Studien zum Physik- und Chemielernen

H. Niedderer, H. Fischler, E. Sumfleth [Hrsg.]

210

Martin Dickmann

Messung von Experimentierfähigkeiten

Validierungsstudien zur Qualität eines computerbasierten Testverfahrens

λογος

Studien zum Physik- und Chemielernen

Herausgegeben von Hans Niedderer, Helmut Fischler und Elke Sumfleth

Diese Reihe im Logos-Verlag bietet ein Forum zur Veröffentlichung von wissenschaftlichen Studien zum Physik- und Chemielernen. In ihr werden Ergebnisse empirischer Untersuchungen zum Physik- und Chemielernen dargestellt, z. B. über Schülervorstellungen, Lehr-/Lernprozesse in Schule und Hochschule oder Evaluationsstudien. Von Bedeutung sind auch Arbeiten über Motivation und Einstellungen sowie Interessensgebiete im Physik- und Chemieunterricht. Die Reihe fühlt sich damit der Tradition der empirisch orientierten Forschung in den Fachdidaktiken verpflichtet. Die Herausgeber hoffen, durch die Herausgabe von Studien hoher Qualität einen Beitrag zur weiteren Stabilisierung der physik- und chemiedidaktischen Forschung und zur Förderung eines an den Ergebnissen fachdidaktischer Forschung orientierten Unterrichts in den beiden Fächern zu leisten.

Hans Niedderer Helmut Fischler Elke Sumfleth

Messung von Experimentierfähigkeiten

Validierungsstudien zur Qualität eines computerbasierten Testverfahrens

Martin Dickmann

Bibliografische Information der Deutschen Nationalbibliothek

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

ISBN 978-3-8325-4356-3

Logos Verlag Berlin GmbH
Comeniushof, Gubener Str. 47,
10243 Berlin
Tel.: +49 (0)30 42 85 10 90
Fax: +49 (0)30 42 85 10 92
INTERNET: http://www.logos-verlag.de

Messung von Experimentierfähigkeiten

Validierungsstudien zur Qualität eines computerbasierten Testverfahrens

Von der Fakultät für Physik der
Universität Duisburg-Essen genehmigte
Dissertation zur Erlangung des
Doktorgrades der Naturphilosophie
Dr. phil. nat.

von Martin Dickmann aus Dinslaken

1. Gutachterin:	Prof. Dr. Heike Theyßen
2. Gutachter:	Prof. Dr. Horst Schecker
3. Gutachter:	Prof. Dr. Roger Erb
Eingereicht am:	08.04.16
Tag der Disputation:	13.07.16

Teile dieser Arbeit sind bereits veröffentlicht in:

Dickmann, M. & Theyßen, H. (2013). Curriculare Validität von Units zur Messung experimenteller Kompetenz. In S. Bernholt (Hrsg.), *Inquiry-based Learning - Forschendes Lernen: Gesellschaft für Didaktik der Chemie und Physik, Jahrestagung in Hannover 2012*. Kiel: IPN, 587-589.

Dickmann, M., Eickhorst, B., Theyßen, H., Neumann, K., Schecker, H. & Schreiber. N. (2014). Measuring experimental skills in large-scale assessments: developing a simulation-based test instrument. In C. P. Constantinou, N. Papadouris & A. Hadjigeorgiou (Eds.), *Science Education Research For Evidence-based Teaching and Coherence in Learning*. *E-Book Proceedings of the ESERA 2013 Conference,* Part 11 (co-ed. Millar, R., Dolin, J.), 1993-2001. Nicosia, Cyprus: European Science Education Research Association.

Theyßen, H., Schecker, H., Neumann, K., Dickmann, M. & Eickhorst, B. (2013). Messung experimenteller Kompetenz in Large-Scale Assessments. In S. Bernholt (Hrsg.), *Inquiry-based Learning - Forschendes Lernen: Gesellschaft für Didaktik der Chemie und Physik, Jahrestagung in Hannover 2012*. Kiel: IPN, 596-598.

Dickmann, M., Eickhorst, B., Theyßen, H., Schecker, H. & Neumann, K. (2015). Testinstrument für experimentelle Kompetenz: Einfluss des Testformats auf konstruktbezogene Denkprozesse. In S. Bernholt (Hrsg.), *Heterogenität und Diversität - Vielfalt der Voraussetzungen im naturwissenschaftlichen Unterricht. Gesellschaft für Didaktik der Chemie und Physik. Jahrestagung in Bremen 2014*. Kiel: IPN, 663-665.

Eickhorst, B., Dickmann, M., Schecker, H., Theyßen, H. & Neumann, K. (2015). Messung experimenteller Kompetenz im Large-Scale: Bewertung experimenteller Aufgaben. In S. Bernholt (Hrsg.), *Heterogenität und Diversität - Vielfalt der Voraussetzungen im naturwissenschaftlichen Unterricht. Gesellschaft für Didaktik der Chemie und Physik. Jahrestagung in Bremen 2014*. Kiel: IPN, 169-171.

Theyßen, H., Schecker, H., Dickmann, M., Eickhorst, B. und Neumann, K. (2016a). Messung experimenteller Kompetenz in Large-Scale-Assessments (MeK-LSA). In Bundesministerium für Bildung und Forschung (Hrsg.), *Forschungsvorhaben in Ankopplung an Large-Scale-Assessments.* Berlin: Bundesministerium für Bildung und Forschung, Bildungsforschung Band 44, 83-96.

Theyßen, H., Schecker, H., Neumann, K., Eickhorst, B. & Dickmann, M. (2016b). Messung experimenteller Kompetenz – ein computergestützter Experimentiertest. *Physik und Didaktik in Schule und Hochschule*, *15 (*1).

Zusammenfassung

Im Projekt MeK-LSA wurde ein computerbasiertes Testverfahren (MeK-LSA Experimentiertest) für den Einsatz in Large-Scale Assessments entwickelt, das experimentelle Kompetenz in den Bereichen Planung, Durchführung und Auswertung in einer tatsächlichen experimentellen Handlungssituation effizient, reliabel und valide erfassen soll. Eine Grundvoraussetzung für die Validität des gesamten Testverfahrens ist, dass bereits die Ergebnisse der Aufgabenbearbeitungen und die daraus berechneten Testwerte (als Maß für die Lösungsqualität) valide als Ausdruck von Experimentierfähigkeiten interpretiert werden können. Um dies zu erreichen, wurde der gesamte Prozess der Testentwicklung bis zur Zuweisung der Testwerte im Projekt MeK-LSA durch umfangreiche Validierungsstudien begleitet.

Ziel der vorliegenden Dissertation ist es, die Ergebnisse der einzelnen Validierungsstudien zu einer kohärenten Argumentation zusammenzuführen, die schlüssig belegt, bis zu welchem Grad die Testwerte valide als Ausdruck von Experimentierfähigkeiten interpretiert werden können. Dazu wird ein Interpretations-Nutzungs-Argument formuliert und theorie- und evidenzbasiert bewertet. Durch diese Vorgehensweise wird ein Prozess transparent dokumentiert, der bisher in der Praxis selten systematisch erfolgt und über den in der Folge – wenn überhaupt – nur bruchstückhaft berichtet wird.

In einem ersten Validierungsschritt wird die Relevanz und Repräsentativität der Testinhalte und Testanforderungen mit Experten- und Schülerbefragungen geprüft. Darauf aufbauend wird untersucht, ob die Aufgaben angemessen in experimentelle Performanz übersetzt worden sind. Eine der zentralen Fragen ist dabei, in welchem Maße Schülerinnen und Schüler bei der Bearbeitung der Aufgaben am Computerbildschirm (on-screen) experimentelle Überlegungen anstellen. Zur Klärung dieser Frage wurde eine Think-Aloud Studie durchgeführt. Ein weiterer wichtiger Validierungsschritt besteht in der Überprüfung, ob Bearbeitungen von on-screen Aufgaben ähnliche Testwerte zugewiesen werden wie Bearbeitungen von inhaltlich identischen Aufgaben mit Realexperimenten (hands-on).

Aus dem Blickwinkel der Validität zeigen die Ergebnisse ein positives Bild des MeK-LSA Experimentiertests. So decken die Aufgaben beispielsweise relevante und repräsentative Inhalte und Anforderungen schulischen Experimentierens aus dem Physikunterricht der Sekundarstufe I ab. Darüber hinaus stellen die Schülerinnen und Schüler bei der Bearbeitung der on-screen Aufgaben überwiegend experimentbezogene Überlegungen an. Beim Vergleich der Testformate (hands-on vs. on-screen) konnte gezeigt werden, dass bei der Bearbeitung von on-screen gestellten Aufgaben in der Regel ähnliche Leistungen erzielt werden wie bei der Bearbeitung inhaltlich identischer hands-on Aufgaben. Die Aussagefähigkeit der Befunde ist allerdings durch die zugrundeliegende Kompetenzmodellierung beschränkt. Dennoch leistet der MeK-LSA Experimentiertest einen Beitrag zum Assessment-Diskurs experimenteller Kompetenz. Mit dem MeK-LSA Experimentiertest liegt nun ein Test vor, der das Spektrum der in Large-Scale Assessments erfassbaren Experimentierfähigkeiten erweitert.

