der Ausgänge in die Abbildung ein.

Die Solarzellen sind über einen roten und einen schwarzen Ausgang verbunden. Die Reihenfolge spielt keine Rolle.



⟨N9E84i05⟩

Code 1: richtige Antwort

Code 0: falsche Antwort

Code 9: keine Antwort

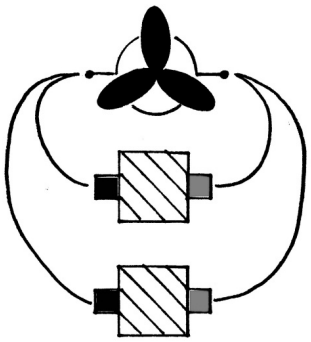
- Baue die abgebildete **Parallel-Schaltung** auf.

⟨N9E84i06⟩

Item wird nicht codiert.

- Wie müssen die roten und schwarzen Solarzellen-Ausgänge zueinander stehen, damit der Ventilator am besten läuft? Finde die beste Schaltung und zeichne die Farben der Ausgänge in die Abbildung ein.

Die roten Ausgänge sind miteinander verbunden. Dasselbe gilt für die schwarzen Ausgänge.



⟨N9E84i07⟩

Code 1: richtige Antwort

Code 0: falsche Antwort

Code 9: keine Antwort

- Vergleiche die Serie-Schaltung mit der Parallel-Schaltung. Finde heraus, bei welcher Schaltung der Ventilator mit weniger Licht betrieben werden kann. Die Stärke des Lichteinfalls auf die Solarzellen veränderst du, indem du die Tischlampe unterschiedlich weit von den Solarzellen weg hältst. Plane eine Messung, um die Frage zu entscheiden. Schreibe deinen Plan auf. Gib an,
  – wie du bei der Messung vorgehen willst,
  – was du messen willst und
  – wie du es genau messen willst.

a) Es werden mindestens zwei Messungen, je eine an der Serie- und eine an der Parallelschaltung vorgeschlagen, die einen Vergleich der beiden Schaltungen ermöglichen.

b) Es wird eine dazu adäquate Messmethode beschrieben.

***Methode 1**: Vergleich, wie weit darf die Lampe von den Solarzellen entfernt werden, bis der Ventilator stoppt.*

***Methode 2**: Vergleich, wie nahe muss die Lampe an die Solarzellen gehalten werden, damit der der Ventilator von selbst startet.*

***Methode 3**: Vergleich, wie schnell dreht der Ventilator bei gleicher Entfernung der Tischlampe.*

***weitere Methoden**: z.B. Einfluss der Beschattung auf Schaltungen*

c) Es wird der Methode entsprechend erklärt, was gemessen wird. Dieser Punkt kann auch aus der Beschreibung der Messmethode klar werden.

***Methode 1**: Entfernungsmessung: Solarzelle-Lampe*

***Methode 2**: Entfernungsmessung: Solarzelle-Lampe*

***Methode 3**: Einschätzung der Rotationsgeschwindigkeit des Ventilators*

***weitere Methoden**: entsprechend*

⟨N9E84i08⟩

Code 2:	vollständige Antwort: a), b) und c) ganz erfüllt
Code 1:	unvollständige Antwort: Ein Punkt von a) b) und c) nicht oder nur teilweise erfüllt
Code 0:	falsche Antwort: Kein Punkt vollständig erfüllt
Code 9:	keine Antwort

- Führe die Messung durch und halte die Messergebnisse fest.

MATRIX 1	***Entfernungs-Methode 1**: Stoppen beim Entfernen der Lampe*	***Annäherungs-Methode 2**: Starten beim Annähern der Lampe*	***Geschwindigkeits-Methode 3**: Geschwindigkeitsvergleich bei konstanter Lampenentfernung*	***Beschattungs-Methode 4**: maximaler Abdeckungsgrad beim Beschatten*
Serieschaltung	*35cm – 55cm*	*10cm – 40cm*	*Von Serie Umbau auf Parallel → Motor läuft langsamer*	*1 Zelle*
Parallelschaltung	*50cm – 65cm*	*10cm – 40cm*	*Von Parallel Umbau auf Serie → Motor läuft schneller*	*mehr als 1 Zelle*
***Ergebnis**: Weniger Licht benötigt die …*	***Parallelschaltung***	*je nach Experimentierset die Parallel- oder die Serieschaltung*	***Serieschaltung***	***Parallelschaltung***

⟨N9E84i09⟩

Code 1: Messergebisse einer vergleichenden Messung (Messmethode aus Item 8 bekannt) festgehalten

UND

Messergebnisse entsprechen den Angaben der Matrix 1

Code 0: andere Antworten

Code 9: keine Antwort

- Welche Schaltung benötigt weniger Licht?

☐ Die Serie-Schaltung benötigt weniger Licht.
☐ Die Parallel-Schaltung benötigt weniger Licht.

MATRIX 2	***Entfernungs-Methode 1**: Stoppen beim Entfernen der Lampe*	***Annäherungs-Methode 2**: Starten beim Annähern der Lampe*	***Geschwindigkeits-Methode 3**: Geschwindigkeitsvergleich bei konstanter Lampenentfernung*	***Beschattungs-Methode 4**: maximaler Abdeckungsgrad beim Beschatten*
Messergebnis aus Item 9	*Entfernung Schaltung A > Entfernung Schaltung B*	*Entfernung Schaltung A > Entfernung Schaltung B*	*Ventilator A schneller als Ventilator B*	*Max. Beschattung Schaltung A > max. Beschattung Schaltung B*
***Schluss**: Weniger Licht benötigt die*	***Schaltung A***	***Schaltung A***	***Schaltung A***	***Schaltung A***

⟨N9E84i10⟩

Bei diesem Item wird bewertet, ob der Schüler/die Schülerin aus seinen/ihren Daten, die auch falsch sein können, den richtigen Schluss zieht. Schaltung A und B können je nach Sch die Serie- bzw. die Parallelschaltung bedeuten.

Code 1: richtige Schlussfolgerung gemäss Matrix 2

Code 0: falsche Schlussfolgerung gemäss Matrix 2

ODER

keine Beurteilung möglich, da in Item 9 (oder 8) keine Messergebnisse festgehalten wurden

Code 9: keine Antwort

Solarzellen werden gelegentlich durch Gegenstände beschattet und liefern dadurch weniger Strom. Untersuche, welche Schaltung (Serie- oder Parallel-Schaltung) durch Beschattung weniger beeinträchtigt wird.

- Halte das Ergebnis deiner Untersuchung fest.

☐ *Die Serie-Schaltung wird durch Beschattung weniger beeinträchtigt*
☒ *Die Parallel-Schaltung wird durch Beschattung weniger beeinträchtigt*

⟨N9E84i11⟩

Code 1: richtige Antwort

Code 0: falsche Antwort

Code 9: keine Antwort

- Wie bist du auf deine Ergebnisse gekommen?
  Beschreibe, wie du beim Experimentieren vorgegangen bist und was du beobachtet hast.

Beschattung der Serieschaltung:
Der Ventilator stoppt, wenn eine Solarzelle ganz beschattet ist.

Beschattung der Parallelschaltung:
Der Ventilator stoppt, wenn eine Solarzelle ganz und ½ – ¾ der zweiten Solarzellen beschattet sind.

⟨N9E84i12⟩

Code 2: vollständige Antwort:
- a) quantitative Angabe des maximal möglichen Beschattungsgrades im Sinne von
  - ***Serieschaltung stoppt beim Abdecken von 1 Zelle***
  - ***Parallelschaltung stoppt erst beim Abdecken von mehr als 1 Zelle***
- b) Schluss gezogen: beste Schaltung = Schaltung mit grösstem erlaubtem Beschattungsgrad (Dieses Kriterium ist erfüllt, wenn im Item 11 die korrekte Lösung angekreuzt wird!)

Code 1: unvollständige Antwort: ein Punkt von a) und b) falsch, unvollständig oder nicht erwähnt

Code 0: falsche Antwort: kein Punkt von a) und b) vollständig beantwortet

Code 9: keine Antwort

Anhang B

Deutschsprachige Itemstichproben: ⟨E08|69d|V⟩, ⟨E08|69d|red⟩, ⟨E08|69d|T⟩

Die drei Analysen der Itemschwierigkeit beziehen sich auf drei verschiedene Itemstichproben:

⟨E08|69d|V⟩ *Vollständige Itemstichprobe*: Diese Stichprobe umfasst alle 96 Items mit experimentellem Inhalt. Gegenüber der Validierung des HarmoS-Kompetenzmodells (Ramseier et al., 2011) haben wir den experimentellen Kompetenzbergriff enger gefasst und mehr Items als nicht experimentell ausgeschieden.

⟨E08|69d|red⟩ *Reduzierte Itemstichprobe*: Diese Stichprobe ist eine Teilmenge der vollständigen Stichprobe, bei der das Experimentiermaterial für die Lösung der Items eine Rolle spielt.

⟨E08|69d|T⟩ *Testlet-Stichprobe*: Zwischen den Items der vollständigen Stichprobe bestehen Abhängigkeiten, weil die Teilprozesse der Items aufeinander aufbauen. Bei der Testlet-Itemstichprobe werden abhängige Items zu einem Summenitem zusammengefasst und neu kodiert. Die Regeln der Umkodierung sind im Anhang D zusammengestellt.

Die Stichproben sind in der nachfolgenden Tabelle B.1 als Teilmengen der 141 Items der Basisstichprobe (cf. S. 109) zusammengestellt. Zur besseren Orientierung sind zusätzlich zu jedem Item der Aufgabentyp (cf. Tab. 7.5 auf S. 139) und die involvierten Teilprozesse (cf. Tab. 7.4 auf S. 138) angegeben.

E-Aufgaben	vollständige Stichprobe	Aufgabentyp	Aufgabenumfang	reduzierte Stichprobe	Testlet-Stichprobe
⟨**Steine**⟩	⟨N1E13i01⟩	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx
(69df)	⟨N1E13i02⟩	Untersuchung	Hypothese	xxxxxxxxxx	⟨N1E13i02T⟩
	⟨N1E13i03⟩	Untersuchung	Planung	manipulativ	⟨N1E13i03T⟩
	⟨N1E13i04⟩	Untersuchung	Planung, Durchführung	manipulativ	⟨N1E13i04T⟩
	⟨N1E13i05⟩		Durchführung, Auswertung	manipulativ	
	⟨N1E13i06⟩		Reflexion	xxxxxxxxxx	
	⟨N1E13i07⟩	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx
	⟨N1E13i08⟩		Auswertung	xxxxxxxxxx	
⟨**Rapex**⟩	⟨N9E14i01⟩	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx
(9d)	⟨N9E14i02⟩	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx
	⟨N9E14i03⟩	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx
	⟨N9E14i04⟩	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx
	⟨N9E14i05⟩	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx
	⟨N9E14i07⟩	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx
	⟨N9E14i08⟩	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx
	⟨N9E14i09⟩	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx
	⟨N9E14i10⟩	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx
	⟨N9E14i11⟩	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx
	⟨N9E14i12⟩	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx
	⟨N9E14i13⟩	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx
⟨**Murmel**⟩	⟨N9E21i01⟩	Messung	Planung, Durchführung	manipulativ	⟨N9E21i01T⟩
(9df)	⟨N9E21i02⟩	Messung	Durchführung	manipulativ	⟨N9E21i02T⟩
	⟨N9E21i03⟩	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx
	⟨N9E21i04⟩	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx
	⟨N9E21i05⟩	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx
	⟨N9E21i06⟩	Untersuchung	Planung, Durchführung	manipulativ	⟨N9E21i06T⟩
	⟨N9E21i07⟩		Auswertung	xxxxxxxxxx	
	⟨N9E21i08⟩		Auswertung	xxxxxxxxxx	
	⟨N9E21i09⟩	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx
⟨**Taschenlampe**⟩	⟨N1E22i01⟩	Vergleich	Planung	manipulativ	⟨N1E22i01T⟩
(69df)	⟨N1E22i02⟩		Duchführung	manipulativ	
	⟨N1E22i03⟩		Auswertung	manipulativ	
	⟨N1E22i04⟩		Reflexion	manipulativ	
	⟨N1E22i05⟩	Messung	Planung, Durchführung	manipulativ	⟨N1E22i05T⟩
	⟨N1E22i06⟩	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx
	⟨N1E22i07⟩	Beobachtung	Planung, Durchführung	manipulativ	⟨N1E22i07T⟩
⟨**Balkenwaage**⟩	⟨N1E23i01⟩	Herstellung	Planung, Durchführung	manipulativ	⟨N1E23i01T⟩
(69dfi)	⟨N1E23i02⟩		Planung, Durchführung	manipulativ	
	⟨N1E23i03⟩		Auswertung	xxxxxxxxxx	
	⟨N1E23i04⟩	Herstellung	Planung, Durchführung	manipulativ	⟨N1E23i04T⟩
	⟨N1E23i05⟩		Planung, Durchführung	manipulativ	
	⟨N1E23i06⟩		Auswertung	xxxxxxxxxx	
	⟨N1E23i07⟩	Herstellung	Planung, Durchführung	manipulativ	⟨N1E23i07T⟩
	⟨N1E23i08⟩		Planung, Durchführung	manipulativ	
	⟨N1E23i09⟩		Auswertung	xxxxxxxxxx	
	⟨N1E23i10⟩	Herstellung	Planung, Durchführung	manipulativ	⟨N1E23i10T⟩
	⟨N1E23i11⟩		Planung, Durchführung	manipulativ	
	⟨N1E23i12⟩		Auswertung	xxxxxxxxxx	
	⟨N1E23i13⟩	Herstellung	Planung, Durchführung	manipulativ	⟨N1E23i13T⟩
	⟨N1E23i14⟩		Planung, Durchführung	manipulativ	
	⟨N1E23i15⟩		Auswertung	xxxxxxxxxx	
	⟨N1E23i16⟩	Herstellung	Planung, Durchführung	manipulativ	⟨N1E23i16T⟩
	⟨N1E23i17⟩		Planung, Durchführung	manipualtiv	
	⟨N1E23i18⟩		Auswertung	xxxxxxxxxx	
	⟨N1E23i19⟩	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx
	⟨N1E23i20⟩	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx

E-Aufgaben	**vollständige Stichprobe**	**Aufgabentyp**	**Aufgabenumfang**	**reduzierte Stichprobe**	**Testlet-Stichprobe**
⟨**Seife**⟩	⟨N9E41i01⟩	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx
(9dfi)	⟨N9E41i02⟩	Messung	Planung, Durchführung	manipulativ	⟨N9E41i02T⟩
	⟨N9E41i03⟩		Auswertung	xxxxxxxxxx	
	⟨N9E41i04⟩	Messung	Durchführung	manipulativ	⟨N9E41i04T⟩
	⟨N9E41i05⟩		Auswertung	xxxxxxxxxx	
	⟨N9E41i06⟩	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx
	⟨N9E41i07⟩	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx
	⟨N9E41i08⟩	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx
	⟨N9E41i09⟩	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx
⟨**Öl**⟩	⟨N9E42i01⟩	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx
(9d)	⟨N9E42i02⟩	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx
	⟨N9E42i03⟩	Beobachtung	Durchführung	manipulativ	⟨N9E42i03T⟩
	⟨N9E42i04⟩	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx
	⟨N9E42i05⟩	Messung	Planung	manipulativ	⟨N9E42i05T⟩
	⟨N9E42i06⟩	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx
	⟨N9E42i07⟩	Messung	Durchführung	manipulativ	⟨N9E42i07T⟩
	⟨N9E42i08⟩	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx
	⟨N9E42i09⟩	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx
	⟨N9E42i10⟩	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx
⟨**Tabletten**⟩	⟨N1E43i01⟩	Untersuchung	Hypothese	xxxxxxxxxx	⟨N1E43i01T⟩
(69df)	⟨N6E43i02⟩	Untersuchung	Planung	manipulativ	⟨N6E43i02T⟩
	⟨N6E43i03⟩		Durchführung	manipulativ	
	⟨N9E43i02⟩	Untersuchung	Planung	manipulativ	⟨N9E43i02T⟩
	⟨N9E43i03⟩		Durchführung	manipulativ	
	⟨N1E43i04⟩		Auswertung	xxxxxxxxxx	
	⟨N1E43i05⟩		Auswertung	xxxxxxxxxx	
	⟨N1E43i06⟩	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx
	⟨N1E43i07⟩		Reflexion	xxxxxxxxxx	
⟨**Schwimmen & Sinken**⟩	⟨N6E44i01⟩	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx
(6d)	⟨N6E44i02⟩	Untersuchung	Planung	manipulativ	⟨N6E44i02T⟩
	⟨N6E44i03⟩		Durchführung	manipulativ	
	⟨N6E44i04⟩	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx
	⟨N6E44i05⟩	Messung	Durchführung	manipulativ	⟨N6E44i05T⟩
	⟨N6E44i06⟩	Messung	Planung	manipulativ	⟨N6E44i06T⟩
	⟨N6E44i07⟩		Durchführung, Auswertung	manipulativ	
	⟨N6E44i08⟩		Reflexion	xxxxxxxxxx	
	⟨N6E44i09⟩	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx
⟨**Asseln**⟩	⟨N1E53i01⟩	Vergleich	Durchführung	manipulativ	⟨N1E53i01T⟩
(69dfi)	⟨N1E53i02⟩		Auswertung	xxxxxxxxxx	
	⟨N1E53i03⟩		Reflexion	xxxxxxxxxx	
	⟨N1E53i04⟩	Vergleich	Hypothese	xxxxxxxxxx	⟨N1E53i04T⟩
	⟨N1E53i05⟩		Planung	manipualtiv	
	⟨N1E53i06⟩		Durchführung	manipulativ	
	⟨N1E53i08⟩		Reflexion	xxxxxxxxxx	
	⟨N1E53i07⟩	Beobachtung	Durchführung	manipulativ	⟨N1E53i07T⟩
	⟨N1E53i09⟩	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx
⟨**Wasserpest**⟩	⟨N9E54i01⟩	Messung	Durchführung	manipulativ	⟨N9E54i01T⟩
(9d)	⟨N9E54i02⟩	Beobachtung	Durchführung	manipulativ	⟨N9E54i02T⟩
	⟨N9E54i03⟩	(Messung)	Durchführung	manipulativ	
	⟨N9E54i04⟩		Auswertung	xxxxxxxxxx	
	⟨N9E54i05⟩	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx

E-Aufgaben	**vollständige Stichprobe**	**Aufgabentyp**	**Aufgabenumfang**	**reduzierte Stichprobe**	**Testlet-Stichprobe**
⟨**Gänseblümchen**⟩	⟨N6E56i01⟩	Beobachtung	Durchführung	manipulativ	⟨N6E56i01T⟩
(6df)	⟨N6E56i02⟩	Beobachtung	Durchführung	manipulativ	⟨N6E56i02T⟩
	⟨N6E56i03⟩	Beobachtung	Durchführung	manipulativ	⟨N6E56i03T⟩
	⟨N6E56i04⟩	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx
	⟨N6E56i05⟩	Beobachtung	Durchführung, Auswertung	manipulativ	⟨N6E56i05T⟩
	⟨N6E56i06⟩		Reflexion	xxxxxxxxxx	
⟨**Laubbäume**⟩	⟨N6E61i01⟩	Beobachtung	Durchführung	manipulativ	⟨N6E61i01T⟩
(6d)	⟨N6E61i02⟩	Beobachtung	Durchführung	manipulativ	⟨N6E61i02T⟩
	⟨N6E61i03⟩	Beobachtung	Durchführung	manipulativ	⟨N6E61i03T⟩
	⟨N6E61i04⟩	Beobachtung	Durchführung	manipulativ	⟨N6E61i04T⟩
	⟨N6E61i05⟩	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx
	⟨N6E61i06⟩	Beobachtung	Durchführung, Auswertung	manipulativ	⟨N6E61i06T⟩
	⟨N6E61i07⟩	Beobachtung	Durchführung	manipulativ	⟨N6E61i07T⟩
	⟨N6E61i08⟩	Beobachtung	Durchführung, Auswertung	manipulativ	⟨N6E61i08T⟩
	⟨N6E61i09⟩	Beobachtung	Durchführung, Auswertung	manipulativ	⟨N6E61i09T⟩
	⟨N6E61i10⟩	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx
⟨**Sparlampe**⟩	⟨N9E83i01⟩	Beobachtung	Durchführung	manipulativ	⟨N9E83i01T⟩
(9d)	⟨N9E83i02⟩	Herstellung	Durchführung	manipulativ	⟨N9E83i02T⟩
	⟨N9E83i03⟩	Herstellung	Durchführung	manipulativ	⟨N9E83i03T⟩
	⟨N9E83i04⟩		Auswertung	xxxxxxxxxx	
	⟨N9E83i05⟩	Herstellung	Durchführung	manipulativ	⟨N9E83i05T⟩
	⟨N9E83i06⟩	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx
	⟨N9E83i07⟩	Beobachtung	Durchführung, Auswertung	manipulativ	⟨N9E83i07T⟩
⟨**Solarzellen**⟩	⟨N9E84i02⟩	Herstellung	Durchführung	manipulativ	⟨N9E84i02T⟩
(9dfi)	⟨N9E84i03⟩		Durchführung	manipulativ	
	⟨N9E84i05⟩	Herstellung	Durchführung, Auswertung	manipulativ	⟨N9E84i05T⟩
	⟨N9E84i07⟩		Durchführung, Auswertung	manipulativ	
	⟨N9E84i08⟩	Untersuchung	Planung	manipulativ	⟨N9E84i08T⟩
	⟨N9E84i09⟩		Durchführung	manipulativ	
	⟨N9E84i10⟩		Auswertung	xxxxxxxxxx	
	⟨N9E84i11⟩	Untersuchung	Auswertung	xxxxxxxxxx	⟨N9E84i11T⟩
	⟨N9E84i12⟩		Planung, Durchführung	manipulativ	
	⟨N9E84i13⟩	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx
	⟨N9E84i14⟩	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx	xxxxxxxxxx

Tabelle B.1 – Übersicht über die Experimentieraufgaben und -items

Anhang C

Item-Test-Analysen: Vollständige Itemstichprobe ⟨E08|69d|V⟩

C.1 Itemparameter

Mit der vollständigen Itemstichprobe wurde eine separate Rasch-Modellierung gerechnet. Die nachfolgende Tabelle C.1 fasst die Ergebnisse zusammen, wobei die Grauunterlegung von Zellen kritische Itemwerte anzeigt.

E-Aufgaben	Item	Item-schwierigkeit	Trenn-schärfe	Infit-Parameter
⟨**Steine**⟩	⟨N1E13i02⟩	0.444	0.52	1.03
(69d)	⟨N1E13i03⟩	0.829	0.58	0.95
	⟨N1E13i04⟩	0.412	0.59	0.97
	⟨N1E13i05⟩	0.930	0.55	1.01
	⟨N1E13i06⟩	0.162	0.46	0.96
	⟨N1E13i08⟩	-0.299	0.52	0.95
⟨**Murmel**⟩	⟨N9E21i01⟩	0.007	0.47	1.25
(9d)	⟨N9E21i02⟩	0.218	0.30	1.14
	⟨N9E21i07⟩	1.278	0.62	0.82
	⟨N9E21i08⟩	0.463	0.64	0.95
⟨**Taschenlampe**⟩	⟨N1E22i01⟩	0.485	0.60	0.99
(69d)	⟨N1E22i02⟩	0.974	0.57	0.97
	⟨N1E22i03⟩	-0.897	0.27	0.98
	⟨N1E22i04⟩	0.761	0.53	1.00
	⟨N1E22i05⟩	-3.929	0.09	1.06
	⟨N1E22i07⟩	0.820	0.41	1.06
⟨**Balkenwaage**⟩	⟨N1E23i01⟩	-1.324	0.34	0.93
(69d)	⟨N1E23i02⟩	-0.445	0.46	0.95
	⟨N1E23i03⟩	-0.578	0.49	0.90
	⟨N1E23i04⟩	-0.196	0.41	1.05
	⟨N1E23i05⟩	0.024	0.30	1.13
	⟨N1E23i06⟩	0.205	0.56	0.88
	⟨N1E23i07⟩	0.130	0.62	0.87
	⟨N1E23i08⟩	0.345	0.54	1.03
	⟨N1E23i09⟩	0.166	0.61	0.89
	⟨N1E23i10⟩	0.142	0.61	0.92
	⟨N1E23i11⟩	0.265	0.53	1.03
	⟨N1E23i12⟩	0.253	0.65	0.85
	⟨N1E23i13⟩	-0.671	0.57	0.86
	⟨N1E23i14⟩	-0.692	0.42	0.99
	⟨N1E23i15⟩	-0.196	0.55	0.88
	⟨N1E23i16⟩	-1.142	0.43	0.82
	⟨N1E23i17⟩	-0.626	0.42	0.99
	⟨N1E23i18⟩	0.071	0.48	1.03
⟨**Seife**⟩	⟨N9E41i02⟩	1.202	0.41	1.13
(9d)	⟨N9E41i03⟩	2.794	0.42	0.98
	⟨N9E41i04⟩	1.286	0.32	1.12
	⟨N9E41i05⟩	0.210	0.44	0.95
⟨**Öl**⟩	⟨N9E42i03⟩	-0.786	0.43	1.02
(9d)	⟨N9E42i05⟩	1.276	0.37	0.99
	⟨N9E42i07⟩	-0.593	0.44	1.04
⟨**Tabletten**⟩	⟨N1E43i01⟩	-0.312	0.45	1.09
(69d)	⟨N6E43i02⟩	0.620	0.61	0.91
	⟨N9E43i02⟩	1.464	0.48	0.94
	⟨N6E43i03⟩	-0.399	0.60	0.86
	⟨N9E43i03⟩	-0.251	0.49	0.92
	⟨N1E43i04⟩	-0.835	0.39	1.07
	⟨N1E43i05⟩	-1.130	0.24	1.02
	⟨N1E43i07⟩	0.159	0.47	1.02

E-Aufgaben	Item	Item-schwierigkeit	Trenn-schärfe	Infit-Parameter
⟨**Schwimmen & Sinken**⟩	⟨N6E44i02⟩	-0.104	0.39	1.11
(6d)	⟨N6E44i03⟩	0.706	0.53	0.97
	⟨N6E44i05⟩	-0.352	0.63	0.81
	⟨N6E44i06⟩	1.327	0.41	0.96
	⟨N6E44i07⟩	-0.330	0.56	0.88
	⟨N6E44i08⟩	-0.946	0.45	0.94
⟨**Asseln**⟩	⟨N1E53i01⟩	-0.951	0.16	1.04
(69d)	⟨N1E53i02⟩	-2.313	0.11	1.01
	⟨N1E53i03⟩	-0.672	0.15	1.08
	⟨N1E53i04⟩	0.813	0.32	1.04
	⟨N1E53i05⟩	0.084	0.44	1.08
	⟨N1E53i06⟩	-0.325	0.18	1.04
	⟨N1E53i07⟩	-0.887	0.23	1.07
	⟨N1E53i08⟩	0.119	0.51	1.02
⟨**Wasserpest**⟩	⟨N9E54i01⟩	2.074	0.49	1.03
(9d)	⟨N9E54i02⟩	1.099	0.72	0.88
	⟨N9E54i03⟩	1.971	0.29	0.98
	⟨N9E54i04⟩	-0.806	0.70	1.00
⟨**Gänseblümchen**⟩	⟨N6E56i01⟩	-1.113	0.06	1.12
(6d)	⟨N6E56i02⟩	2.657	0.18	0.98
	⟨N6E56i03⟩	-0.548	0.41	1.09
	⟨N6E56i05⟩	-1.443	0.32	1.03
	⟨N6E56i06⟩	-0.116	0.47	0.99
⟨**Laubbäume**⟩	⟨N6E61i01⟩	0.204	0.36	1.14
(6d)	⟨N6E61i02⟩	0.707	0.19	1.20
	⟨N6E61i03⟩	-1.420	0.37	0.98
	⟨N6E61i04⟩	0.053	0.35	1.19
	⟨N6E61i06⟩	-0.777	0.24	0.99
	⟨N6E61i07⟩	-0.738	0.10	1.08
	⟨N6E61i08⟩	-0.801	-0.03	1.12
	⟨N6E61i09⟩	0.115	0.22	1.04
⟨**Sparlampe**⟩	⟨N9E83i01⟩	-0.773	0.53	0.87
(9d)	⟨N9E83i02⟩	0.385	0.43	0.94
	⟨N9E83i03⟩	3.152	0.08	1.02
	⟨N9E83i04⟩	1.472	0.52	0.97
	⟨N9E83i05⟩	1.294	0.51	1.09
	⟨N9E83i07⟩	2.140	0.28	1.05
⟨**Solarzellen**⟩	⟨N9E84i02⟩	-2.222	0.28	0.99
(9d)	⟨N9E84i03⟩	-1.960	0.29	0.91
	⟨N9E84i05⟩	-2.216	0.25	1.07
	⟨N9E84i07⟩	-2.493	0.33	1.07
	⟨N9E84i08⟩	0.292	0.65	0.78
	⟨N9E84i09⟩	0.856	0.66	0.92
	⟨N9E84i10⟩	-0.003	0.51	0.92
	⟨N9E84i11⟩	-0.543	0.49	1.01
	⟨N9E84i12⟩	*0.195	0.57	1.13

Tabelle C.1 – Item-Test-Analysen: Übersicht über die Parameter der vollständigen Itemstichprobe

Anhang D

Item-Test-Analysen: Testlet-Stichprobe ⟨E08|69d|T⟩

D.1 Kodierung

Die Kodierung der Testlet-Items basiert auf den 96 Items der vollständigen Stichprobe ⟨E08|69df|V⟩. Aus den Grunditem, die ein Testlet-Item bilden, wurde das Summenitem berechnet. Dabei ergaben sich bis zu 11 Codes. Bei der Umkodierung wurden diese Summencodes auf maximal vier Testlet-Codes reduziert (Code 0, Code 1, Code 2, Code 3). Die jeweilige Umkodierung wird in den vier rechten Spalten der nachfolgenden Tabelle D.1 beschrieben.