Summary

In the MeK-LSA project a computer-based test instrument (*MeK-LSA Experimentiertest*) for the use in large-scale assessments was developed that can efficiently, reliably and validly assess experimental skills in the areas of planning, performing and analyzing experiments in an actual experimental operational situation. A basic prerequisite for the validity of the entire test instrument is that already the results of the tasks and the calculated test values can be validly interpreted as an expression of experimental skills (as a measure of solution quality). In order to ensure measurement validity, the entire process of test development to the assignment of test values was accompanied by multiple validation studies in the MeK-LSA project.

The aim of this PhD-thesis is to merge the results of these validation studies into a coherent argument, that is conclusive, to the degree to which the test values can be validly interpreted as an expression of experimental skills. To achieve this aim an interpretation-use-argument is formulated and evaluated theory- and evidence-based. By doing so, a process is documented transparently, that is so far in practice rarely done systematically and subsequently, if at all, reported fragmentary.

In a first validation step, the relevance and representativeness of test content and test requirements is checked by expert and student surveys. Based on this study it will be examined if the tasks have been adequately translated into experimental performance. One of the key questions is the extent to which students do experimental considerations while working on the tasks in front of the computer screen (on-screen). To clarify this question a think-aloud study was carried out. Another important validation step is to consider if similar test values are assigned to on-screen tasks and hands-on tasks with the same content.

The results show - from the perspective of validity - a positive image of the *MeK-LSA Experimentiertest*. The tasks cover, for example, relevant and representative content and requirements of curricular experimentation from physics classes of secondary education. In addition, the students do predominantly experiment related considerations while working on the on-screen tasks. When comparing the test formats (hands-on vs. on-screen) it was shown that persons working on the on-screen tasks generally achieve similar performance as well as working on hands-on tasks with identical content. The validity of the findings, however, is limited by the underlying competence modeling. Nevertheless, the *MeK-LSA Experimentiertest* contributes to the assessment discourse of experimental skills. The *MeK-LSA Experimentiertest* is a test, which extends the range of measurable experimental skills in large-scale assessments.

INHALTSVERZEICHNIS

1 Einleitung

Ein zentrales Bildungsziel des naturwissenschaftlichen Unterrichts ist der Erwerb von Experimentierfähigkeiten. Obwohl über dieses Bildungsziel national und international ein breiter Konsens besteht, werden Experimentierfähigkeiten in zentralen Prüfungen und im Bildungsmonitoring bisher kaum bzw. nur unvollständig überprüft. Das hat Auswirkungen auf die Bedeutung, die diesem Bildungsziel in der Unterrichtspraxis beigemessen wird. Um das Erreichen des Bildungsziels *„wirksamer lenken zu können“* (Klieme, Maag-Merki & Hartig, 2007, S. 8), muss die Ausprägung von Experimentierfähigkeiten auch in großflächigen Erhebungen (Large-Scale Assessments) zum Bildungsmonitoring vollständig überprüfbar sein (vgl. Abschnitt 1.1). Eine vollständige Überprüfung von Experimentierfähigkeiten in Large-Scale Assessments ist bisher jedoch nicht möglich, da Testverfahren fehlen, die Experimentierfähigkeiten zur Durchführung von Experimenten effizient, reliabel und valide messen können (vgl. Abschnitt 1.2). In der vorliegenden Arbeit werden die Entwicklung und eine Argumentation zur Validitätsbewertung eines Experimentiertests beschrieben, der diese Lücke schließen soll.

1.1 Ausgangslage

Die *„ungünstigen Ergebnisse“* (Pant, Stanat, Pöhlmann & Böhme, 2013, S. 13) deutscher Schülerinnen und Schüler in den ersten *TIMS*- und *PISA-Studien* (vgl. Baumert et al. 1997 & Artelt et al., 2001) haben zu weitreichenden Änderungen in der administrativen Steuerung des deutschen Bildungswesens geführt (vgl. Breakspear, 2012, S. 14). Diese Veränderungen wurden durch eine neue Strategie zum Bildungsmonitoring der Kultusministerkonferenz der Länder (KMK) eingeleitet (Zeitler, Asbrand & Heller, 2013, S. 127). Zentrale Merkmale der neuen Strategie sind die 2004 verabschiedeten output- und kompetenzorientierten Bildungsstandards für den mittleren Schulabschluss, die Entwicklung von Testverfahren zur Überprüfung der in den Standards formulierten Kompetenzerwartungen, sowie die Entwicklung einer Gesamtstrategie zur Qualitätssicherung des deutschen Schulsystems (Köller & Schöps, 2013, S. 72-73). Nach Pant, Böhme und Köller (2013) werden in den Bildungsstandards *„Kompetenzerwartungen durchgängig als Beschreibungen konkreter Fähigkeiten formuliert (im Sinne von Can-do-Statements) […]“* (S. 53). Als Ausgangspunkt für die *Kompetenzmodellierung* hat sich im deutschsprachigen Raum die Kompetenzdefinition von Klieme und Leutner (2006) etabliert, wonach Kompetenzen als *„kontextspezifische kognitive Leistungsdispositionen, die sich funktional auf Situationen und Anforderungen in einer bestimmten Domäne beziehen“* (S. 879), zu verstehen sind. Dieses Kompetenzverständnis verweist auf einen funktionalen Handlungsbezug sowie die Erlern- und Förderbarkeit von Kompetenzen (Pant et al., 2013, S. 54) und stimmt in wesentlichen Teilen mit dem Kompetenzverständnis nationaler und internationaler Schulleistungsstudien überein (Wendt & Bos, 2011, S. 16). Auch für die naturwissenschaftlichen Unterrichtsfächer (Biologie, Chemie und Physik) sind Bildungsstandards für den mittleren Schulabschluss (KMK, 2005a, 2005b, 2005c) verabschiedet worden. In diesen sind normativ Kompetenzerwartungen formuliert, die Schülerinnen und Schüler in der Regel in den vier Kompetenzbereichen *Umgang mit Fachwissen*, *Erkenntnisgewinnung*, *Kommunikation* und

Bewertung, am Ende der Sekundarstufe I erfüllen sollten. Im Kompetenzbereich *Erkenntnisgewinnung* nehmen Kompetenzerwartungen zum Experimentieren als eine Methode naturwissenschaftlicher Erkenntnisgewinnung einen hohen Stellenwert ein. Im Zusammenhang mit naturwissenschaftlichen Fragestellungen sollen Schülerinnen und Schüler u.a. dazu befähigt werden, experimentelle Untersuchungen selbstständig zu planen, durchzuführen und auszuwerten. Die Bedeutung des Erwerbs von Experimentierfähigkeiten wird auch in internationalen Science-Curricula (u.a. USA: NRC, 1996, 2012; England: DfEE, 1999, DfE 2013) betont. Für ein erfolgreiches Bildungsmonitoring muss die Ausprägung dieser Experimentierfähigkeiten auch in Large-Scale Assessments vollständig, reliabel, valide und effizient messbar sein. Das gilt gleichermaßen für nationales und internationales Bildungsmonitoring.