E-Aufgaben	Items	Codes	Testlet-Items	Testlet-Codes			
				0	**1**	**2**	**3**
⟨**Steine**⟩	⟨N1E13i02⟩	0,1,2	⟨N1E13i02T⟩	0	1	2	
(69d)	⟨N1E13i03⟩	0,1,2	⟨N1E13i03T⟩	0	1	2	
	⟨N1E13i04⟩	0,1,2,3	⟨N1E13i04T⟩	0,1	2,3,4	5,6,7	8,9,10
	⟨N1E13i05⟩	0,1,2,3					
	⟨N1E13i06⟩	0,1,2					
	⟨N1E13i08⟩	0,1,2					
⟨**Murmel**⟩	⟨N9E21i01⟩	0,1,2	⟨N9E21i01T⟩	0	1	2	
(9d)	⟨N9E21i02⟩	0,1,2	⟨N9E21i02T⟩	0	1	2	
	⟨N9E21i06⟩	0,1,2	⟨N9E21i06T⟩	0	1,2,	3,4,	5,6,7
	⟨N9E21i07⟩	0,1,2					
	⟨N9E21i08⟩	0,1,2					
⟨**Taschenlampe**⟩	⟨N1E22i01⟩	0,1,2	⟨N1E22i01T⟩	0,1	2,3	4,5	6,7
(69d)	⟨N1E22i02⟩	0,1,2					
	⟨N1E22i03⟩	0,1					
	⟨N1E22i04⟩	0,1,2					
	⟨N1E22i05⟩	0,1	⟨N1E22i05T⟩				
	⟨N1E22i07⟩	0,1,2	⟨N1E22i07T⟩	0	1	2	
⟨**Balkenwaage**⟩	⟨N1E23i01⟩	0,1,2	⟨N1E23i01T⟩	0	1,2	3,4	5,6
(69d)	⟨N1E23i02⟩	0,1,2					
	⟨N1E23i03⟩	0,1,2					
	⟨N1E23i04⟩	0,1,2	⟨N1E23i04T⟩	0	1,2	3,4	5,6
	⟨N1E23i05⟩	0,1,2					
	⟨N1E23i06⟩	0,1,2					
	⟨N1E23i07⟩	0,1,2	⟨N1E23i07T⟩	0	1,2	3,4	5,6
	⟨N1E23i08⟩	0,1,2					
	⟨N1E23i09⟩	0,1,2					
	⟨N1E23i10⟩	0,1,2	⟨N1E23i10T⟩	0	1,2	3,4	5,6
	⟨N1E23i11⟩	0,1,2					
	⟨N1E23i12⟩	0,1,2					
	⟨N1E23i13⟩	0,1,2	⟨N1E23i13T⟩	0	1,2	3,4	5,6
	⟨N1E23i14⟩	0,1,2					
	⟨N1E23i15⟩	0,1,2					
	⟨N1E23i16⟩	0,1,2	⟨N1E23i16T⟩	0	1,2	3,4	5,6
	⟨N1E23i17⟩	0,1,2					
	⟨N1E23i18⟩	0,1,2					
⟨**Seife**⟩	⟨N9E41i02⟩	0,1,2	⟨N9E41i02T⟩	0	1	2	3,4
(9d)	⟨N9E41i03⟩	0,1,2					
	⟨N9E41i04⟩	0,1,2	⟨N9E41i04T⟩	0	1	2	3
	⟨N9E41i05⟩	0,1					
⟨**Öl**⟩	⟨N9E42i03⟩	0,1,2	⟨N9E42i03T⟩	0	1	2	
(9d)	⟨N9E42i05⟩	0,1	⟨N9E42i05T⟩	0	1		
	⟨N9E42i07⟩	0,1	⟨N9E42i07T⟩	0	1		
⟨**Tabletten**⟩	⟨N1E43i01⟩	0,1,2	⟨N1E43i01T⟩	0	1	2	
(69d)	⟨N6E43i02⟩	0,1,2	⟨N6E43i02T⟩	0,1	2,3,4	5,6,7	8,9
	⟨N6E43i03⟩	0,1,2					
	⟨N1E43i04⟩	0,1,2					
	⟨N1E43i05⟩	0,1					
	⟨N1E43i07⟩	0,1,2					
	⟨N9E43i02⟩	0,1,2	⟨N6E43i02T⟩	0,1	2,3,4	5,6,7	8,9
	⟨N9E43i03⟩	0,1,2					
	⟨N1E43i04⟩	0,1,2					
	⟨N1E43i05⟩	0,1					
	⟨N1E43i07⟩	0,1,2					

E-Aufgaben	Items	Codes	Testlet-Items	Testlet-Codes			
				0	**1**	**2**	**3**
⟨**Schwimmen & Sinken**⟩	⟨N6E44i02⟩	0,1,2	⟨N6E44i02T⟩	0	1	2	3,4
(6d)	⟨N6E44i03⟩	0,1,2					
	⟨N6E44i05⟩	0,1,2	⟨N6E44i05T⟩	0	1	2	
	⟨N6E44i06⟩	0,1,2	⟨N6E44i06T⟩	0	1	2	3,4
	⟨N6E44i07⟩	0,1					
	⟨N6E44i08⟩	0,1					
⟨**Asseln**⟩	⟨N1E53i01⟩	0,1	⟨N1E53i01T⟩	0	1	2	3
(69d)	⟨N1E53i02⟩	0,1					
	⟨N1E53i03⟩	0,1					
	⟨N1E53i04⟩	0,1	⟨N1E53i04T⟩	0	1,2	3,4	5,6
	⟨N1E53i05⟩	0,1,2					
	⟨N1E53i06⟩	0,1					
	⟨N1E53i08⟩	0,1,2					
	⟨N1E53i07⟩	0,1	⟨N1E53i07T⟩	0	1		
⟨**Wasserpest**⟩	⟨N9E54i01⟩	0,1,2	⟨N9E54i01T⟩	0	1	2	
(9d)	⟨N9E54i02⟩	0,1,2	⟨N9E54i02T⟩	0	1,2	3,4	5
	⟨N9E54i03⟩	0,1					
	⟨N9E54i04⟩	0,1,2					
⟨**Gänseblümchen**⟩	⟨N6E56i01⟩	0,1	⟨N6E56i01T⟩	0	1		
(6d)	⟨N6E56i02⟩	0,1	⟨N6E56i02T⟩	0	1		
	⟨N6E56i03⟩	0,1,2	⟨N6E56i03T⟩	0	1	2	
	⟨N6E56i05⟩	0,1	⟨N6E56i05T⟩	0	1	2	3
	⟨N6E56i06⟩	0,1,2					
⟨**Laubbäume**⟩	⟨N6E61i01⟩	0,1,2	⟨N6E61i01T⟩	0	1	2	
(6d)	⟨N6E61i02⟩	0,1,2	⟨N6E61i02T⟩	0	1	2	
	⟨N6E61i03⟩	0,1,2	⟨N6E61i03T⟩	0	1	2	
	⟨N6E61i04⟩	0,1,2	⟨N6E61i04T⟩	0	1	2	
	⟨N6E61i06⟩	0,1	⟨N6E61i06T⟩	0	1		
	⟨N6E61i07⟩	0,1	⟨N6E61i07T⟩	0	1		
	⟨N6E61i08⟩	0,1	⟨N6E61i08T⟩	0	1		
	⟨N6E61i09⟩	0,1	⟨N6E61i09T⟩	0	1		
⟨**Sparlampe**⟩	⟨N9E83i01⟩	0,1	⟨N9E83i01T⟩	0	1		
(9d)	⟨N9E83i02⟩	0,1	⟨N9E83i02T⟩	0	1		
	⟨N9E83i03⟩	0,1,2	⟨N9E83i03T⟩	0	1	2	3
	⟨N9E83i04⟩	0,1					
	⟨N9E83i05⟩	0,1,2	⟨N9E83i05T⟩	0	1	2	
	⟨N9E83i07⟩	0,1	⟨N9E83i07T⟩	0	1		
⟨**Solarzellen**⟩	⟨N9E84i02⟩	0,1	⟨N9E84i02T⟩	0	1	2	
(9d)	⟨N9E84i03⟩	0,1					
	⟨N9E84i05⟩	0,1	⟨N9E84i05T⟩	0	1	2	
	⟨N9E84i07⟩	0,1					
	⟨N9E84i08⟩	0,1,2	⟨N9E84i08T⟩	0	1	2	3,4
	⟨N9E84i09⟩	0,1					
	⟨N9E84i10⟩	0,1					
	⟨N9E84i11⟩	0,1	⟨N9E84i11T⟩	0	1	2	3
	⟨N9E84i12⟩	0,1,2					

Tabelle D.1 – Übersicht über die Testlet-Items und deren Codes

D.2 Itemparameter

Mit den umkodierten Testlet-Items wurde eine separate Rasch-Modellierung gerechnet. Die nachfolgende Tabelle D.2 fasst die Ergebnisse zusammen, wobei die Grauunterlegung von Zellen kritische Itemwerte anzeigt.

E-Aufgaben	Item	Item-schwierigkeit	Trenn-schärfe	Infit-Parameter
⟨**Steine**⟩	⟨N1E13i02T⟩	0.426	0.61	0.98
(69d)	⟨N1E13i03T⟩	0.826	0.62	0.99
	⟨N1E13i04T⟩	-0.151	0.66	0.87
⟨**Murmel**⟩	⟨N9E21i01T⟩	0.024	0.49	1.04
(9d)	⟨N9E21i02T⟩	0.251	0.26	1.07
	⟨N9E21i06T⟩	0.658	0.48	1.00
⟨**Taschenlampe**⟩	⟨N1E22i01T⟩	0.485	0.66	1.01
(69d)	⟨N1E22i05T⟩	-3.267	0.15	1.06
	⟨N1E22i07T⟩	0.781	0.50	0.96
⟨**Balkenwaage**⟩	⟨N1E23i01T⟩	-0.721	0.61	0.95
(69d)	⟨N1E23i04T⟩	-0.155	0.44	1.06
	⟨N1E23i07T⟩	0.261	0.65	1.07
	⟨N1E23i10T⟩	0.277	0.62	1.08
	⟨N1E23i13T⟩	-0.519	0.66	0.91
	⟨N1E23i16T⟩	-0.605	0.69	0.88
⟨**Seife**⟩	⟨N9E41i02T⟩	1.161	0.62	0.93
(9d)	⟨N9E41i04T⟩	1.017	0.48	1.17
⟨**Öl**⟩	⟨N9E42i03T⟩	-0.836	0.38	0.97
(9d)	⟨N9E42i05T⟩	1.383	0.37	1.06
	⟨N9E42i07T⟩	-0.277	0.43	1.00
⟨**Tabletten**⟩	⟨N1E43i01T⟩	-0.293	0.64	1.10
(69d)	⟨N6E43i02T⟩	-0.405	0.74	0.97
	⟨N9E43i02T⟩	-0.176	0.43	1.14
⟨**Schwimmen & Sinken**⟩	⟨N6E44i02T⟩	-0.115	0.60	1.06
(6d)	⟨N6E44i05T⟩	-0.399	0.63	0.81
	⟨N6E44i06T⟩	0.001	0.68	0.88
⟨**Asseln**⟩	⟨N1E53i01T⟩	-1.026	0.40	1.12
(69d)	⟨N1E53i04T⟩	-0.403	0.61	1.00
	⟨N1E53i07T⟩	-0.969	0.30	1.01
⟨**Wasserpest**⟩	⟨N9E54i01T⟩	2.002	0.69	1.10
(9d)	⟨N9E54i02T⟩	1.079	0.86	1.05
⟨**Gänseblümchen**⟩	⟨N6E56i01T⟩	-1.123	0.12	1.07
(6d)	⟨N6E56i02T⟩	2.617	0.17	1.07
	⟨N6E56i03T⟩	-0.556	0.48	1.07
	⟨N6E56i05T⟩	-0.529	0.58	0.92
⟨**Laubbäume**⟩	⟨N6E61i01T⟩	0.256	0.39	1.20
(6d)	⟨N6E61i02T⟩	0.683	0.22	1.13
	⟨N6E61i03T⟩	-1.443	0.40	0.91
	⟨N6E61i04T⟩	0.028	0.41	1.07
	⟨N6E61i06T⟩	-0.720	0.29	0.89
	⟨N6E61i07T⟩	-0.780	0.15	1.03
	⟨N6E61i08T⟩	-0.842	0.04	1.06
	⟨N6E61i09T⟩	0.135	0.24	1.06
⟨**Sparlampe**⟩	⟨N9E83i01T⟩	-0.822	0.51	0.92
(9d)	⟨N9E83i02T⟩	0.376	0.43	0.98
	⟨N9E83i03T⟩	1.537	0.44	1.02
	⟨N9E83i05T⟩	1.203	0.50	1.09
	⟨N9E83i07T⟩	2.200	0.27	1.04
⟨**Solarzellen**⟩	⟨N9E84i02T⟩	-1.440	0.40	0.95
(9d)	⟨N9E84i05T⟩	-1.463	0.39	1.01
	⟨N9E84i08T⟩	0.152	0.73	1.01
	⟨N9E84i11T⟩	*0.218	0.60	1.20

Tabelle D.2 – Item-Test-Analysen: Übersicht über die Parameter der Testlet-Items

Anhang E

Personen-Test-Analysen: Zweisprachige Itemstichprobe ⟨E08|69df⟩

E.1 Itemselektion

Für den Vergleich von Personenstichproben wurde eine Itemselektion vorgenommen, wobei in Korrespondenz zur Auswertung des HarmoS-Papier-und-Bleistifttests gemäss Ramseier et al. (2011) folgende Auswahlkriterien verwandt wurden: Items mit Trennschärfen < 0.3 wurden gestrichen, sofern sie weder besonders einfach (Itemschwierigkeit < -1.0) oder besonders schwierig (Itemschwietrigkeit > 1.0) waren. Ein Item wurde weggelassen, wenn der Infit-Parameter zu stark von 1.00 abwich ($|\text{wMNSQ}-1| > 0.30$). Items, die in der französisch- und deutschsprachigen Schweiz eingesetzt wurden, wurden in zwei Items gesplittet, wenn der DIF-Parameter > 0.4 ist.