1.2 Messung von Experimentierfähigkeiten in Large-Scale Assessments - ein Problemaufriss

In bisherigen Large-Scale Assessments werden Experimentierfähigkeiten aus vorwiegend testökonomischen Gründen mit rein schriftlichen Testverfahren erfasst (bis auf wenige Ausnahmen, z. B. Schweiz: *HarmoS*, vgl. Adamina et al., 2009). Rein schriftliche Testverfahren können aber nur Experimentierfähigkeiten erfassen *„die streng genommen alle am Schreibtisch“* (Höttecke & Rieß, 2015, S. 136) gezeigt werden können, wie zum Beispiel das Ziehen von Schlussfolgerungen aus experimentellen Daten. Fähigkeiten zur Durchführung experimenteller Untersuchungen, wie zum Beispiel das eigenständige Aufbauen einer Versuchsanordnung, können mit diesen Verfahren nicht erfasst werden (vgl. IQB, 2013, S. 16; Meier & Mayer, 2012, S. 95). Diese Fähigkeiten sind jedoch ein wichtiger Bestandteil experimenteller Kompetenz. Das schränkt die Validität bisheriger Messungen experimenteller Kompetenz in Large-Scale Assessments ein, da mit ihnen der Anspruch verbunden wird, die inhaltlich und theoretisch bedeutsamen Facetten einer Kompetenz abbilden zu können (Pellegrino, Chudowsky & Glaser, 2001, S. 13). Im Hinblick auf die Erfassung von Fähigkeiten zur Durchführung experimenteller Untersuchungen wird Experimentiertests (oder zumindest Tests mit praktischen Anteilen) aufgrund ihres authentischen Charakters nicht selten eine *verführerische Augenscheinvalidität* (Aschbacher, 1991, S. 277) unterstellt. Ob Experimentiertests im Vergleich zu rein schriftlichen Testverfahren tatsächlich eine validere Messung von Experimentierfähigkeiten ermöglichen, ist evidenzbasiert weiter abzusichern. Aus messmethodischer Sicht bestehen bei Experimentiertests nach wie vor Probleme, die noch nicht zufriedenstellend gelöst werden konnten. Der Zeit- und Kostenaufwand für Experimentiertests liegt (noch) deutlich über dem Aufwand für rein schriftliche Kompetenztests (vgl. Stecher & Klein, 1997, S. 11). Die wahre Ausprägung der Kompetenz kann in Experimentiertests häufig nicht ausreichend reliabel geschätzt werden, da die gezeigte und bewertete Performanz durch eine Vielzahl von Faktoren beeinflusst wird (u.a. Shavelson, Baxter & Gao, 1993, S. 217; Übersicht: Schreiber, Theyßen & Schecker, 2014, S. 163). Mit Blick auf Fragen zur Validität der Messung ergeben sich in Anlehnung an Messick (1996, S. 5) für Experimentiertests zwei wesentliche Einschränkungen: die Erzeugung kompetenzirrelevanter Varianz (z. B. Verständnisanforderungen) und die unvollständige Abdeckung der Facetten eines Fähigkeitskonstrukts (Unterrepräsentation des eigentlichen Kompetenzumfangs). Diese Einschränkungen können allerdings auch auf stärker

standardisierte (schriftliche) Testverfahren zutreffen und sind kein Alleinstellungsmerkmal von Experimentiertests. Bezogen auf die gesamte Kompetenzforschung stellen Wendt und Bos (2011, S. 17) in Frage, ob eine valide Erfassung von Kompetenzen durch ein einzelnes Verfahren überhaupt sinnvoll ist, oder ob im Hinblick auf eine größere Varianzaufklärung eine Triangulation verschiedener Testverfahren vorzuziehen ist. Eine Triangulation setzt allerdings voraus, dass mehrere geeignete Testverfahren für alle wesentlichen Aspekte einer Kompetenz verfügbar sind. Sind geeignete Testverfahren verfügbar, kann ein theorie- und evidenzbasierter Diskurs initiiert werden, der im Kern auf die angemessene Erfassung der interessierenden Kompetenz fokussiert und nicht durch pragmatische und ökonomische Entscheidungen dominiert wird. Das Ziel muss es daher sein, sowohl die Testverfahren als auch die Kompetenzmodellierung kontinuierlich weiterzuentwickeln (Shavelson, 2010, S. 62). Die (Weiter-)Entwicklung von Testverfahren zur Kompetenzerfassung lässt sich im Assessment-Diskurs verorten (vgl. Gut, 2012, S. 20-23). Das Projekt *Messung experimenteller Kompetenz in Large-Scale Assessments* (kurz: MeK-LSA[1]), in dem die vorliegende Arbeit angesiedelt ist, soll einen Beitrag zum Assessment-Diskurs experimenteller Kompetenz (vgl. ebenda, S. 39-55) leisten.

1.3 Projekt MeK-LSA - Kurzdarstellung

Ziel des Projekts MeK-LSA (vgl. Theyßen, Schecker, Neumann, Dickmann & Eickhorst, 2013, S. 596) ist die Entwicklung eines Testverfahrens …

> … das *experimentelle Kompetenz* in den Bereichen Planung, Durchführung und Auswertung in einer tatsächlichen experimentellen Handlungssituation *reliabel* und *valide* erfasst
>
> … und gleichzeitig *effizient* genug für den Einsatz in *Large-Scale Assessments* ist.

Zielgruppe des Testverfahrens sind Schülerinnen und Schüler am *Ende der Sekundarstufe I*.

Das Projekt MeK-LSA definiert *experimentelle Kompetenz* auf folgende Weise:

> *„Experimentelle Kompetenz ist eine latente Fähigkeit zur mindestens intuitiv regelbasierten Planung und Durchführung von Versuchen, die der Klärung einer physikalischen Fragestellung dienen, sowie zur methodisch bewussten Auswertung der damit gewonnen Daten. In Performanz realisiert sich experimentelle Kompetenz, wenn zusätzlich das für den Themenbereich notwendige Fachwissen vorhanden i*st“ (Eickhorst, Dickmann, Schecker, Theyßen & Neumann, 2015, S. 169).

Die Definition basiert auf vorliegender Literatur zu experimenteller Kompetenz. Bezüglich der inneren Struktur experimenteller Kompetenz folgt die Definition dem weit verbreiteten Teilprozessansatz, der den Prozess des Experimentierens idealtypisch als lineare Abfolge von

[1] Die vorliegende Dissertation ist im Rahmen des vom Bundesministerium für Bildung und Forschung (FKZ LSA005) geförderten Verbundprojekts „Messung experimenteller Kompetenz in Large-Scale Assessments“ (MeK-LSA) der Universitäten Duisburg-Essen und Bremen sowie des IPN Kiel entstanden.

Teilprozessen beschreibt (vgl. Gut, Hild, Metzger & Tardent, 2014a, S. 171; Abschnitt 2.1). Nach Emden und Sumfleth (2012, S. 69) lassen sich die Teilprozesse in drei übergeordnete Bereiche gliedern: Ideen- bzw. Hypothesenfindung, Durchführung und Auswertung. Von Aufschnaiter und Rogge (2010, S. 106) betonen, dass sich aus rein explorativem Handeln noch keine Kompetenz ableiten lässt. Diese Auffassung wird in der Definition aufgenommen, indem auf Ebene der erwarteten kognitiven Prozesse eine mindestens intuitiv regelbasierte Vorgehensweise gefordert wird. Darüber hinaus wird in der Definition angenommen, dass *experimentelle Kompetenz* ein eigenständiges Konstrukt ist, *Fachwissen* (Gott & Duggan, 2002) und *kognitive Fähigkeiten* aber einen Einfluss auf die Fähigkeit zum Zeigen experimenteller Performanz haben. Auch die Ergebnisse einer von Erb, Neumann und Härtig (2015) durchgeführten Expertenbefragung (N=74 Fachdidaktikerinnen und Fachdidaktiker sowie Psychologinnen und Psychologen) zur übergeordneten Frage *Welche Fähigkeiten sind für Teilschritte des Experimentierens wichtig?* zeigen, dass Fachwissen von zentraler Bedeutung für erfolgreiches Experimentieren ist.

Die so definierte Kompetenz soll mit dem MeK-LSA Experimentiertest (Theyßen, Schecker, Neumann, Eickhorst & Dickmann, 2016b) in Large-Scale Assessments gemessen werden. Der MeK-LSA Experimentiertest ist gekennzeichnet durch curricular valide experimentelle Aufgabenstellungen, ein konsekutives Aufgabenformat[2] und ein darauf abgestimmtes Auswertungsverfahren für die Schülerlösungen aller Teilschritte. Die Aufgaben sind vollständig am Computerbildschirm (on-screen) zu bearbeiten. Neben on-screen zu bearbeitenden schriftlichen Aufgaben kommen Aufgaben mit interaktiven Elementen (Simulationen, Zeichentools etc.) zur Planung und Durchführung von Versuchen sowie zur Auswertung der damit gewonnen Daten zum Einsatz.

1.4 Ziel und Ertrag der vorliegenden Arbeit

Im Rahmen einer Large-Scale Studie mit 1194 Schülerinnen und Schülern in vier Bundesländern wurde der MeK-LSA Experimentiertest eingesetzt (Theyßen et al., 2016b). Die damit gewonnenen Daten bilden die Grundlage für eine Skalierung des Tests (ebenda), ein Standardsetting zur Beschreibung von Stufen experimenteller Kompetenz (Schecker, Neumann, Theyßen, Eickhorst & Dickmann, im Druck) sowie Untersuchungen zur Struktur und Stabilität experimenteller Kompetenz (Eickhorst, in Vorbereitung).

Die Grundvoraussetzung für all diese Analysen besteht darin, dass bereits die Ergebnisse der Aufgabenbearbeitungen und die daraus berechneten Testwerte (als Maß für die Lösungsqualität) valide als Ausdruck von Experimentierfähigkeiten interpretiert werden können. Der gesamte Prozess der Testentwicklung bis zur Zuweisung der Testwerte wurde im Projekt MeK-LSA daher durch umfangreiche Validierungsstudien begleitet. Über die Methodik und die Ergebnisse dieser Studien wird in der vorliegenden Arbeit berichtet.