E-Aufgaben	Item	Trenn-schärfe	Item-schwierigkeit	Infit-Parameter	0.5×DIF-Parameter	Massnahme
⟨**Steine**⟩	⟨N1E13i02⟩	0.41	0.443	1.09	-0.042	
(69df)	⟨N1E13i03⟩	0.51	1.375	0.92	-0.417	D/F-Split
	⟨N1E13i04⟩	0.53	0.604	0.96	-0.220	D/F-Split
	⟨N1E13i05⟩	0.51	1.010	0.99	-0.158	
	⟨N1E13i06⟩	0.38	0.073	1.02	-0.031	
	⟨N1E13i08⟩	0.41	0.240	1.03	-0.419	D/F-Split
⟨**Murmel**⟩	⟨N9E21i01⟩	0.49	0.191	1.16	-0.204	D/F-Split
(9df)	⟨N9E21i02⟩	0.44	0.509	0.96	-0.246	D/F-Split
	⟨N9E21i06⟩	0.29	-4.294	0.97	1.335	streichen
	⟨N9E21i07⟩	0.69	0.679	0.92	0.281	D/F-Split
	⟨N9E21i08⟩	0.63	0.750	0.90	-0.282	D/F-Split
⟨**Taschenlampe**⟩	⟨N1E22i01⟩	0.56	0.170	0.99	0.290	D/F-Split
(69df)	⟨N1E22i02⟩	0.62	0.656	1.01	0.287	D/F-Split
	⟨N1E22i03⟩	0.26	-1.006	1.02	0.003	
	⟨N1E22i04⟩	0.48	0.948	0.98	-0.288	D/F-Split
	⟨N1E22i05⟩	0.08	-4.129	0.99	-0.102	streichen
	⟨N1E22i07⟩	0.31	1.188	1.08	-0.297	D/F-Split
⟨**Balkenwaage**⟩	⟨N1E23i01⟩	0.28	-1.219	1.07	-0.243	D/F-Split
(69df)	⟨N1E23i02⟩	0.41	-0.461	1.03	-0.078	
	⟨N1E23i03⟩	0.43	-0.648	1.00	0.013	
	⟨N1E23i04⟩	0.32	0.007	1.07	-0.314	D/F-Split
	⟨N1E23i05⟩	0.32	0.129	1.08	-0.200	
	⟨N1E23i06⟩	0.48	0.585	0.99	-0.429	D/F-Split
	⟨N1E23i07⟩	0.62	0.044	0.85	0.091	
	⟨N1E23i08⟩	0.58	0.221	0.90	0.130	
	⟨N1E23i09⟩	0.64	0.065	0.83	0.098	
	⟨N1E23i10⟩	0.69	0.124	0.79	-0.016	
	⟨N1E23i11⟩	0.62	0.239	0.88	-0.012	
	⟨N1E23i12⟩	0.68	0.233	0.84	-0.005	
	⟨N1E23i13⟩	0.43	-0.798	0.94	0.138	
	⟨N1E23i14⟩	0.39	-0.743	1.04	-0.019	
	⟨N1E23i15⟩	0.45	-0.010	1.00	-0.270	D/F-Split
	⟨N1E23i16⟩	0.33	-1.275	0.91	0.260	streichen
	⟨N1E23i17⟩	0.38	-0.750	0.99	0.160	
	⟨N1E23i18⟩	0.42	-0.049	1.04	-0.019	
⟨**Seife**⟩	⟨N9E41i02⟩	0.37	1.573	1.04	-0.307	D/F-Split
(9df)	⟨N9E41i03⟩	0.35	2.164	1.03	0.273	D/F-Split
	⟨N9E41i04⟩	0.33	1.333	1.06	-0.147	
	⟨N9E41i05⟩	0.40	0.777	0.99	-0.912	D/F-Split
⟨**Öl**⟩	⟨N9E42i03⟩	0.43	-0.657	1.00	xxxxx	
(9d)	⟨N9E42i05⟩	0.37	1.394	0.99	xxxxx	
	⟨N9E42i07⟩	0.44	-0.467	1.03	xxxxx	
⟨**Tabletten**⟩	⟨N1E43i01⟩	0.44	-0.169	1.03	-0.194	
(69df)	⟨N6E43i02⟩	0.50	0.512	0.92	0.006	
	⟨N9E43i02⟩	0.94	1.260	0.94	-0.005	
	⟨N6E43i03⟩	0.48	-0.202	1.03	-0.200	
	⟨N9E43i03⟩	0.93	-0.419	0.93	0.088	
	⟨N1E43i04⟩	0.36	-0.677	1.09	-0.147	
	⟨N1E43i05⟩	0.23	-0.677	1.02	-0.451	streichen
	⟨N1E43i07⟩	0.45	0.305	0.93	-0.190	

E-Aufgaben	Item	Trenn-schärfe	Item-schwierigkeit	Infit-Parameter	0.5×DIF-Parameter	Massnahme
⟨**Schwimmen & Sinken**⟩	⟨N6E44i02⟩	0.39	0.005	1.08	xxxxx	
(6d)	⟨N6E44i03⟩	0.81	0.810	0.81	xxxxx	
	⟨N6E44i05⟩	0.63	-0.241	0.95	xxxxx	
	⟨N6E44i06⟩	0.41	1.428	1.05	xxxxx	
	⟨N6E44i07⟩	0.56	-0.220	0.93	xxxxx	
	⟨N6E44i08⟩	0.45	-0.833	0.97	xxxxx	
⟨**Asseln**⟩	⟨N1E53i01⟩	0.23	-0.262	1.03	-0.580	streichen
(69df)	⟨N1E53i02⟩	0.15	-2.134	1.00	-0.155	
	⟨N1E53i03⟩	0.23	-0.362	1.05	-0.276	streichen
	⟨N1E53i04⟩	0.24	1.736	1.03	-0.905	D/F-Split
	⟨N1E53i05⟩	0.34	0.370	1.03	-0.278	D/F-Split
	⟨N1E53i06⟩	0.22	-0.316	1.06	-0.017	streichen
	⟨N1E53i07⟩	0.22	-1.148	1.05	0.293	streichen
	⟨N1E53i08⟩	0.44	-0.085	1.00	0.078	
⟨**Wasserpest**⟩	⟨N9E54i01⟩	0.49	2.184	1.01	xxxxx	
(9d)	⟨N9E54i02⟩	0.72	1.214	0.92	xxxxx	
	⟨N9E54i03⟩	0.29	2.086	0.99	xxxxx	
	⟨N9E54i04⟩	0.70	-0.676	1.02	xxxxx	
⟨**Gänseblümchen**⟩	⟨N6E56i01⟩	0.16	-0.898	1.05	-0.285	streichen
(6df)	⟨N6E56i02⟩	0.07	2.212	1.07	0.329	streichen
	⟨N6E56i03⟩	0.39	-0.508	1.10	-0.043	
	⟨N6E56i05⟩	0.23	-1.398	1.05	-0.114	
	⟨N6E56i06⟩	0.46	0.193	1.05	-0.353	D/F-Split
⟨**Laubbäume**⟩	⟨N6E61i01⟩	0.36	0.274	1.13	xxxxx	
(6d)	⟨N6E61i02⟩	0.19	0.779	1.09	xxxxx	streichen
	⟨N6E61i03⟩	0.37	-1.349	1.11	xxxxx	
	⟨N6E61i04⟩	0.35	0.122	1.21	xxxxx	
	⟨N6E61i06⟩	0.24	-0.706	1.07	xxxxx	streichen
	⟨N6E61i07⟩	0.10	-0.688	1.08	xxxxx	streichen
	⟨N6E61i08⟩	-0.03	-0.731	1.14	xxxxx	streichen
	⟨N6E61i09⟩	0.22	0.225	1.01	xxxxx	streichen
⟨**Sparlampe**⟩	⟨N9E83i01⟩	0.53	-0.649	0.87	xxxxx	
(9d)	⟨N9E83i02⟩	0.43	0.504	0.95	xxxxx	
	⟨N9E83i03⟩	0.08	3.260	1.01	xxxxx	streichen
	⟨N9E83i04⟩	0.52	1.586	0.95	xxxxx	
	⟨N9E83i05⟩	0.51	1.405	1.08	xxxxx	
	⟨N9E83i07⟩	0.28	2.251	1.04	xxxxx	
⟨**Solarzellen**⟩	⟨N9E84i02⟩	0.22	-3.003	0.97	3.640	streichen
(9df)	⟨N9E84i03⟩	0.23	-2.547	0.97	0.845	streichen
	⟨N9E84i05⟩	0.28	-2.707	0.98	0.685	streichen
	⟨N9E84i07⟩	0.35	-2.472	0.98	-0.044	
	⟨N9E84i08⟩	0.67	0.413	0.88	-0.147	
	⟨N9E84i09⟩	0.64	0.592	0.92	0.205	D/F-Split
	⟨N9E84i10⟩	0.63	-0.232	0.85	0.201	D/F-Split
	⟨N9E84i11⟩	0.50	-0.828	0.93	0.230	D/F-Split
	⟨N9E84i12⟩	0.52	*0.180	0.97	*0.023	

Tabelle E.1 – Übersicht über die Parameter für die Itemselektion

E.2 Itemparameter

Mit der durch Selektion und Splitting korrigierten Itemstichprobe wurde eine zweite eindimensionale Rasch-Analyse gerechnet. Die nachfolgende Tabelle E.2 fasst die Ergebnisse zusammen, wobei die Grauunterlegung von Zellen kritische Itemwerte anzeigt.

E-Aufgaben	Item	Item-schwierigkeit	Trenn-schärfe	Infit-Parameter
⟨**Steine**⟩	⟨N1E13i02⟩	0.122	0.43	1.10
(69df)	⟨N1E13i03D⟩	0.611	0.58	0.95
	⟨N1E13i03F⟩	1.562	0.46	0.90
	⟨N1E13i04D⟩	0.189	0.64	0.90
	⟨N1E13i04F⟩	0.359	0.51	0.97
	⟨N1E13i05⟩	0.707	0.54	0.96
	⟨N1E13i06⟩	-0.258	0.42	1.02
	⟨N1E13i08D⟩	-0.548	0.53	1.02
	⟨N1E13i08F⟩	0.242	0.33	1.10
⟨**Murmel**⟩	⟨N9E21i01D⟩	-0.175	0.49	1.29
(9df)	⟨N9E21i01F⟩	-0.078	0.50	1.06
	⟨N9E21i02D⟩	0.023	0.29	1.25
	⟨N9E21i02F⟩	0.285	0.57	0.87
	⟨N9E21i07D⟩	1.091	0.61	0.89
	⟨N9E21i07F⟩	0.101	0.72	0.75
	⟨N9E21i08D⟩	0.442	0.73	0.84
	⟨N9E21i08F⟩	0.571	0.77	0.71
⟨**Taschenlampe**⟩	⟨N1E22i01D⟩	0.292	0.62	0.96
(69df)	⟨N1E22i01F⟩	-0.427	0.53	1.09
	⟨N1E22i02D⟩	0.791	0.58	1.12
	⟨N1E22i02F⟩	0.13o	0.67	0.92
	⟨N1E22i03⟩	-1.280	0.25	1.01
	⟨N1E22i04D⟩	0.570	0.52	0.91
	⟨N1E22i04F⟩	0.817	0.47	1.09
	⟨N1E22i07D⟩	0.630	0.41	1.04
	⟨N1E22i07F⟩	1.274	0.26	1.14
⟨**Balkenwaage**⟩	⟨N1E23i01D⟩	-1.637	0.34	1.04
(69df)	⟨N1E23i01F⟩	-1.519	0.21	1.15
	⟨N1E23i02⟩	-0.773	0.43	1.03
	⟨N1E23i03⟩	-0.999	0.44	0.97
	⟨N1E23i04D⟩	-0.498	0.31	1.08
	⟨N1E23i04F⟩	-0.023	0.33	1.05
	⟨N1E23i05⟩	-0.168	0.33	1.09
	⟨N1E23i06D⟩	0.285	0.53	0.95
	⟨N1E23i06F⟩	0.702	0.46	0.99
	⟨N1E23i07⟩	-0.261	0.62	0.86
	⟨N1E23i08⟩	-0.075	0.59	0.94
	⟨N1E23i09⟩	-0.246	0.65	0.85
	⟨N1E23i10⟩	-0.176	0.68	0.83
	⟨N1E23i11⟩	-0.054	0.61	0.90
	⟨N1E23i12⟩	-0.064	0.68	0.84
	⟨N1E23i13⟩	-1.117	0.42	1.00
	⟨N1E23i14⟩	-1.062	0.39	1.01
	⟨N1E23i15D⟩	-0.504	0.55	0.92
	⟨N1E23i15F⟩	-0.137	0.30	1.03
	⟨N1E23i17⟩	-1.068	0.34	1.06
	⟨N1E23i18⟩	-0.448	0.39	1.06
⟨**Seife**⟩	⟨N9E41i02D⟩	0.953	0.41	1.10
(9df)	⟨N9E41i02F⟩	2.055	0.25	0.97
	⟨N9E41i03D⟩	2.590	0.42	1.00
	⟨N9E41i03F⟩	1.382	0.36	0.99
	⟨N9E41i04⟩	1.044	0.33	1.15
	⟨N9E41i05D⟩	-0.050	0.44	0.95
	⟨N9E41i05F⟩	1.652	0.15	1.02