[2] konsekutives Aufgabenformat: Bearbeitungsleitfaden zur Abfolge von Teilschritten ist vorgegeben.

Ziel der vorliegenden Arbeit ist es, …

> … die Ergebnisse der einzelnen Validierungsstudien zu einer kohärenten Argumentation zusammenzuführen,
>
> … die schlüssig belegt, bis zu welchem Grad die Testwerte valide als Ausdruck von Experimentierfähigkeiten interpretiert werden können.

Validität gilt als das wichtigste Qualitätsmerkmal eines Testverfahrens (z. B. Bortz & Döring, 2006, S. 200). Auf theoretischer Ebene wird das Konzept Validität entsprechend differenziert und umfangreich beschrieben (vgl. Kapitel 5). Bei der praktischen Anwendung wird die Überprüfung der Validität allerdings häufig auf statistisch zu beschreibende Merkmale eines Testverfahrens reduziert (Leuders, 2014, S. 12). An dieser Vorgehensweise wird kritisiert, dass wesentliche Validierungsschritte häufig unberücksichtigt bleiben und somit Sprünge in der Validitätsargumentation entstehen.

> "*One of the essential leaps in the assessment process is from performance to [...] scoring the underlying attribute from what students do [...]. Theoretically, this leap is problematic but most approaches to the use of test data fail to problematise it"* (Matters, 2009, S. 210).

Um Sprünge in der Validitätsargumentation zu vermeiden, sollte bei der Überprüfung der Validität die Frage im Vordergrund stehen, welche Merkmale berücksichtigt werden müssen, um eine kohärente und schlüssige Argumentation für bzw. gegen die beabsichtigte Interpretation der Testwerte aufbauen zu können. Das schließt explizit auch Merkmale ein, die auf qualitativen Urteilen beruhen (Leuders, 2014, S. 12). Dieser Grundidee folgend wird in der vorliegenden Dissertation für den MeK-LSA Experimentiertest der Prozess der Testentwicklung bis zur Zuweisung von Testwerten systematisch validiert. Durch diese Vorgehensweise wird ein Prozess transparent dokumentiert, der in der Praxis bisher selten systematisch erfolgt und über den in der Folge – wenn überhaupt – nur bruchstückhaft berichtet wird.

1.5 Aufbau der Dissertation

Die Dissertation gliedert sich in zwei Teile (vgl. Abbildung 1.1): einen Teil zur Vorstellung und Begründung des MeK-LSA Experimentiertests (Teil I) und einen Teil mit empirischen Studien zur Validität des MeK-LSA Experimentiertests (Teil II).

In Teil I bildet die Beschreibung der theoretischen Grundlagen zur Messung experimenteller Kompetenz (Kapitel 2) die Basis für die Vorstellung und Begründung des – durch das MeK-LSA Projektteam – entwickelten Testverfahrens (Kapitel 3). Ein zentraler Schritt bei der Entwicklung des MeK-LSA Experimentiertests war die Überführung der Testkonzeption in konkrete Testaufgaben (Kapitel 4). Die für diesen Schritt erforderlichen empirischen Grundlagen sind vom Autor dieser Dissertation erarbeitet worden.

Teil II bildet den Kern der vorliegenden Dissertation. In diesem Teil wird der MeK-LSA Experimentiertest aus dem Blickwinkel der Validität bewertet. Die theoretischen Grundlagen zur Validität von Testverfahren (Kapitel 5) zeigen das der Dissertation zugrundeliegende Verständnis von Validität. Daran anknüpfend wird ein Bezugsrahmen für die Validitätsbewertung des MeK-LSA Experimentiertests beschrieben, der die notwendigen Validierungsschritte systematisiert und gleichzeitig einen konsistenten Rahmen für den Bewertungsprozess schafft. Der Bezugsrahmen bildet das zentrale Element zur weiteren Strukturierung der Dissertation (vgl. Kapitel 6). Ausgehend vom Bezugsrahmen wird theorie- und evidenzbasiert bewertet, bis zu welchem Grad die Testwerte des MeK-LSA Experimentiertests valide als Ausdruck von Experimentierfähigkeiten aufgefasst werden können (Kapitel 7 bis 17). Die Arbeit endet mit einer Schlussbemerkung und einem Ausblick (Kapitel 18).

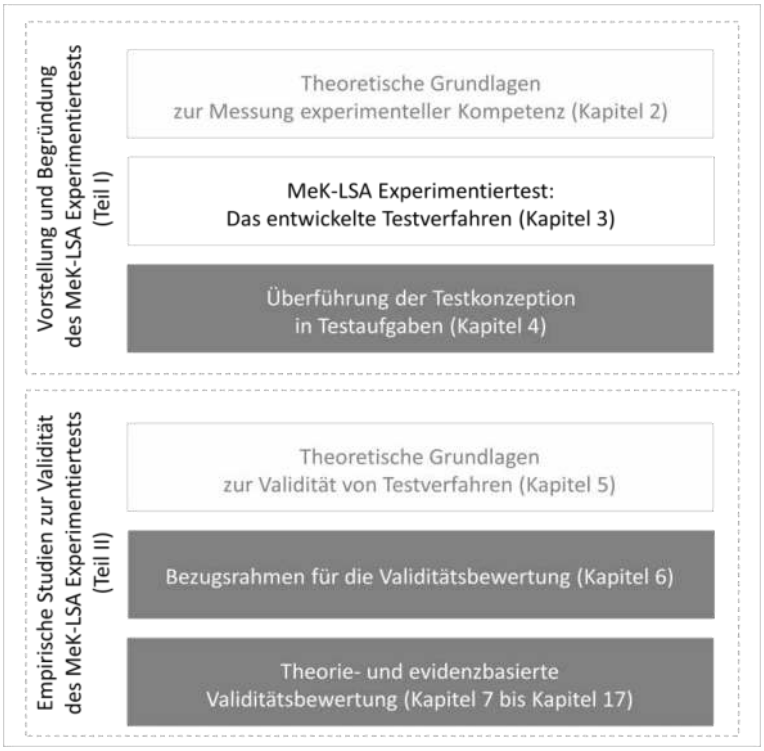


Abbildung 1.1: struktureller Aufbau der Dissertation (weiße Kästen mit grauer Schrift: theoretische Grundlagen; weißer Kasten mit schwarzer Schrift: Darstellung des vom Projektteam erarbeiteten Tests; graue Kästen mit weißer Schrift: zentraler Gegenstand der vorliegenden Arbeit)

Teil I: Vorstellung und Begründung des MeK-LSA Experimentiertests

In Teil I dieser Dissertation wird der – durch das MeK-LSA Projektteam – entwickelte MeK-LSA Experimentiertest vorgestellt und begründet (Kapitel 3). Den Bezugspunkt für die Vorstellung des Testverfahrens bilden theoretische Grundlagen zur Messung experimenteller Kompetenz (Kapitel 2). Das zentrale Element in Teil I ist die detaillierte Beschreibung der systematischen Überführung der Testkonzeption in konkrete Testaufgaben (Kapitel 4). Die detaillierte Beschreibung des Überführungsprozesses ermöglicht größtmögliche Transparenz im Hinblick auf die Inhalte und Anforderungen der Testaufgaben. Der Überführungsprozess zeichnet sich dadurch aus, dass neben Entscheidungen des MeK-LSA Projektteams insbesondere auch kriteriengeleitete Dokumentanalysen und Einschätzungen von Expertinnen und Experten, die nicht operativ am Projekt MeK-LSA beteiligt waren, den Überführungsprozess mitbestimmt haben.

2 Theoretische Grundlagen zur Messung experimenteller Kompetenz

Testentwicklerinnen und Testentwickler müssen bei der Entwicklung eines Testverfahrens zur Messung experimenteller Kompetenz im Wesentlichen drei Entscheidungen treffen:

- Die Auswahl (ggf. auch Konzeption) eines Modells experimenteller Kompetenz als Grundlage für die Testentwicklung,
- die Auswahl (ggf. Konzeption) einer Testsituation, in der alle im Modell experimenteller Kompetenz beschriebenen Fähigkeiten erfasst werden können,
- die Auswahl (ggf. Konzeption) eines geeigneten Auswertungsverfahren zur Bewertung der (Schüler-)Lösungen.

In den Abschnitten 2.1 bis 2.3 werden die theoretischen Grundlagen beschrieben, die für diese drei Entscheidungen bei der Entwicklung des MeK-LSA Experimentiertests relevant sind.

2.1 Modelle experimenteller Kompetenz

Bei der Modellierung experimenteller Kompetenz unterscheiden Gut, Metzger, Hild und Tardent (2014b, S. 2) u.a. zwischen der inneren und äußeren Abgrenzung experimenteller Kompetenz. Bei der inneren Abgrenzung gehe es um die Frage, welche Teilbereiche unterschieden werden sollen. Bei der äußeren Abgrenzung müsse die Frage geklärt werden, welche Fähigkeiten zur Kompetenz zählen und welche nicht. Bezüglich der inneren Abgrenzung dominiert in bestehenden Modellen experimenteller Kompetenz der Teilprozessansatz, der den Prozess des Experimentierens idealtypisch als lineare Abfolge von Teilprozessen beschreibt (vgl. Gut et al., 2014a, S. 171). Nach Emden und Sumfleth (2012, S. 69) lassen sich die Teilprozesse in drei übergeordnete Bereiche gliedern: Ideen- bzw. Hypothesenfindung, Durchführung und Auswertung. Eine Verortung von Kompetenzmodellen im Teilprozessansatz bedeutet allerdings nicht automatisch eine Vergleichbarkeit der Modelle untereinander.

Die Modelle unterscheiden sich zum Teil erheblich bezüglich der genauen Bezeichnungen der drei übergeordneten Bereiche und des Auflösungsgrades (vgl. Neumann, 2013, S. 35) innerhalb der Bereiche (Übersichten über Kompetenzmodelle geben z. B. Emden, 2011, S. 11, und Schreiber, 2012, S. 28).

Der unterschiedliche Auflösungsgrad von Modellen innerhalb des Teilprozessansatzes wird exemplarisch am Bereich der Durchführung verdeutlicht. Während im ursprünglichen Strukturmodell zum Wissenschaftlichen Denken von Mayer (2007, S. 181) der Bereich der Durchführung gar nicht berücksichtigt oder in anderen Modellen nicht näher ausdifferenziert wird (z. B. Hammann, 2004; Walpuski, 2006), lösen wiederum andere Modelle gerade diesen Bereich genauer auf (z. B.: Schreiber, Theyßen & Schecker, 2009; S. 93; Nawrath, Maiseyenka & Schecker, 2011, S. 43 ; Meier & Mayer, 2014, S. 9; vgl. Abschnitt 2.1.2). Neben Modellen experimenteller Kompetenz, die sich ausschließlich auf den Teilprozessansatz beziehen, sind in der Biologiedidaktik (z. B. Wellnitz, 2012, S. 41) und der Chemiedidaktik (z. B. Nehring, Nowak, Tiemann & Upmeier zu Belzen, 2012, S. 302) für den Bereich der Erkenntnisgewinnung Modelle entwickelt worden, die neben der Unterscheidung von Teilprozessen auch zwischen verschiedenen naturwissenschaftlichen Arbeitsweisen (z. B. Modelle nutzen – Experimentieren - Beobachten, Vergleichen, Ordnen, vgl. Nehring, 2014, S. 49) differenzieren. In der Schweiz (vgl. Gut et al., 2014a,b) wird im Rahmen des Projekts *ExKoNawi* ein zum Teilprozessansatz alternativer Problemtypenansatz erprobt. Kompetenz wird in diesem Ansatz in Anlehnung an Gott und Duggan (2002) als auf transferfähigem Wissen basierende Problemlösefähigkeit verstanden (Hild, Tardent, Gut & Metzger, 2015, S. 145). Eine Klassifikation von Experimentieraufgaben erfolgt über den jeweiligen Problemtyp (Gut et al., 2014a, S. 171-172). Beispielsweise werden die folgenden Problemtypen unterschieden: kategoriengeleitetes Beobachten, skalenbasiertes Messen, fragengeleitetes Untersuchen und effektbasiertes Vergleichen.

Es lässt sich festhalten, dass Testverfahren, die bezüglich der inneren Abgrenzung dem Teilprozessansatz folgen, anschlussfähig an die Mehrzahl bestehender Modelle experimenteller Kompetenz sind. Sollen mit einem Testverfahren Experimentierfähigkeiten auch im Bereich der Durchführung differenziert erfasst werden, muss das als Grundlage für die Testentwicklung verwendete Modell auch den Bereich der Durchführung genau auflösen. In Abschnitt 2.1.1 werden Modelle mit hohem Auflösungsgrad im Bereich der Durchführung beschrieben und diskutiert.

2.1.1 Modelle mit hohem Auflösungsgrad im Bereich der Durchführung

Schreiber et al. (2009, S. 93) entwickelten ein Modell experimenteller Kompetenz (*eXkomp-Modell*; vgl. Abbildung 2.1 auf Seite 19), das 13 experimentelle Teilfähigkeiten unterscheidet. Der Bereich der Durchführung wird in Anlehnung an Kempa (1986, S. 70) und Lunetta (2003, S. 255) durch sechs (vier plus zwei) Teilfähigkeiten aufgelöst.

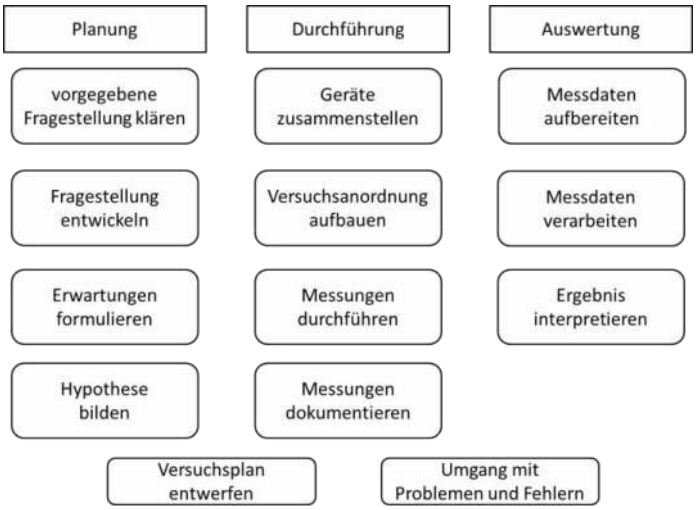


Abbildung 2.1: eXkomp-Modell experimenteller Kompetenz (Schreiber et al., 2009, S. 93)

Der Versuchsplan kann zum Beispiel gedanklich vor der Durchführung entworfen werden oder während der Durchführung des Versuchs modifiziert werden (Schreiber, 2012, S. 36). Die Teilfähigkeit *Versuchsplan entwerfen* wird daher sowohl dem Bereich der Planung, als auch dem Bereich der Durchführung zugeordnet (ebenda). Der Umgang mit Problemen und Fehlern kann sowohl bei der Durchführung des Versuchs, als auch bei der Interpretation von Messergebnissen eine Rolle spielen und wird daher den Bereichen Durchführung und Auswertung zugeordnet (ebenda, S. 37). Das *eXkomp-Modell* impliziert nicht, dass die drei Bereiche Planung, Durchführung und Auswertung während des Experimentierens in einer bestimmten Reihenfolge durchlaufen werden müssen. Die im Modell beschriebenen Komponenten können durchaus in unterschiedlichen oder unvollständigen Abfolgen auftreten (Schreiber et al., 2009, S. 94) Die Komponenten des *eXkomp-Modells* wurden von 25 Lehrkräften sowohl für die tatsächlich stattfindende Unterrichtspraxis (*Realfall*), als auch für die angestrebte Unterrichtspraxis (*Idealfall*) als relevant eingeschätzt (vgl. Schreiber, 2012, S. 43).

Ausgehend vom *eXkomp-Modell* wurde im Rahmen des Projekts *alles>>könner* (z. B. Maiseyenka, 2014, S. 20 - 23.) in enger Zusammenarbeit zwischen Fachdidaktikern und Fachdidaktikerinnen sowie Lehrkräften ein Modell experimenteller Kompetenz für die Schulpraxis entwickelt (*Spinnennetzmodell*; vgl. Abbildung 2.2 auf Seite 20). In die Entwicklung des *Spinnennetzmodells* sind normative und deskriptive Modellaspekte eingeflossen (Maiseyenka, 2014, S. 21). Ähnlich zur Unterscheidung zwischen *Realfall* und *Idealfall* bei Schreiber (2012, S. 43) bezeichnen normative Aspekte in diesem Zusammenhang Experimentierfähigkeiten, die Schülerinnen und Schüler aus Sicht von Lehrkräften im Unterricht idealerweise erwerben sollten, während deskriptive Aspekte Experimentierfähigkeiten beschreiben, die in der tatsächlichen Unterrichtspraxis häufig betont werden. An der Entwicklung waren im Kern 14 Lehrkräfte mit Unterrichtserfahrung in mindestens einem der Fächer Biologie, Chemie, Naturwissenschaft und Physik beteiligt (vgl. Maiseyenka, Schecker & Nawrath, 2013, S. 3).