E-Aufgaben	Item	Item-schwierigkeit	Trenn-schärfe	Infit-Parameter
⟨**Öl**⟩	⟨N9E42i03⟩	-0.963	0.46	1.00
(9d)	⟨N9E42i05⟩	1.125	0.39	1.05
	⟨N9E42i07⟩	-0.763	0.43	1.05
⟨**Tabletten**⟩	⟨N1E43i01⟩	-0.442	0.43	1.10
(69df)	⟨N6E43i02⟩	0.258	0.53	0.95
	⟨N9E43i02⟩	1.021	0.42	1.09
	⟨N6E43i03⟩	-0.475	0.51	1.15
	⟨N9E43i03⟩	-0.684	0.57	0.91
	⟨N1E43i04⟩	-0.959	0.36	1.09
	⟨N1E43i07⟩	0.046	0.44	1.03
⟨**Schwimmen & Sinken**⟩	⟨N6E44i02⟩	-0.340	0.39	1.24
(6d)	⟨N6E44i03⟩	0.485	0.58	1.16
	⟨N6E44i05⟩	-0.598	0.65	0.98
	⟨N6E44i06⟩	1.123	0.44	1.03
	⟨N6E44i07⟩	-0.564	0.58	0.87
	⟨N6E44i08⟩	-1.181	0.43	0.97
⟨**Asseln**⟩	⟨N1E53i02⟩	-2.474	0.14	1.03
(69df)	⟨N1E53i04D⟩	0.596	0.37	1.01
	⟨N1E53i04F⟩	2.283	0.24	1.00
	⟨N1E53i05D⟩	-0.161	0.48	1.08
	⟨N1E53i05F⟩	-1.866	0.20	1.04
	⟨N1E53i08⟩	-0.399	0.44	1.08
⟨**Wasserpest**⟩	⟨N9E54i01⟩	1.937	0.49	0.99
(9d)	⟨N9E54i02⟩	0.944	0.72	0.89
	⟨N9E54i03⟩	1.813	0.29	0.98
	⟨N9E54i04⟩	-0.990	0.70	0.98
⟨**Gänseblümchen**⟩	⟨N6E56i03⟩	-0.819	0.40	1.16
(6df)	⟨N6E56i04⟩	-1.705	0.19	1.11
	⟨N6E56i06D⟩	-0.365	0.47	1.12
	⟨N6E56i06F⟩	0.253	0.44	1.02
⟨**Laubbäume**⟩	⟨N6E61i01⟩	-0.026	0.43	0.94
(6d)	⟨N6E61i03⟩	-1.719	0.44	1.10
	⟨N6E61i04⟩	-0.196	0.34	1.26
⟨**Sparlampe**⟩	⟨N9E83i01⟩	-0.947	0.55	0.88
(9d)	⟨N9E83i02⟩	0.222	0.46	0.97
	⟨N9E83i04⟩	1.320	0.48	1.02
	⟨N9E83i05⟩	1.149	0.52	1.06
	⟨N9E83i07⟩	2.006	0.30	1.04
⟨**Solarzellen**⟩	⟨N9E84i07⟩	-2.815	0.27	1.01
(9df)	⟨N9E84i08⟩	0.101	0.67	0.95
	⟨N9E84i09D⟩	0.663	0.69	0.94
	⟨N9E84i09F⟩	-0.020	0.53	0.89
	⟨N9E84i10D⟩	-0.202	0.55	0.89
	⟨N9E84i10F⟩	-0.891	0.71	0.76
	⟨N9E84i11D⟩	-0.866	0.53	0.90
	⟨N9E84i11F⟩	-1.599	0.43	0.95
	⟨N9E84i12⟩	*-0.137	0.47	1.10

Tabelle E.2 – Personen-Test-Analysen: Übersicht über die Itemparameter

Anhang F

Fragebogen

Der folgende Fragebogen wurde begleitend zum HarmoS-Experimentiertest 2008 eingesetzt. Die Beantwortung der Fragen war für die Schülerinnen und Schüler fakultativ.

Fragebogen: N9d HarmoS Naturwissenschaften+

	Name:	.....
	Vorname:	.....
⟨i01⟩	Schulort:	.....
⟨i02⟩	Klasse:	.....
⟨i03⟩	Alter:	..... Jahre
⟨i04⟩	Geschlecht:	männlich ◯ weiblich ◯

1. Welche Sprache sprichst du am meisten zu Hause?

⟨i11⟩ .....

2. Setze ein Kreuz bei den Stichworten, die zutreffen!

⟨i21⟩	Hast du bei dir zu Hause ein eigenes Zimmer?	◯
⟨i22⟩	Hast du bei dir zu Hause einen eigenen Schreibtisch zum Lernen?	◯
⟨i23⟩	Hast du bei dir zu Hause einen Taschenrechner, der dir gehört?	◯
⟨i24⟩	Hast du bei dir zu Hause einen Computer mit Internetanschluss, den du benutzen darfst?	◯
⟨i25⟩	Hast du bei dir zu Hause einen ruhigen Ort zum Lernen?	◯
⟨i26⟩	Hast du bei dir zu Hause Sachbücher, die dir bei deinen Schularbeiten weiterhelfen?	◯
⟨i27⟩	Hast du bei dir zu Hause Klassische Literatur (z. B. Goethe, Schiller)?	◯
⟨i28⟩	Hast du bei dir zu Hause Kunstwerke (z. B. Bilder)?	◯

3. Kreuze für jede Aussage das Zutreffende an.

		sehr viel	viel	ab und zu	fast nie
⟨i31⟩	Wir sprechen zu Hause über Themen aus der Schule.	◯	◯	◯	◯
⟨i32⟩	Wir gehen mit der Familie oft nach draussen (in den Wald, an Seen, in die Berge usw.).	◯	◯	◯	◯
⟨i33⟩	Wir besuchen mit der Familie Museen, Lehrpfade usw.	◯	◯	◯	◯
⟨i34⟩	Wenn wir in den Ferien sind, schauen wir uns Sachen an diesen Orten an und unternehmen dort Ausflüge.	◯	◯	◯	◯
⟨i36⟩	Ich gehe in Bibliotheken, Mediotheken und sehe mir Bücher und Filme zu Themen an oder leihe sie aus.	◯	◯	◯	◯
⟨i37⟩	Ich sehe im Fernsehen Nachrichtensendungen (z. B. Tagesschau).	◯	◯	◯	◯
⟨i38⟩	Ich lese Tageszeitungen.	◯	◯	◯	◯

4. Wie gross ist dein Interesse an folgenden Fächern?

Kreuze für jedes Fach das Zutreffende an.

		sehr gross	gross	ein wenig	gering
⟨i41⟩	Deutsch	◯	◯	◯	◯
⟨i42⟩	Mathematik	◯	◯	◯	◯
⟨i43⟩	Natur-Mensch-Mitwelt / Mensch und Umwelt	◯	◯	◯	◯
⟨i43.1⟩	Natur+Technik / Naturkunde / Sciences naturelles	◯	◯	◯	◯
⟨i43.2⟩	Geschichte	◯	◯	◯	◯
⟨i43.3⟩	Geografie	◯	◯	◯	◯
⟨i43.4⟩	Hauswirtschaft	◯	◯	◯	◯
⟨i43.5⟩	Religion, Lebenskunde, Ethik	◯	◯	◯	◯
⟨i43.6⟩	Sciences expérimentales	◯	◯	◯	◯
⟨i43.7⟩	Physique	◯	◯	◯	◯
⟨i43.8⟩	Chimie	◯	◯	◯	◯
⟨i43.9⟩	Biologie	◯	◯	◯	◯
⟨i44⟩	Sport	◯	◯	◯	◯
⟨i45⟩	Bildnerisches Gestalten / Zeichnen	◯	◯	◯	◯
⟨i46⟩	Technisches Gestalten / Textiles Gestalten	◯	◯	◯	◯
⟨i47⟩	Musik	◯	◯	◯	◯

5. Welche Note hattest du im letzten Zeugnis im Fach Naturkunde, im Bereich Natur?

Kreuze die zutreffende Note an.

⟨i51⟩ 6 ◯ 5-6 ◯ 5 ◯ 4-5 ◯ 4 ◯ 3-4 ◯ 3 und tiefer ◯ weiss nicht ◯

6. Kreuze zu jeder Frage die für dich richtige Einschätzung an.

	stimmt ...	genau	eher	eher nicht	gar nicht
⟨i61⟩	Diese Aufgaben haben mich interessiert.	◯	◯	◯	◯
⟨i62⟩	Zu diesen Aufgaben wusste ich schon viel.	◯	◯	◯	◯
⟨i63⟩	Diese Aufgaben waren für mich schwierig.	◯	◯	◯	◯
⟨i64⟩	Es ist wichtig, dass ich solche Aufgaben lösen kann.	◯	◯	◯	◯
⟨i65⟩	Diese Aufgaben waren für mich langweilig.	◯	◯	◯	◯
⟨i66⟩	Bei diesen Aufgaben habe ich nichts Neues gelernt.	◯	◯	◯	◯
⟨i67⟩	Diese Aufgaben finde ich spannend.	◯	◯	◯	◯

⟨i68⟩ Dieses Thema hat mich am meisten interessiert: .....

⟨i69⟩ Dieses Thema war für mich am schwierigsten: .....

7. Wie sehr interessieren dich Themen zu Natur, Technik, Gesundheit und Umwelt?

		sehr stark	stark	ein wenig	gar nicht
⟨i71⟩	Mich interessieren Themen zur Erde (Steine, Wasser, Wetter usw.).	◯	◯	◯	◯
⟨i72⟩	Mich interessieren Themen zum Universum (Sonne, Mond, Planeten, Sterne).	◯	◯	◯	◯
⟨i73⟩	Mich interessieren Themen zu Tieren, Pflanzen und Lebensräumen.	◯	◯	◯	◯
⟨i74⟩	Mich interessieren Themen zu Energie, Elektrizität, Maschinen und Geräte usw.	◯	◯	◯	◯
⟨i75⟩	Mich interessieren Themen zu Materialien (Holz, Steine, Metall, Stoffe usw.).	◯	◯	◯	◯
⟨i76⟩	Mich interessieren Themen zu Gesundheit, Essen und Trinken, Bewegung.	◯	◯	◯	◯
⟨i77⟩	Mich interessieren Themen zu unserer Umwelt, wie sich die Umgebung verändert.	◯	◯	◯	◯

Abbildungsverzeichnis

Tabellenverzeichnis

Literaturverzeichnis

Adamina, M. (2008). *Vorstellungen von Schülerinnen und Schülern zu raum-, zeit- und geschichtsbezogenen Themen.* Unveröffentlichte Dissertation, Universität Münster.

APU: Archenhold, F. (Hrsg.). (1988b). *Science at age 15: A review of APU survey findings 1980-84.* London: Her Majesty's Stationery Office.

APU: Gott, R., Davey, A., Gamble, R., Head, J., Khaligh, N., Murphy, P., ... Welford, G. (1985). *Science in schools: Ages 13 and 15* (Bericht Nr. 3). Department of Education and Science, Department of Education for Northern Ireland, Welsh Office.

APU: Harlen, W., Black, P. & Johnson, S. (1981). *Science in schools: Age 11* (Bericht Nr. 1). Department of Education and Science, Department of Education for Northern Ireland, Welsh Office.

APU: Harlen, W., Black, P., Johnson, S., Palacio, D. & Russell, T. (1984). *Science in schools: Age 11* (Bericht Nr. 3). Department of Education and Science, Department of Education for Northern Ireland, Welsh Office.

APU: Russell, T. (Hrsg.). (1988a). *Science at age 11: A review of APU survey findings 1980-84.* London: Her Majesty's Stationery Office.

APU: Schofield, B. (Hrsg.). (1989). *Science at age 13: A review of APU survey findings 1980-84.* London: Her Majesty's Stationery Office.

Artelt, C., Baumert, J., Julius-McElvany, N. & Peschar, J. (2003). *Das Lernen lernen. Voraussetzungen für lebensbegeleitendes Lernen. Ergebnisse von PISA 2000.* Paris: OECD.

Ayala, C. C., Shavelson, R. J. & Ayala, M. A. (2001). *On the cognitive interpretation of performance assessment scores* (CSE Technical Report Nr. 546). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing, University of California.

Ayala, C. C., Shavelson, R. J., Yin, Y. & Schultz, S. E. (2002). Reasoning dimensions underlying science achievement: The case of performance assessment. *Educational Assessment, 8* (2), 101-121.

Backhausen, K., Erichson, B., Plinke, W. & Weiber, R. (2006). *Multivariate Analysemethoden* (11. Aufl.). Berlin: Springer.

Baker, E. L., 0'Neil, H. F. & Linn, R. L. (1993). Policy and validity prospects for performance-based assessment. *American Psychologist*, *48* (12), 1210-1218.

Bass, K. M., Magone, M. E. & Glaser, R. (2002). *Informing the design of performance assessments using a content-process analysis of two NAEP science tasks* (CSE Technical Report Nr. 564). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing, University of California.

Baxter, G. P., Elder, A. D. & Glaser, R. (1995). *Cognitive analysis of a science performance assessment* (CSE Technical Report Nr. 398). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing, University of California.

Baxter, G. P. & Glaser, R. (1998). Investigating the cognitive complexity of science assessments. *Educational Measurement: Issues and Practice*, *17* (3), 37-45.

Baxter, G. P. & Shavelson, R. J. (1994). Science performance assessments: Benchmarks and surrogates. *International Journal of Educational Research*, *21*, 279-298.

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.

Becker, M., Lüdtke, O., Trautwein, U. & Baumert, J. (2006). Leistungszuwachs in Mathematik. Evidenz für einen Schereneffekt im mehrgliedrigen Schulsystem? *Zeitschrift für Pädagogische Psychologie*, *20*, 233-242.

Bernholt, S., Parchmann, I. & Commons, M. L. (2009). Kompetenzmodellierung zwischen Forschung und Unterrichtspraxis. *Zeitschrift für Didaktik der Naturwissenschaften*, *15*, 219-245.

Bond, T. G. & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences.* Mahwah: Lawrence Erlbaum.

Borg, I. & Staufenbiel, T. (2007). *Lehrbuch – Theorien und Methoden der Skalierung.* Bern: Verlag Hans Huber.

Börlin, J. (2012). *Das Experiment als Lerngelegenheit. Vom interkulturellen Vergleich des Physikunterrichts zu Merkmalen seiner Qualität.* Berlin: Logos Verlag.

Britton, E. D. & Schneider, S. A. (2007). Large-scale assessments in science education. In S. K. Abell & N. G. Lederman (Hrsg.), *Handbook of research on science education* (S. 1007-1040). Mahwah: Lawrence Erlbaum.

Brown, J. H. & Shavelson, R. J. (1996). *Assessing hands-on science: A teacher's guide to performance assessment.* Thousand Oaks: Corwin Press.

Brühwiler, C., Kis-Fedi, P., Buccheri, G. & Mariotta, M. (2009). Engagement in den Naturwissenschaften, Berufserwartung und Geschlechterunterschiede. In Bundesamt für Statistik (Hrsg.), *PISA 2006: Analysen zum Kompetenzbereich Naturwissenschaften* (S. 41-92). Neuchâtel: Bundesamt für Statistik, BFS.

Bruner, J. S., Goodnow, J. J. & Austin, G. A. (1956). *A study of thinking.* New York: NY Science Editions.

Bühl, A. (2010). *PASW 18. Einführung in die moderne Datenanalyse.* München: Pearson Studium.

Burbules, N. C. & Linn, M. C. (1988). Response to contradiction: Scientific reasoning during adolescence. *Journal of Educational Psychology*, *80* (1), 67-75.

Bybee, R. W. (1997). Toward an understanding of scientific literacy. In W. Gräber & C. Bolte (Hrsg.), *Scientific literacy: An international symposium* (S. 37-68). Kiel: Institut für die Didaktik der Naturwissenschaften (IPN), Universität Kiel.

Bybee, R. W. (2002). Scientific Literacy – Mythos oder Realität? In W. Gräber, P. Nentwig, T. Koballa & R. Evans (Hrsg.), *Scientific Literacy* (S. 21-43). Opladen: Leske + Buderich.

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.

Carrier, M. (2006). *Wissenschaftstheorie.* Hamburg: Junius Verlag.

Chinn, C. A. & Brewer, W. F. (1992). Psychological responses to anomalous data. In *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society* (S. 165-170). Hillsdale: Lawrence Erlbaum.