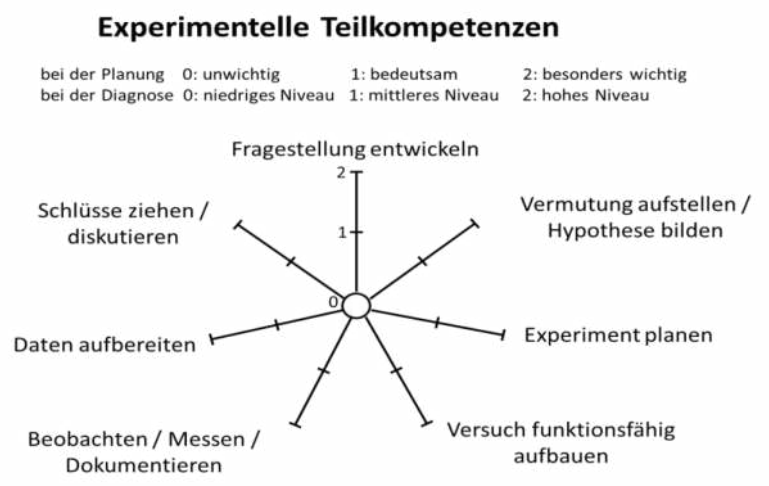


Abbildung 2.2: Spinnennetzmodell experimenteller Kompetenz (Maiseyenka et al., 2013, S. 6)

Im Vergleich zum *eXkomp-Modell* wird die Anzahl der Teilfähigkeiten von 13 auf sieben reduziert. Hintergrund für die Reduktion der Teilfähigkeiten war die aus Lehrersicht kritisierte Komplexität von strukturell stark aufgeschlüsselten Modellen im Hinblick auf die Praktikabilität ihrer Nutzung für die Unterrichtsplanung (Schecker, Nawrath, Elvers, Borgstädt, Einfeldt & Maiseyenka, 2013, S. 9). Einzelne Komponenten des *eXkomp-Modells* werden daher zusammengefasst und zum Teil inhaltlich weiter ausdifferenziert, wobei sich die inhaltliche Ausdifferenzierung bewusst nicht in der Struktur des *Spinnennetzmodells* widerspiegelt. Zum Beispiel gehört zur Fähigkeit einen Versuch funktionsfähig aufbauen zu können,

> *„das Zusammenstellen der in der Planung vorgesehenen Geräte, der Aufbau der Versuchsanordnung, das Testen der Funktion des Aufbaus […]. Beim Versuchsaufbau und der Kontrolle der Funktionsfähigkeit kann es erforderlich sein, Fehlerquellen ausfindig zu machen und ggf. den Aufbau zu variieren. Dies setzt eine systematische Fehlersuche voraus"* (Nawrath et al., 2011, S. 47).

Eine umfassende Beschreibung aller Teilfähigkeiten des *Spinnennetzmodells* findet sich in Nawrath et al. (2011, S. 47). Das *Spinnennetzmodell* kann einerseits zur Planung experimenteller Untersuchungen im Unterricht und andererseits zur Diagnose von Experimentierfähigkeiten von Schülerinnen und Schülern genutzt werden (vgl. Maiseyenka et al., 2013, S. 6). Sowohl das *Spinnennetzmodell* als auch das *eXkomp-Modell* sind anschlussfähig an nationale und internationale Standardbeschreibungen und weisen eine hohe Passung zu anderen im Teilprozessansatz verorteten Kompetenzmodellen auf.

Die Berücksichtigung und genaue Auflösung des Bereichs der Durchführung wird auch in den neueren Arbeiten von Meier und Mayer (2012, 2014; vgl. Abbildung 2.3 auf Seite 21) aus der Biologiedidaktik aufgegriffen. Im Rahmen einer qualitativen Videostudie mit 13 Schülergruppen konnten Meier und Mayer (2012, S. 81-98) einen Bereich der Durchführung identifizieren, der die von Mayer (2007, S. 181) identifizierten Prozessvariablen ergänzt.

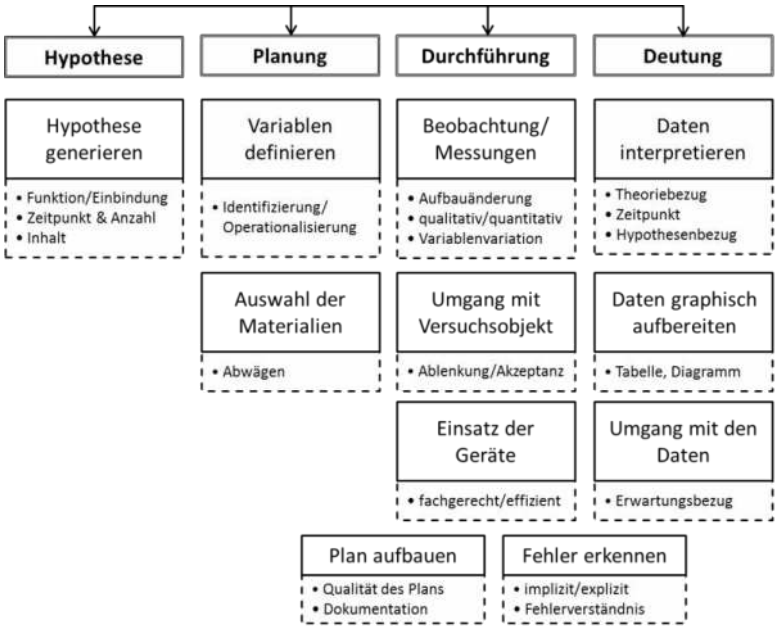


Abbildung 2.3: Fähigkeiten zur praktischen Durchführung eines Experiments (nach Meier & Mayer, 2014, S. 9; Einfärbung aus Original nicht übernommen)

In den drei beschriebenen Modellen werden zwar praktische Experimentierfähigkeiten betont, es geht jedoch nicht um die Operationalisierung von manuellem Geschick, sondern um die Beschreibung von handlungsbezogenen Teilfähigkeiten. Beim Durchführen von Messungen geht es zum Beispiel um die Fähigkeiten, sich Gedanken über die Qualität der Messdaten, die Anzahl der Messwerte oder einen sinnvollen Wertebereich während der Aufnahme einer eigenen Messreihe zu machen. Die beschriebenen Fähigkeiten können folglich im Sinne der Kompetenzdefinition von Klieme und Leutner (2006, S. 879) als kognitive Leistungsdispositionen in einer tatsächlichen Experimentiersituation aufgefasst werden.

Die drei beschriebenen Modelle unterscheiden sich im Hinblick auf die gewählten Entwicklungsansätze. Das *eXkomp-Modell* (vgl. Abbildung 2.1) wurde primär theoriegeleitet entwickelt. Das *Spinnennetzmodell* (vgl. Abbildung 2.2) wurde durch Lehrkräfte mitentwickelt und berücksichtigt insbesondere deren Vorstellungen über experimentelle Anforderungen aus der Unterrichtspraxis. Das Modell von Meier und Mayer (vgl. Abbildung 2.3) wurde auf Basis beobachteter Schülerhandlungen während der Bearbeitung einer experimentellen Aufgabenstellung (*Vertikalwanderung von Wasserflöhen*; vgl. Meier & Mayer, 2014, S. 5) erarbeitet. Trotz der unterschiedlichen Entwicklungsansätze sind deutliche Gemeinsamkeiten zwischen den Modellen zu erkennen. Das ist ein Argument dafür, dass es sich bei den in den Modellen beschriebenen Fähigkeiten um relevante Bestandteile experimenteller Kompetenz handelt.

2.2 Testsituationen zur Erfassung experimenteller Kompetenz

Bei der Entwicklung von Testverfahren zur Messung experimenteller Kompetenz muss diskutiert werden, in welcher Situation die Kompetenzmessung erfolgen soll. Dabei stellt sich die Frage, ob das Vorhandensein des Wissens über experimentelles Vorgehen (*Wissen wie*) getestet werden soll, oder ob Experimentierfähigkeiten über die tatsächliche Anwendung in repräsentativen Situationen (*Zeigen wie*) erfasst werden sollen (vgl. auch Miller, 1990, S. 63; Blömeke, 2013, S. 4).

Die Diskussion um die *Art der Testsituati*on (*Wissen wie* und *Zeigen wie*) muss insbesondere für Large-Scale Assessments geführt werden. Auf der einen Seite wird mit Large-Scale Assessments der Anspruch verbunden, die inhaltlich und theoretisch bedeutsamen Facetten einer Kompetenz abbilden zu können (Pellegrino et al., 2001, S. 13). Auf der anderen Seite müssen Large-Scale Assessments mit vertretbarem zeitlichem und ökonomischem Aufwand administrier- und auswertbar sein. Das führt dazu, dass häufig *„hochgradig standardisierte Messverfahren im Sinne von […] Kompetenztests“* (Wendt & Bos, 2011, S. 17) eingesetzt werden, die sich auf solche Fähigkeitsbereiche einer Kompetenz beschränken, die effizient und *„psychometrisch erfolgsversprechend“* (Gut, 2012, S. 23) gemessen werden können. In Deutschland wurden zum Beispiel bei der Überprüfung der Bildungsstandards zu den Naturwissenschaften (Mittlerer Bildungsabschluss) aus testökonomischen Gründen praktische Fähigkeiten wie das eigenständige Durchführen einer Untersuchung, *„zunächst als prozedurales Wissen („Wissen, wie“) interpretiert und entsprechend* [schriftlich] *getestet“* (Wellnitz et al., 2012, S. 273).

Rein schriftliche Testverfahren sind in der Lage, einzelne anwendungsbezogene Experimentierfähigkeiten, wie beispielsweise das *Aufstellen von Hypothesen*, das *Planen eines Versuchs* oder das *Analysieren von Daten,* ausreichend reliabel und vergleichsweise valide zu erfassen (z. B. Glug, 2009, S. 221-237). Weitere wichtige Experimentierfähigkeiten (vgl. Abschnitt 2.1), wie das eigenständige *Aufbauen einer Versuchsanordnung* oder das *Durchführen und Dokumentieren von Messungen,* können mit rein schriftlichen Testverfahren nicht adäquat erfasst werden (z. B. Schreiber, 2012, S. 142-145). Darüber hinaus weist Meier (2014, S. 137) darauf hin, dass auch die als schriftlich erfassbar angesehenen Fähigkeiten durch die Einbettung in eine tatsächliche experimentelle Handlungssituation beeinflusst werden.

> *„[So] ist die Planung eines eigenen Experimentes beispielsweise im hohen Maße an die gegebenen Materialien und der damit verbundenen Auswahl gebunden, d. h. über Abwägen oder Ausprobieren wird von den Lernenden ein Plan konstruiert und die Variablen operationalisiert“* (Meier, 2014, S. 137).

Sollen Experimentierfähigkeiten, insbesondere im Bereich der Durchführung, in einer tatsächlichen experimentellen Handlungssituation erfasst werden (*Zeigen wie*), ist ein rein schriftlicher Test kein geeignetes Format (vgl. IQB, 2013, S. 16).

Experimentiertests mit hands-on Aufgaben sind bisher, insbesondere für Large-Scale Assessments, keine zufriedenstellende Lösung. Neben methodischen Schwierigkeiten, z. B. vergleichsweise niedrigen Reliabilitäten aufgrund einer geringen Anzahl bearbeiteter Testaufgaben pro Testteilnehmendem, die sich in der Vergangenheit beim Einsatz von Experimentiertests mit hands-on Aufgaben gezeigt haben (Übersicht: Schreiber et al., 2014, S. 163; vgl. Abschnitt 1.2), gilt vor allem die direkte Beobachtung von Experimentierfähigkeiten als sehr ressourcenintensiv. In Large-Scale Assessments wurde diese ressourcenintensive Vorgehensweise bisher nur im Rahmen des *Assessment of Performance Unit* (APU, z. B. Johnson, 1989) umgesetzt. Um eine ressourcenschonendere Erfassung von Experimentierfähigkeiten zu ermöglichen, werden verschiedene Formate diskutiert, die eine direkte Beobachtung des Experimentierprozesses ersetzen sollen (z. B. Emden, 2011; Shavelson, Baxter & Pine, 1991). Beim *TIMSS-Experimentiertest* (vgl. Stebler, Reusser & Ramseier, 1998) und beim *HarmoS-Experimentiertest* (vgl. Adamina et al., 2009) wurden die Experimentierfähigkeiten beispielsweise ausschließlich auf Basis der von den Schülerinnen und Schülern auf Protokollbögen eingetragenen Antworten bewertet, ohne dass der Bearbeitungsprozess aufgezeichnet und analysiert wurde. Bei einem solchen Format kann beispielsweise nicht entschieden werden, ob von Schülerinnen und Schülern dokumentierte Messwerte mit den tatsächlich gemessenen Werten übereinstimmen. In diesem Sinne erlauben rein auf Schülerauskünften basierende Formate nur eine indirekte Messung anwendungsbezogener Experimentierfähigkeiten (vgl. Abrahams, Reiss & Sharpe, 2013).

Als weiteres alternatives Format zur Messung von Experimentierfähigkeiten werden on-screen Tests mit interaktiven Simulationen diskutiert. Dieses Format gilt als pragmatische Alternative zu Experimentiertests mit hands-on Aufgaben, da Schülerinnen und Schüler innerhalb interaktiver Simulationen zumindest virtuelle Handlungen vornehmen können und zusätzlich die Möglichkeit besteht, alle Eingaben und Handlungen automatisch zu speichern. On-screen Tests mit interaktiven Simulationen liegen z. B. von Baxter und Shavelson (1994), Schreiber et al. (2014) oder Quellmalz, Timms und Buckley (2010) vor. Ob on-screen Tests mit interaktiven Simulationen die gleichen Fähigkeiten wie Experimentiertests mit hands-on Aufgaben erfassen, ist noch nicht abschließend geklärt (vgl. Shavelson, Ruiz-Primo & Wiley, 1999, Schreiber, 2012; NAGB, 2008, S. 108). Es gibt allerdings empirische Hinweise darauf, dass on-screen Experimentiertests zumindest auf Populationsebene hands-on Experimentiertests ersetzen können (Schreiber et al., 2014, S. 171). Die Aussagen zur Austauschbarkeit beziehen sich jedoch bereits auf die Ebene der Testwerte. Selbst wenn man auf Ebene der Testwerte hohe positive Zusammenhänge zwischen den Formaten findet, muss das nicht notwendigerweise bedeuten, dass die beiden Formate das gleiche messen. Hohe positive Zusammenhänge können auch dann bestehen, wenn sich die kognitiven Anforderungen unterscheiden (vgl. NAGB, 2008, S. 97). Ob bei der Bearbeitung von hands-on Aufgaben und on-screen Aufgaben aufseiten der Schülerinnen und Schüler vergleichbare kognitive Prozesse ablaufen, ist bisher nicht näher untersucht worden. Bisherige Untersuchungen zu kognitiven Prozessen beschränken sich entweder auf hands-on Aufgaben (vgl. Hamilton, 1994) oder auf on-screen Aufgaben (vgl. Quellmalz, Silberglitt & Timms, 2011, S. 6-7).

Trotz dieser noch offenen Fragen sind 2009 im Rahmen des landesweiten NAEP Science Assessments in den USA neben rein schriftlichen und hands-on Aufgaben in einer Teilstichprobe erstmals auch on-screen Aufgaben mit interaktiven Simulationen eingesetzt worden (vgl. NCES, 2012). Zu diesem Zweck wurden hands-on Aufgaben mit Realexperimenten in on-screen Aufgaben mit interaktiven Simulationen überführt. Die Aufgaben erfordern von den Schülerinnen und Schülern das Planen und Durchführen einer Untersuchung, um daraus Schlussfolgerungen bezogen auf ein Problem zu ziehen (NAGB, 2008, S. 108). Dabei wird die Bearbeitung der Aufgaben durch Teilaufgaben vorstrukturiert. Die Aufgabe *Bottling Honey* (Zielgruppe: grade 8) wird beispielsweise in fünf Teilaufgaben gegliedert. Abbildung 2.4 zeigt beispielhaft eine Teilaufgabe, in der die Schülerinnen und Schüler den Zusammenhang zwischen der Honigtemperatur und der Flussrate des Honigs (über eine Viskositätsmessung) bestimmen sollen. Die Aufgabe ist stark vorstrukturiert und zur Lösung der Aufgabe muss in der Simulation eine Messreihe aufgenommen werden. Die Schülerinnen und Schüler haben innerhalb der Simulation die Möglichkeit, die Temperatur zu verändern und die Kugel fallen zu lassen. Die Messwerte zu den Messgrößen Zeit und Temperatur werden automatisch dokumentiert. Aus den Messwerten kann nur auf die richtige Antwort (D: Graph 4) geschlossen werden, wenn in einem geeigneten Messbereich gemessen wurde.