Chinn, C. A. & Brewer, W. F. (1993). The role of anomalous data in knowledge acquisition: A theoretical framework and implications for science instruction. *Review of Educational Research*, *63* (1), 1-49.

Chinn, C. A. & Brewer, W. F. (1998). An empirical test of a taxonomy of responses to anomalous data in science. *Journal of Research in Science Teaching*, *35* (6), 623-654.

Chinn, C. A. & Malhotra, B. A. (2002). Children's responses to anomalous scientific data: How is conceptual change impeded. *Journal of Educational Psychology*, *94* (2), 327-343.

Clagett, M. (1959). *The science of mechanics in the Middle Ages.* Madison: The University of Wisconsin Press.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Hrsg.), *Educational measurement* (2. Aufl., S. 443-507). Washington DC: American Council on Education.

Dolin, J. (2007). Science education standards and science assessment in Denmark. In D. Waddington, P. Nentwig & S. Schanze (Hrsg.), *Making it comparable: Standards in science education* (S. 71-82). Münster: Waxmann.

Dörner, D. (1976). *Problemlösen als Informationsverarbeitung.* Stuttgart: Kohlhammer.

Dörner, D. (2006). *Die Logik des Misslingens. Strategisches Denken in komplexen Situationen* (5. Aufl.). Hamburg: Rowohlt.

Driessen, H. (2007). Development and evaluation of standards in secondary science education in the Netherlands. In D. Waddington, P. Nentwig & S. Schanze (Hrsg.), *Making it comparable: Standards in science education* (S. 221-236). Münster: Waxmann.

Dunbar, K. (1993). Concept discovery in a scientific domain. *Cognitive Science*, *17*, 397-434.

Dunbar, K. & Klahr, D. (1989). Developmental differences in scientific discovery processes. In D. Klahr & K. Kotovsky (Hrsg.), *Complex information processing: The impact of Herbet A. Simon* (S. 109-143). Hillsdale: Lawrence Erlbaum.

EDK: Schweizerische Konferenz der kantonalen Erziehungsdirektoren. (2011). *Nationale Bildungsstandards: Grundkompetenzen für die Naturwissenschaften* (Bericht). Schweizerische Konferenz der kantonalen Erziehungsdirektoren (EDK). Zugriff auf `http://edudoc.ch/record/96787/files/grundkomp_nawi_d.pdf` (Zugriff: 20.1.2012)

Ehmer, M. & Hammann, M. (2007). Alternative Argumentationsstrategien als Ursache methodischer Schülerfehler beim Experimentieren. In H. Bayrhuber et al. (Hrsg.), *Ausbildung und Professionalisierung von Lehrkräften: Internationale Tagung der Fachgruppe Biologiedidkatik im VBIO* (S. 27-30). Kassel: Universität Kassel.

Einhaus, E. (2007). *Schülerkompetenzen im Bereich Wärmelehre. Entwicklung eines Testinstruments zur Überprüfung und Weiterentwicklung eines normativen Modells fachbezogener Kompetenzen.* Berlin: Logs.

Fairbrother, B. (1991). Principles of practical assessment. In B. E. Woolnough (Hrsg.), *Practical science* (S. 153-166). Milton Keynes: Open University Press.

Fairbrother, R. W. (1988). *Assessment of practical work for the GCSE.* York: Longman Resources Unit.

Feyerabend, P. (1999 [1976]). *Wider den Methodenzwang* (7. Aufl.). Frankfurt am Main: Suhrkamp.

Finke, E. (1999). Faktoren der Entwicklung von Biologieinteressen in der Sekundarstufe I. In R. Duit & J. Mayer (Hrsg.), *Studien zur naturwissenschaftsdidaktischen Lern- und Interessenforschung* (S. 103-117). Kiel: Institut für die Pädagogik der Naturwissenschaften (IPN), Universität Kiel.

Fischer, G. H. (1997). Unidimensional linear logistic Rasch models. In W. J. van der Linden & R. K. Hambleton (Hrsg.), *Handbook of modern item response theory* (S. 225-243). New York: Springer.

Fischer, H. E. & Draxler, D. (2007). Konstruktion und Bewertung von Physikaufgaben. In E. Kicher, R. Girwidz & P. Häußler (Hrsg.), *Physikdidaktik: Theorie und Praxis*

(S. 639-655). Berlin, Heidelberg: Springer.

Fischer, H. E., Klemm, K., Leutner, D., Sumfleth, E., Tiemann, R. & Wirth, J. (2003). Naturwissenschaftsdidaktische Lehr- und Lernforschung: Defizite und Desiderata. *Zeitschrift für Diaktik der Naturwissenschaften*, *9*, 179-209.

Fraser, B. J. (1980). Development and validation of a test oh enquiry skills. *Journal of Research in Science Teaching*, *17* (1), 7-16.

Frege, G. (1892). Über Sinn und Bedeutung. *Zeitschrift für Philosophie und philosophische Kritik*, *NF 100*, 25-50.

Funke, J. (2003). *Problemlösendes Denken*. Stuttgart: Kohlhammer.

Gao, X., Shavelson, R. J. & Baxter, G. P. (1994). Generalizability of large-scale performance assessments in science: Promises and problems. *Applied Measurement in Education*, *7* (4), 323-342.

Garrett, R. M. (1986). Problem-solving in science education. *Studies in Science Education*, *13*, 70-95.

Genz, H. (1996). Buridans Esel und die Spontane Symmetriebrechung. *Physik in unserer Zeit*, *27* (5), 218-220.

Gott, R. & Duggan, S. (1995). *Investigative work in the science curriculum*. Buckingham, Philadelphia: Open University Press.

Grube, C., Möller, A. & Mayer, J. (2007). Dimensionen eines Kompetenzstrukturmodells zum Experimentieren. In H. Bayrhuber et al. (Hrsg.), *Ausbildung und Professionalisierung von Lehrkräften: Internationale Tagung der Fachgruppe Biologiedidkatik im VBIO* (S. 31-34). Kassel: Universität Kassel.

Gut, C. & Labudde, P. (2010). Assessment of students' practical performance in science: The Swiss HarmoS project. In G. Çakmakci & M. Taşar (Hrsg.), *Contemporary science education research: Learning and assessment* (S. 295-298). Istanbul: Pegem Akademi, ESERA.

Gut, C., Labudde, P. & Ramseier, E. (2010). Large-scale Experimentiertests: Ansätze zur Analyse von Itemschwierigkeiten. In D. Höttecke (Hrsg.), *Entwicklung naturwissenschaftlichen Denkens zwischen Phänomen und Systematik* (S. 245-247). Berlin: LIT Verlag.

Haertel, E. H. & Linn, R. L. (1996). Comparability. In G. W. Phillips (Hrsg.), *Technical issues in large-scale performance assessment* (S. 59-78). Washington DC: National Center for Education Statistics, U.S. Department of Education.

Hafner, R. (2007). Standards in science education in Australia. In D. Waddington, P. Nentwig & S. Schanze (Hrsg.), *Making it comparable: Standards in science education* (S. 23-59). Münster: Waxmann.

Hamilton, L. S., Nussbaum, E. M. & Snow, R. E. (1997). Interview procedures for

validating science assessments. *Applied Measurement in Education*, *10*, 181-200.

Hammann, M. (2004). Kompetenzentwicklungsmodelle. *Der mathematische und naturwissenschaftliche Unterricht*, *57* (4), 196-203.

Hammann, M. (2007). Das Scientific Discovery as Dual Search-Modell. In D. Krüger & H. Vogt (Hrsg.), *Theorien in der biologiedidaktischen Forschung* (S. 187-196). Berlin: Springer.

Hammann, M., Phan, T. T. H. & Bayrhuber, H. (2007). Experimentieren als Problemlösen: Lässt sich das SDDS-Modell nutzen, um unterschiedliche Dimensionen beim Experimentieren zu messen? *Zeitschrift für Erziehungswissenschaft*, *10* (Sonderheft 8), 33-49.

Hammann, M., Phan, T. T. H., Ehmer, M. & Bayrhuber, H. (2006). Fehlerfreies Experimentieren. *Der mathematische und naturwissenschaftliche Unterricht*, *59* (5), 292-299.

Hammann, M., Phan, T. T. H., Ehmer, M. & Grimm, T. (2008). Assessing pupils' skills in experimetation. *Journal of Biological Education*, *42* (2), 66-72.

HarmoS: Konsortium HarmoS Naturwissenschaften. (2008). *HarmoS Naturwissenschaften+: Kompetenzmodell und Vorschläge für Bildungsstandards* (Bericht). Konsortium HarmoS Naturwissenschaften.

HarmoS: Konsortium HarmoS Naturwissenschaften. (2010). *Basisstandards für die Naturwissenschaften: Unterlagen für den Anhörungsprozess* (Bericht). Schweizerische Konferenz der kantonalen Erziehungsdirektoren (EDK).

Häußler, P., Bünder, W., Duit, R., Gräber, W. & Mayer, J. (1998). *Naturwissenschaftsdidaktische Forschung: Perspektiven für die Unterrichtspraxis.* Kiel: Institut für die Pädagogik der Naturwissenschaften (IPN), Universität Kiel.

Häußler, P. & Hoffmann, L. (1995). Physikunterricht – an den Interessen von Mädchen und Jungen orientiert. *Unterrichtswissenschaften*, *23* (2), 107-126.

Herzog, W. (2007). Pro und Kontra Bildungsstandards. Die Perspektive eines Skeptikers. In P. Labudde (Hrsg.), *Bildungsstandards am Gymnasium – Korsett oder Katalysator?* (S. 57-64). Bern: h.e.p. verlag.

Hodson, D. (1990). A critical look at practical work in school science. *School Science Review*, *70* (256), 33-40.

Hodson, D. (1992). Assessment of practical work. *Science and Education*, *1*, 115-144.

Hoffmann, L., Häußler, P. & Lehrke, M. (1998). *Die IPN-Interessenstudie Physik.* Kiel: Institut für die Pädagogik der Naturwissenschaften (IPN), Universität Kiel.

Hoffmann, L. & Lehrke, M. (1986). Eine Untersuchung über Schülerinteressen an Physik und Technik. *Zeitschrift für Pädagogik*, *32* (2), 189-204.

Holstermann, N. & Bögeholz, S. (2007). Interesse von Jungen und Mädchen an naturwis-

senschaftlichen Themen am Ende der Sekundarstufe I. *Zeitschrift für Didaktik der Naturwissenschaften*, *13*, 71-86.

Höttecke, D. (2001). Die Vorstellungen von Schülern und Schülerinnen von der "Natur der Naturwissenschaften". *Zeitschrift für Didaktik der Naturwissenschaften*, *7*, 7-23.

Johsua, S. & Dupin, J. J. (1987). Taking into account student conceptions in instructional strategy: An example in physics. *Cognition and Instruction*, *4* (2), 117-135.

Jonassen, D. H. (2000). Toward a design theory of problem solving. *Educational Technology – Research & Development*, *48* (4), 1042-1629.

Jovanovic, J., Solano-Flores, G. & Shavelson, R. J. (1994). Performace-based assessments: Will gender differences in science achievement be eliminated? *Education and Urban Society*, *26* (4), 352-366.

Kauertz, A. (2008). *Schwierigkeitserzeugende Merkmale physikalischer Leistungstestaufgaben.* Berlin: Logos Verlag.

Kauertz, A., Fischer, H. E., Lau, A. & Neumann, K. (2008). Kompetenzmessung durch Leistungstests – Hilfe oder Druckmittel? *Der mathematische und naturwissenschaftliche Unterricht*, *61* (2), 75-79.

Kauertz, A., Fischer, H. E., Mayer, J., Sumfleth, E. & Walpuski, M. (2010). Standardbezogene Kompetenzmodellierung in den Naturwissenschaften der Sekundarstufe I. *Zeitschrift für Didaktik der Naturwissenschaften*, *16*, 135-153.

Kintsch, W. & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, *85*, 363-394.

Klahr, D. (2000). *Exploring science. the cognition and development of discovery process.* Cambridge: The MIT Press.

Klahr, D. & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Science*, *12*, 1-48.

Klayman, J. & Ha, Y.-W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, *94* (2), 211-228.

Klayman, J. & Ha, Y.-W. (1989). Hypothesis testing in rule discovery: Strategy, structure, and content. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15* (4), 596-604.

Klein, S. P., Jovanovic, J., Stecher, B. M., McCafferey, D., Shavelson, R. J., Haertel, E., ... Comfort, K. (1997). Gender and racial/ethnic differences on performance assessments in science. *Educational Evaluation and Policy Analysis*, *19* (2), 83-97.

Klieme, E. (2007). Bildungsstandards, Leistungsmessung und Unterrichtsqualität. In P. Labudde (Hrsg.), *Bildungsstandards am Gymnasium* (S. 77-84). Bern: h.e.p. verlag.

Klieme, E., Avenarius, H., Blum, W., Döbrich, P., Gruber, H., Prenzel, M., ... Voll-

mer, H. J. (2007). *Zur Entwicklung nationaler Bildungsstandards: Eine Expertise* (Bericht). Bonn, Berlin: Bundesministerium für Bildung und Forschung.

Klieme, E. & Hartig, J. (2007). Kompetenzkonzepte in den Sozialwissenschaften und im erziehungswissenschaftlichen Diskurs. *Zeitschrift für Erziehungswissenschaft*, *10* (Sonderheft 8), 11-29.

Klieme, E. & Leutner, D. (2006). Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen. Beschreibung eines neu eingerichteten Schwerpunktproramm der DFG. *Zeitschrift für Pädagogik*, *52* (6), 876-903.

Klos, S., Henke, C., Kieren, C., Walpuski, M. & Sumfleth, E. (2008). Naturwissenschaftliches Experimentieren und chemisches Fachwissen – zwei verschiedene Kompetenzen. *Zeitschrift für Pädagogik*, *54* (3), 304-321.

KMK: Kultusministerkonferenz. (2005a). *Bildungsstandards im Fach Biologie für den Mittleren Schulabschluss.* München: Wolters Kluwer.

KMK: Kultusministerkonferenz. (2005b). *Bildungsstandards im Fach Chemie für den Mittleren Schulabschluss.* München: Wolters Kluwer.

KMK: Kultusministerkonferenz. (2005c). *Bildungsstandards im Fach Physik für den Mittleren Schulabschluss.* München: Wolters Kluwer.

Koslowski, B. (1996). *Theory and evidence. The development of scientific reasoning.* Cambridge: The MIT Press.

Kuhn, D. (1989). Children and adults as intuitive scientists. *Psychological Review*, *96* (4), 674-689.

Kuhn, D., Amsel, E. A. & O'Loughlin, M. (1988). *The development of scientific thinking.* New York: Academic Press.

Kuhn, D. & Phelps, E. (1982). The development of problem-solving strategies. *Advances in Child Development and Behavior*, *17*, 1-44.

Kuhn, T. S. (1997 [1962]). *Die Struktur wissenschaftlicher Revolution* (5. Aufl.). Frankfurt am Main: Suhrkamp Verlag.

Kulgemeyer, C. (2009). *PISA-Aufgaben im Vergleich.* Norderstedt: Books on Demand.

Kulgemeyer, C. & Schecker, H. (2007). PISA 2000 bis 2006 – Ein Vergleich anhand eines Strukturmodells für naturwissenschaftliche Aufgaben. *Zeitschrift für Didaktik der Naturwissenschaften*, *13*, 199-220.

Labudde, P. (2007). How to develop, implement and assess standards in science education? 12 challanges from a Swiss perspective. In D. Waddington, P. Nentwig & S. Schanze (Hrsg.), *Making it comparable: Standards in science education* (S. 277-301). Münster: Waxmann.

Labudde, P., Duit, R., Fickermann, D., Fischer, H., Harms, U., Mikelskis, H., ... Weigl-

hofer, H. (2009). Schwerpunkttagung "Kompetenzmodelle und Bildungsstandards: Aufgaben für die naturwissenschaftliche Forschung". *Zeitschrift für Didaktik der Naturwissenschaften*, *15*, 343-370.

Labudde, P., Metzger, S. & Gut, C. (2009). Bildungsstandards: Validierung des Schweizer Kompetenzmodells. In D. Höttecke (Hrsg.), *Chemie- und Physikdidaktik für die Leramtsausbildung* (S. 307-317). Berlin: LIT Verlag.

Labudde, P. & Stebler, R. (1999). Lern- und Prüfungsaufgaben für den Physikunterricht: Erträge aus dem TIMSS-Experimentiertest. *Unterricht Physik*, *10* (54), 23-31.

Leach, J. (2002). Students' understanding of the nature of science and its influence on labwork. In D. Psillos & H. Niedderer (Hrsg.), *Teaching and learning in the science laboratory* (S. 41-48). Dordrecht: Kluwer.

Linn, R. L., Baker, E. L. & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher*, *20* (8), 15-21.

Lunetta, V. N. (1998). The school science laboratory: Historical perspectives and contextes for contemporary teaching. In B. J. Fraser & K. G. Tobin (Hrsg.), *International handbook of science education* (S. 249-262). Dordrecht: Kluwer.

Maag Merki, K. (2007). Bildungsstandards – Konzept und Begrifflichkeiten. In P. Labudde (Hrsg.), *Bildungsstandards am Gymnasium – Korsett oder Katalysator?* (S. 17-25). Bern: h.e.p. verlag.

Mach, E. (1963 [1883]). *Die Mechanik: Historisch-kritisch dargestellt* (9. Aufl.). Darmstadt: Wissenschaftliche Buchgesellschaft.

Mannel, S., Sumfleth, E. & Walpuski, M. (2010). Student assessment in the area of acquirement of knowledge. In G. Çakmakci & M. Taşar (Hrsg.), *Contemporary science education research: Learning and assessment* (S. 299-305). Istanbul: Pegem Akademi, ESERA.

Masters, G. N. & Wright, B. D. (1997). The partial credit model. In W. J. van der Linden & R. K. Hambleton (Hrsg.), *Handbook of modern item response theory* (S. 101-121). New York: Springer.

Mayer, J. (2007). Erkenntnisgewinnung als wissenschaftliches Problemlösen. In D. Krüger & H. Vogt (Hrsg.), *Theorien in der biologiedidaktischen Forschung* (S. 177-186). Berlin: Springer.

Mayer, J., Grube, C. & Möller, A. (2008). Kompetenzmodell naturwissenschaftlicher Erkenntnisgewinnung. In U. Harms & A. Sandmann (Hrsg.), *Lehr- und Lernforschung in der Biologiedidaktik – Ausbildung und Professionalisierung von Lehrkräften* (Bd. 3, S. 63-79). Innsbruck: StudienVerlag.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, *23* (2), 13-23.

Messick, S. (1996). Validity of performance assessments. In G. W. Phillips (Hrsg.), *Technical issues in large-scale performance assessment* (S. 1-18). Washington DC: National Center for Education Statistics, U.S. Department of Education.

Millar, R. (1989). What is scientific method and can it be taught? In J. J. Wellington (Hrsg.), *Skills and processes in science education* (S. 249-262). London: Routledge.

Millar, R. (2007). How standards in science education are set and monitored in the English education system. In D. Waddington, P. Nentwig & S. Schanze (Hrsg.), *Making it comparable: Standards in science education* (S. 83-100). Münster: Waxmann.

Millar, R. & Driver, R. (1987). Beyond processes. *Studies in Science Education*, *14*, 33-62.

Miller, M. D. & Linn, R. L. (2000). Validation of performance-based assessments. *Applied Psychological Measurement*, *24* (4), 367-378.

Möller, A., Grube, C. & Mayer, J. (2007). Kompetenzniveaus der Erkenntnisgewinnung bei Schülerinnen und Schülern der Sekundarstufe I. In H. Bayrhuber et al. (Hrsg.), *Ausbildung und Professionalisierung von Lehrkräften: Internationale Tagung der Fachgruppe Biologiedidkatik im VBIO* (S. 55-58). Kassel: Universität Kassel.

Moser, U. & Angelone, D. (2009). Unterrichtszeit, Unterrichtsorganisation, Leistung und Interesse. In Bundesamt für Statistik (Hrsg.), *PISA 2006: Analysen zum Kompetenzbereich Naturwissenschaften* (S. 9-39). Neuchâtel: Bundesamt für Statistik, BFS.

Moser, U. & Angelone, D. (2011). *PISA 2009: Porträt des Kantons Zürich* (Bericht). Zürich. Zugriff auf `http://pisa.educa.ch/sites/default/files/20111205/pisa2009-zh_0.pdf` (Zugriff: 20.1.2012)

Murphy, P. & Gott, R. (1984). *Science: Assessment framework age 13 and 15* (Bd. 2). Letchworth: The Garden City Press.

NAEP: National Assessment Governing Board. (2008). *Science framework for the 2009 national assessment of educational progress.* Washington, DC: U.S. Goverment Printing Office. Zugriff auf `http://www.nagb.org/publications/frameworks/science-09.pdf` (Zugriff am 20.1.2012)

National Standards: National Research Counsil. (1996). *National science education standards.* Washington, DC: National Acadamy Press. Zugriff auf `http://www.nap.edu/openbook.php?record_id=4962` (Zugriff am 20.1.2012)

Nawrath, D., Maiseyenka, V. & Schecker, H. (2011). Experimentelle Kompetenz: Ein Modell für die Unterrichtspraxis. *Physik in der Schule*, *60* (6), 42-48.

Nentwig, P. & Waddington, D. (2007). Standards: an international comparison. In D. Waddington, P. Nentwig & S. Schanze (Hrsg.), *Making it comparable: Standards*

in science education (S. 375-403). Münster: Waxmann.

Neuhaus, B. & Vogt, H. (2004). Forschungsergebnisse zur Neugestaltung des Unterrichts in den Naturwissenschaften. In H. Bayrhuber, B. Ralle, K. Reiss, L.-H. Schön & H. J. Vollmer (Hrsg.), *Konsequenzen aus PISA: Perspektiven der Fachdidaktik* (S. 197-198). Innsbruck: StudienVerlag.

Neumann, K., Kauertz, A., Lau, A., Notarp, H. & Fischer, H. E. (2007). Die Modellierung physikalischer Kompetenz und ihre Entwicklung. *Zeitschrift für Didaktik der Naturwissenschaften*, *13*, 101-121.

Nidegger, C., Moreau, J. & Gingins, F. (2009). Kompetenzen der Schülerinnen und Schüler in den Naturwissenschaften: Erkenntnisse aus PISA und HarmoS. In Bundesamt für Statistik (Hrsg.), *PISA 2006: Analysen zum Kompetenzbereich Naturwissenschaften* (S. 93-118). Neuchâtel: Bundesamt für Statistik, BFS.

Oelkers, J. (2007). Bildungsstandards im Gymnasium – Ein neues Problem? In P. Labudde (Hrsg.), *Bildungsstandards am Gymnasium – Korsett oder Katalysator?* (S. 27-36). Bern: h.e.p. verlag.

PISA: Konsortium PISA.ch. (2011). *PISA 2009: Regionale und kantonale Ergebnisse* (Bericht). Bern, Neuchâtel: BBT/EDK und Konsortium PISA.ch. Zugriff auf `http://pisa.educa.ch/sites/default/files/20111205/pisa_de_0.pdf` (Zugriff am 20.1.2012)

PISA: Prenzel, M., Schöps, K., Rönnebeck, S., Senkbeil, M., Walter, O., Carstensen, C. H. & Hammann, M. (2007). Die naturwissenschaftliche Kompetenz im internationalen Vergleich. In PISA-Konsortium Deutschland (Hrsg.), *PISA 2006: Die Ergebnisse der dritten internationalen Vergleichsstudie* (S. 63-105). Münster: Waxmann.

PISA: Prenzel, M. & Schütte, K. (2008). Interesse an den Naturwissenschaften. In PISA-Konsortium Deutschland (Hrsg.), *PISA 2006 in Deutschland. Die Kompetenzen der Jugendlichen im dritten Ländervergleich* (S. 95-106). Münster: Waxmann.

PISA: Rost, J., Walter, O., Carstensen, C. H., Senkbeil, M. & Prenzel, M. (2004). Naturwissenschaftliche Kompetenz. In PISA-Konsortium Deutschland (Hrsg.), *PISA 2003: Der Bildungsstand der Jugendlichen in Deutschland – Ergebnisse des zweiten internationalen Vergleichs* (S. 111-146). Münster: Waxmann.

Polanyi, M. (1969). *Knowing and being.* London: Routledge and Kegan Paul.

Prenzel, M., Häußler, P., Rost, J. & Senkbeil, M. (2002). Der PISA-Naturwissenschaftstest: Lassen sich die Aufgabenschwierigkeiten vorhersagen? *Unterrichtswissenschaft*, *30* (1), 120-135.

Ramseier, E., Labudde, P. & Adamina, M. (2011). Validierung des Kompetenzmodells HarmoS Naturwissenschaften: Fazite und Defizite. *Zeitschrift für Didaktik der Naturwissenschaften*, *17*, 7-33.

Rosenquist, A., Shavelson, R. J. & Ruiz-Primo, M. A. (2000). *On the "exchangeability" of hands-on and computer-simulated science performance assessments* (CSE Technical Report Nr. 531). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing, Stanford University.

Rost, J. (2004a). *Lehrbuch Testtheorie – Testkonstruktion* (2. Aufl.). Bern: Verlag Hans Huber.

Rost, J. (2004b). Psychometrische Modelle zur Überprüfung von Bildungsstandards anhand von Kompetenzmodellen. *Zeitschrift für Pädagogik*, *50* (5), 662-678.

Rost, J. & Carstensen, C. H. (2002). Multidimensional Rasch measurement via item component models and faceted designs. *Applied Psychological Measurement*, *26* (1), 42-56.

Ruiz-Primo, M. A., Baxter, G. P. & Shavelson, R. J. (1993). On the stability of performance assessments. *Journal of Educational Measurement*, *30* (1), 41-53.

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.

Russell, B. (1905). On denoting. *Mind*, *14*, 479-493.

Samarapungavan, A. (1992). Children's judgements in theory choice tasks: Scientific rationality in childhood. *Cognition*, *45*, 1-32.

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.

Schauble, L., Glaser, R., Raghavan, K. & Reiner, M. (1991). Causal models and experimentation strategies in scientific reasoning. *The Journal of the Learning Sciences*, *1* (2), 201-238.

Schauble, L., Glaser, R., Raghavan, K. & Reiner, M. (1992). The integration of knowledge and experimentation strategies in understanding a physical systems. *Applied Cognitive Psychology*, *6* (4), 321-343.

Schauble, L., Klopfer, L. E. & Raghavan, K. (1991). Students' transition from an engineering model to a science model of experimentation. *Journal of Research in Science Teaching*, *28* (9), 859-882.

Schecker, H. & Parchmann, I. (2006). Modellierung naturwissenschaftlicher Kompetenz. *Zeitschrift für Didaktik der Naturwissenschaften*, *12*, 45-66.

Schecker, H. & Wiesner, H. (2007). Die Bildungsstandards Physik. *Praxis der Naturwissenschaften: Physik in der Schule*, *56* (6), 5-13.

Schmiemann, P. (2011). Fachsprache in biologischen Testaufgaben. *Zeitschrift für Didaktik der Naturwissenschaften*, *17*, 115-136.

Schreiber, N., Theyßen, H. & Schecker, H. (2009). Experimentelle Kompetenz messen?!

Physik und Didaktik in Schule und Hochschule, *8* (3), 92-101.

Senkbeil, M., Rost, J., Carstensen, C. H. & Walter, O. (2005). Der nationale Naturwissenschaftstest PISA 2003: Entwicklung und empirische Überprüfung eines zweidimensionalen Facettendesigns. *Empirische Pädagogik*, *19* (2), 166-189.

Shavelson, R. J., Baxter, G. P. & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement*, *30* (3), 215-232.

Shavelson, R. J., Baxter, G. P. & Pine, J. (1991). Performance assessment in science. *Applied Measurement in Education*, *4* (4), 347-362.

Shavelson, R. J., Baxter, G. P. & Pine, J. (1992). Performance assessments: Political rhetoric and measurment reality. *Educational Researcher*, *21* (4), 22-27.

Shavelson, R. J., Roeser, R. W., Kupermintz, H., Lau, S., Ayala, C., Haydel, A., ... Quihuis, G. (2002). Richard E. Snow's remaking of the concept of aptitude and multidimensional test validity: Introduction to the special issue. *Educational Assessment*, *8* (2), 77-99.

Shavelson, R. J. & Ruiz-Primo, M. A. (1999). Leistungsbewertung im naturwissenschaftlichen Unterricht. *Unterrichtswissenschaft*, *27* (2), 102-127.

Shavelson, R. J., Ruiz-Primo, M. A. & Wiley, E. W. (1999). Note on sources of sampling variability in science performance assessments. *Journal of Educational Measurement*, *36* (1), 61-71.

Shavelson, R. J., Solano-Flores, G. & Ruiz-Primo, M. A. (1998). Toward a science performance assessment technology. *Evaluation and Program Planning*, *21*, 171-184.

Siegler, R. S. & Liebert, R. M. (1975). Acquisition of formal scientific reasoning by 10- and 13-year-olds: Designing a factorial experiment. *Developmental Psychology*, *11* (3), 401-402.

Simon, H. A. (1977). *Models of discovery.* Dordrecht: D. Reidel.

Simon, S. A. & Jones, A. T. (1992). *Open work in science: A review of existing practice.* London: King's College.

Solano-Flores, G. (1994). *A logical model for the development of science performance assessments.* Unveröffentlichte Dissertation, University of California, Santa Barbara.

Solano-Flores, G., Jovanovic, J., Shavelson, R. J. & Bachman, M. (1999). On the development and evaluation of a shell for generating performance assessments. *International Journal of Science Education*, *21* (3), 293-315.

Solano-Flores, G. & Shavelson, R. J. (1997). Development of performance assessments in science: Conceptual, practical, and logistical issues. *Educational Measurement: Issues and Practice*, *16* (3), 16-25.

Solano-Flores, G., Shavelson, R. J., Ruiz-Primo, M. A., Schultz, S. E. & Wiley, E. W.

(1997). *On the development and scoring of classification and observation science performance assessments* (CSE Technical Report Nr. 458). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing, University of Los Angeles.

Stebler, R., Reusser, K. & Ramseier, E. (1997). Spitzenleistungen der Schweizer Siebtklässler im TIMSS-Experimentiertest. *Schweizer Lehrerinnen- und Lererzeitung*, *10*, 18-21.

Stebler, R., Reusser, K. & Ramseier, E. (1998). Praktische Anwendungsaufgaben zur integrierten Förderung formaler und materialer Kompetenzen: Erträge aus dem TIMMS-Experimentiertest. *Bildungsforschung und Bildungspraxis*, *20* (1), 28-53.

Stein, W. (1930). *Der Begriff des Schwerpunktes bei Archimedes*. Berlin: Julius Springer.

Szlovak, B. (2005). *HarmoS – Lehrplanvergleich Naturwissenschaften* (Bericht). Bern: Schweizerische Konferenz der kantonalen Erziehungsdirektoren (EDK).

Tannenbaum, R. S. (1971). The development of the test of science processes. *Journal of Research in Science Teaching*, *8* (2), 123-136.

Theyssen, H., von Aufschnaiter, S. & Schumacher, D. (2001). Kategoriengeleitete Analyse und Komplexitätsanalyse von Lernprozessen im Physikalischen Praktikum. In S. von Aufschnaiter & M. Wenzel (Hrsg.), *Nutzung von Videodaten zur Untersuchung von Lehr-Lern-Prozessen* (S. 101-114). Münster: Waxmann.

TIMSS Performance Assessment Coding Committee. (1994). *TIMSS main survey: Coding guide for performance assessment: Populations 1 and 2* (Bericht). Chestnut Hill: Boston College.

TIMSS: Beaton, A. E., Martin, M. O., Mullis, I. V. S., Gonzales, E. J., Smith, T. A. & Kelly, D. L. (1996). *Science achievement in the middle school years: IEA's Third International Mathematics and Science Study (TIMSS)*. Chestnut Hill: Boston College.

TIMSS: Harmon, M., Smith, T. A., Martin, M. O., Kelly, D. L., Beaton, A. E., Mullis, I. V. S., ... Orpwood, G. (1997). *Performance Assessment in IEA's Third International Mathematics and Science Study*. Chestnut Hill: Boston College.

TIMSS: Klieme, E. (2000a). Fachleistung im voruniversitären Mathematik- und Physikunterricht: Theoretische Grundlagen, Kompetenzstufen und Unterrichtsschwerpunkte. In J. Baumert, W. Bos & R. Lehmann (Hrsg.), *TIMSS/III: Dritte Internationale Mathematik und Naturwissenschaftsstudie. Band 2: Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn* (S. 57-128). Opladen: Leske und Budrich.

TIMSS: Klieme, E., Baumert, J., Köller, O. & Bos, W. (2000b). Mathematische und naturwissenschaftliche Grundbildung: Konzeptuelle Grundlagen und die Erfassung

und Skalierung von Kompetenzen. In J. Baumert, W. Bos & R. Lehmann (Hrsg.), *TIMSS/III: Dritte Internationale Mathematik- und Naturwissenschaftsstudie. Band 1: Mathematische und naturwissenschaftliche Grundbildung am Ende der Pflichtschulzeit* (S. 85-133). Leske + Budrich.

Toh, K.-A. & Woolnough, B. E. (1990). Assessing, through reporting, the outcomes of scientific investigations. *Educational Research*, *32* (1), 59-65.

Tschirgi, J. E. (1980). Sensible reasoning: A hypothesis about hypotheses. *Child Development*, *51*, 1-10.

Verband Deutscher Ingenieure. (2007). *Bildungsstandards Technik für den Mittleren Schulabschluss* (Bericht). Düsseldorf: Verband Deutscher Ingenieure. Zugriff auf `http://www.vdi-jutec.de/medienarchiv/ablage/original/bildungsstandards_2007.pdf` (Zugriff: 20.1.2012)

Vogt, P., Müller, A. & Kuhn, J. (2011). *Experimental competence rubrics: Conceptual synthesis and psychometric validation.* (Vortrag an der ESERA-Konferenz 2011 in Lyon)

von Aufschnaiter, C. & Rogge, C. (2010). Wie lassen sich Verläufe der Entwicklung von Kompetenz modellieren? *Zeitschrift für Didaktik der Naturwissenschaften*, *16*, 95-114.

von Aufschnaiter, S., von Aufschnaiter, C. & Schoster, A. (2000). Zur Dynamik von Bedeutungsentwicklung unterschiedlicher Schüler(innen) bei der Bearbeitung derselben Physik-Aufgaben. *Zeitschrift für Didaktik der Naturwissenschaften*, *6*, 37-57.

von Aufschnaiter, S. & Welzel, M. (1997). Wissensvermittlung durch Wissensentwicklung: Das Bremer Komplexitätsmodell zur quantitativen Beschreibung von Bedeutungsentwicklung und Lernen. *Zeitschrift für Didaktik der Naturwissenschaften*, *3* (2), 43-58.

Waddington, D., Nentwig, P. & Schanze, S. (Hrsg.). (2007). *Making it comparable: Standards in science education.* Münster: Waxmann.

Wainer, H., Bradlow, E. T. & Wang, X. (2007). *Testlet response theory and its applications.* Cambridge: Cambridge University Press.

Walpuski, M. (2010). Modelling scientific inquiry for large scale assessments. In G. Çakmakci & M. F. Taşar (Hrsg.), *Contemporary science education research: Learning and assessment* (S. 283-287). Istanbul: Pegem Akademi, ESERA.

Walpuski, M., Kampa, N., Kauertz, A. & Wellnitz, N. (2008). Evaluation der Bildungsstandards in den Naturwissenschaften. *Der mathematische und naturwissenschaftliche Unterricht*, *61* (6), 323-326.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, *12* (3), 129-140.

Weinert, F. E. (2001a). Concept of competence: A conceptual clarification. In D. S. Rychen & L. H. Salganik (Hrsg.), *Defining and selecting key competencies* (S. 45-65). Göttingen: Hogrefe & Huber.

Weinert, F. E. (2001b). Vergleichende Leistungsmessung in Schulen – Eine umstrittene Selbstverständlichkeit. In F. E. Weinert (Hrsg.), *Leistungsmessungen in Schulen* (S. 17-31). Weinheim: Beltz Verlag.

Wellenreuther, M. (2005). *Lehren und Lernen – aber wie? Empirisch-experimentelle Forschungen zum Lehren und Lernen im Unterricht.* Hohengehren: Schneider Verlag.

Wellnitz, N., Hartmann, S. & Mayer, J. (2010). Developing a paper-and-pencil-test to assess students' skills in scientific inquiry. In G. Çakmakci & M. Taşar (Hrsg.), *Contemporary science education research: Learning and assessment* (S. 289-294). Istanbul: Pegem Akademi, ESERA.

Wellnitz, N. & Mayer, J. (2008). Evaluation von Kompetenzstruktur und -niveaus zum Beobachten, Vergleichen, Ordnen und Experimentieren. *Erkenntnisweg Biologiedidaktik*, *7*, 129-143. Zugriff auf `http://www.biologie.fu-berlin.de/didaktik/Erkenntnisweg/2008/2008_09_Wellnitz.pdf` (Zugriff: 20.1.2012)

Woolnough, B. E. & Allsop, T. (1985). *Practical work in science.* Cambridge: Cambridge University Press.

Woolnough, B. E. & Toh, K.-A. (1990). Alternative approaches to assessment of practical work in science. *School Science Review*, *71* (256), 127-131.

Wright, B. D. & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8* (3), 370. Zugriff auf `http://www.rasch.org/rmt/rmt83b.htm` (Zugriff am 20.1.2012)

Wu, M. L. & Adams, R. J. (2007). *Applying the Rasch model to psycho-social measurement: A practical approach.* Melbourne: Educational Measurement Solutions. Zugriff auf `http://www.edmeasurement.com.au/_docs/RaschMeasurement_Complete.pdf` (Zugriff: 20.1.2012)

Wu, M. L., Adams, R. J., Wilson, M. R. & Haldane, S. A. (2007). ACER Conquest Version 2.0 [Software-Handbuch]. Camberwell: ACER Press.

Zberg, U. (preprint). *Experimentierkompetenz im tri-nationalen Vergleich (Arbeitstitel)* [Masterarbeit].

Zimmerman, C. & Glaser, R. (2001). *Testing positive versus negative claims: A preliminary investigation of the role of cover story on the assessment of experimental design skills* (CSE Technical Report Nr. 554). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing, University of Pittsburgh.

Dank

Ohne die Vorarbeit, Unterstützung und Beratung vieler Personen hätte diese Dissertation nicht geschrieben werden können. Ihnen allen gilt mein herzlicher Dank.

Insbesondere danke ich Peter Labudde für die Betreuung der Dissertation, die wertvollen Rückmeldungen und das Vertrauen in meine Arbeit ganz allgemein. Ich danke für die wissenschaftliche Förderung, von der ich in den vergangenen viereinhalb Jahren reichlich profitieren durfte, und für die familienfreundlichen Arbeitsbedingungen, die mir als Vater zweier kleiner Buben das Forschen ermöglichten.

Ich danke den Kolleginnen, Kollegen und den Assistierenden des Konsortiums HarmoS Naturwissenschaften für ihre Arbeit, auf der die Dissertation aufbaut. Insbesondere danke ich Marco Adamina, Johannes Börlin, François Gingin, Susanne Metzger, Reto Müller, Kathleen Raths, Armin Rempfler, Markus Vetterli und Urs Wagner für die bereichernde Zusammenarbeit bei der Testentwicklung. Ohne ihr Engagement hätte der HarmoS-Experimentiertests nicht stattgefunden. Marco Adamina und Peter Labudde, den Co-Leitern des Konsortiums, danke ich für das Vertrauen, mir die Daten des Experimentiertests und des Fragebogens – ohne Wenn und Aber – zur Verfügung zu stellen.

Ich danke Erich Ramseier und Kathleen Raths für die Vorarbeiten bei der Datenbearbeitung. Erich Ramseier danke ich für die Beratung bei Fragen zu Rasch und ConQuest.

Ich danke Johannes Börlin für die gute Kollegenschaft am Zentrum für Naturwissenschafts- und Technikdidaktik, den wissenschaftlichen Austausch und die Hilfe bei LaTeX-Problemen.

Ich danke Christoph Bruder für die Begleitung der Dissertation als Fakultätsverantwortlicher der Universtät Basel und Horst Schecker für die Begutachtung der Dissertation.

Ganz besonders danke ich meiner Frau Jutta Glanzmann. Für die Partnerschaft, die stete Unterstützung und Begleitung. Für die Bereitschaft, mir mit zusätzlicher Familienarbeit den Rücken frei zu halten. Für das Korrekturlesen letztlich, das zwischen Familie und eigenem Job auch noch Platz fand.

Selbständigkeitserklärung

Die vorliegende Studie baut auf Arbeiten auf, die im Rahmen des Projekts HarmoS Naturwissenschaften entstanden sind. Dies betrifft die Entwicklung, Durchführung, Kodierung sowie die erste Aufbereitung der Daten des HarmoS-Experimentiertests und des Fragebogens. An der Aufgabenentwicklung und Herstellung der Experimentiersets beteiligt waren neben mir massgeblich Marco Adamina, Johannes Börlin, François Gingins, Peter Labudde, Susanne Metzger, Reto Müller, Kathleen Raths, Armin Rempfler, Markus Vetterli und Urs Wagner. Die Schulbesuche für die Pilotierung und für die Haupttests wurden von den oben genannten Konsortiumsmitgliedern (mit Ausnahme von Armin Rempfler) unter Mithilfe von Hilfsassistierenden der Pädagogischen Hochschulen in Zürich, Bern und Lausanne durchgeführt. Die Pilotierung der Aufgaben für das 9. Schuljahr sowie die Tests in der Region Bern für das 9. Schuljahr wurden von mir selbst organisiert. Die Testbögen wurden in zwei Gruppen kodiert, eine unter der Leitung von Marco Adamina und Kathleen Raths an der PHBern sowie eine unter meiner Leitung an der PH FHNW in Basel. Die Daten der zwei Kodiergruppen wurden von Birte Knierim zusammengetragen und von Erich Ramseier in einer ersten Itemauswertung angepasst und in einem SPSS-Dokument abgelegt, das ich als Datenbasis für meine Analysen übernahm. Der Fragebogen wurde von Marco Adamina entwickelt. Kathleen Raths trug die kodierten Daten in einem SPSS-Dokument zusammen, das ebenfalls von mir unverändert übernommen wurde. Darüber hinaus habe ich keine weiteren fremden Arbeiten übernommen oder substantielle Hilfe Dritter in Anspruch genommen. Dies gilt sowohl für die konzeptuelle Arbeit als auch für die quantitativen Auswertungen. Die Kodierungen, die für die verschiedenen Analysen notwendig waren, wurden von mir alleine durchgeführt.

26. März 2012
Christoph Gut-Glanzmann

Curriculum vitae

Personalia

Christoph Gut-Glanzmann
Tramstrasse 88
CH-8050 Zürich

geboren am 22. Juni 1970
Bürger von Affoltern am Albis, Zürich
verheiratet, Vater zweier Söhne

Ausbildung

1996 - 1998	Universität Zürich, Höheres Lehramt mit Lehrdiplom in Physik
1990 - 1996	Eidgenössische Technische Hochschule Zürich, Studium der Physik, Diplom in theoretischer Festkörperphysik bei Prof. Dr. T. Maurice Rice
1983 - 1989	Kantonsschule Wiedikon Zürich, Matura Typus B

Lehr- und Forschungstätikeit

2011 - dato	Dozent für Didaktik der Physik, PH Zürich
2009 - 2011	wissenschaftlicher Mitarbeiter, PH Zürich (Abteilung Sekundarstufe I)
2007 - 2011	wissenschaftlicher Mitarbeiter, Doktorand von Prof. Dr. Peter Labudde, PHBern (Institut Sekundarstufe II), PH FHNW (Zentrum für Naturwissenschafts- und Technikdidaktik)
2002 - 2007	Hauptlehrer für Physik und Mathematik, Kantonsschule Olten
1996 - 2002	Lehrbeauftragter für Physik, Mathematik und Informatik, Kantonsschulen Frauenfeld und Olten