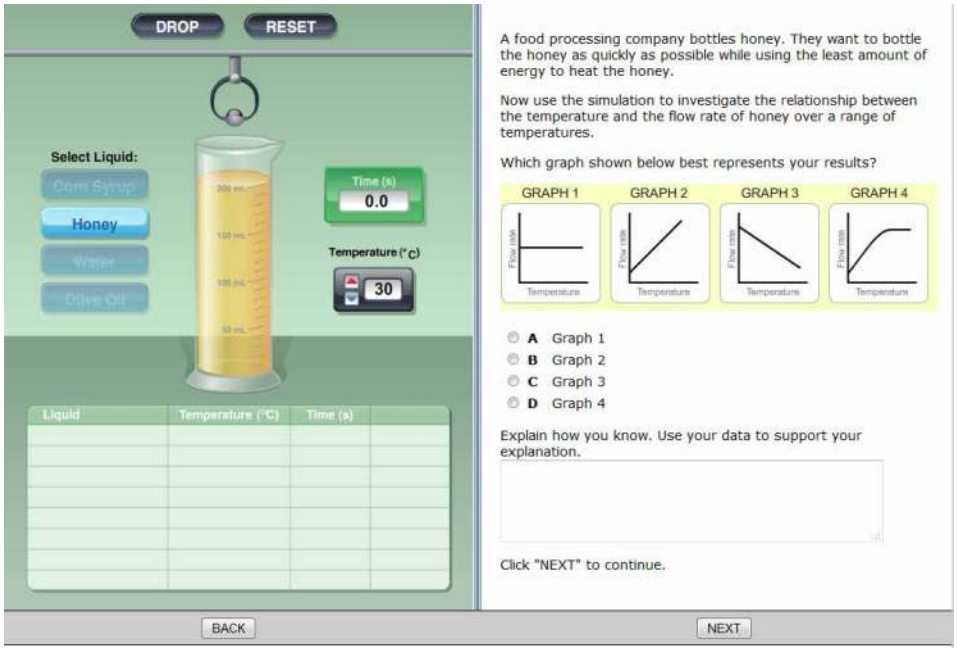


Abbildung 2.4: on-screen Teilaufgabe aus dem NAEP Science Assessment 2009 (Screenshot: http://www.nationsreportcard.gov/science2009ict/bottlinghoney/bottlinghoney4a.aspx; Datum: 22.01.16)

Durch die Vorstrukturierung können trotz testökonomischer Herausforderungen auch in Large-Scale Assessments anwendungsbezogene Experimentierfähigkeiten (z. B. Durchführen von Messungen) miterfasst werden. Im Gegensatz zu den NAEP Science Assessments wurde im *eXkomp-Projekt* (Schreiber et al., 2009) ein weniger vorstrukturiertes Aufgabenformat eingesetzt, bei dem offene Aufgaben ohne Bearbeitungsleitfaden gestellt wurden (vgl. Abbildung 2.5 auf Seite 25; Aufgabenstellung in der Abbildung).

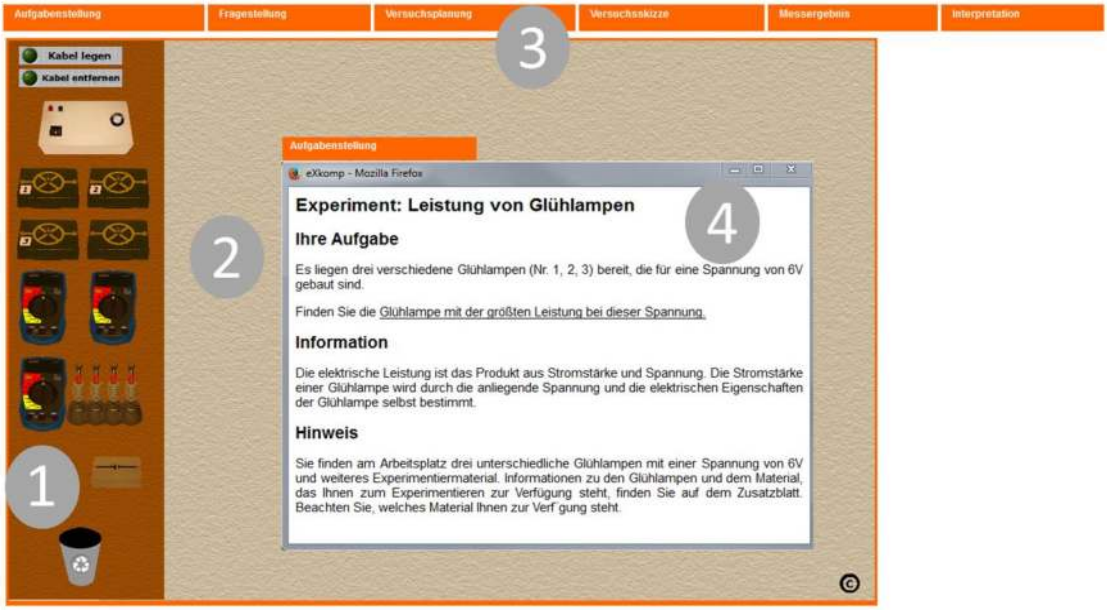


Abbildung 2.5: Aufgabe aus dem eXkomp Projekt (Schreiber et al., 2014); 1: zur Verfügung stehendes Experimentiermaterial; 2: Simulationsfläche; 3: Reiter zum Öffnen von Pop-up Fenstern mit Aufgabenstellung bzw. Möglichkeiten zur Bearbeitung von weiteren Teilaufgaben (z. B. Versuchsskizze anfertigen); 4: geöffnetes Pop-up Fenster mit Aufgabenstellung (hier: „Leistung von Glühlampen bestimmen"); eigener Screenshot)

Die Schülerinnen und Schüler mussten den gesamten Bearbeitungsprozess eigenständig strukturieren und bekamen beim Auftreten von Problemen keine Hilfestellung. Das führte dazu, dass viele Schülerinnen und Schüler die Aufgaben nicht vollständig bearbeiteten bzw. bearbeiten konnten, da Folgefehler nicht abgefangen wurden (vgl. Schreiber, 2012, S. 136). Ist es z. B. einem Schüler nicht gelungen, einen funktionsfähigen Versuch herzustellen, konnten in der Folge auch keine Messungen durchgeführt werden. Über die Fähigkeit dieses Schülers, Messungen durchzuführen und zu dokumentieren, kann folglich keine Aussage getroffen werden, da er nicht die Möglichkeit hatte, diese Fähigkeit zu zeigen. An diesem Beispiel zeigen sich zwei Probleme. Zum einen fehlen Hilfestellungen, die bei auftretenden Problemen (z. B. Versuch wird nicht funktionsfähig aufgebaut) den Wiedereinstieg in die Bearbeitung ermöglichen (z. B. das Bereitstellen eines funktionsfähigen Aufbaus, mit dem Messungen durchgeführt werden könnten). Zum anderen werden Teilfähigkeiten (z. B. Versuch funktionsfähig aufbauen und Messungen durchführen) erst durch die Bewertung der Lösung unterschieden. Auf diese Weise erfolgt eine nachträgliche Sequenzierung in Teilaufgaben. Diese Teilaufgaben und die Lösungen zu den Teilaufgaben sind jedoch nicht mehr unabhängig voneinander. Das hat u.a. Auswirkungen auf die Bewertung der (Schüler-)Lösungen einzelner Teilaufgaben. So ist es beispielsweise schwierig, die korrekte Umsetzung einer eigenen falschen Skizze im experimentellen Aufbau zu bewerten (Problem: Umgang mit Folgefehlern). Das Problem der Abhängigkeit von Teilaufgaben beschreibt auch Gut (2012, S. 206) für den *HarmoS-Experimentiertest*. Sollen mit einem Testverfahren einzelne Teilfähigkeiten unterschieden werden, sind offene Aufgaben ohne Bearbeitungsleitfaden nicht geeignet.

Die in diesem Abschnitt genannten Ergebnisse sprechen dafür, dass on-screen Experimentiertests mit interaktiven Simulationen eine Alternative zu hands-on

Experimentiertests darstellen, wobei weiter zu prüfen ist, bis zu welchem Grad die Performanz der Schülerinnen und Schüler durch das Testformat (hands-on, on-screen) beeinflusst wird. Darüber hinaus legen die Ergebnisse nahe, dass Strukturierungsmaßnahmen getroffen werden sollten, damit Schülerinnen und Schüler die Aufgabenanforderungen bewältigen können und die Unterscheidung einzelner Teilfähigkeiten möglich ist.

2.3 Auswertungsverfahren zur Bewertung der Lösungsqualität experimenteller Aufgaben

Die Lösungsqualität experimenteller Aufgaben muss kriteriengeleitet bewertet werden. Nach Baxter, Shavelson, Goldman und Pine (1992, S. 2) sollte der zugewiesene Testwert (als Maß für die Lösungsqualität) sowohl den Prozess, der zur Lösung geführt hat, als auch die Lösung selbst widerspiegeln. Schreiber (2012, S. 51-55) und Gut (2012, S. 87-91) haben Auswertungsverfahren verschiedener Experimentiertests qualitativ analysiert. Schreiber (2012, S. 51-55) unterscheidet bei seiner Analyse zwischen produktbezogenen und prozessbezogenen Auswertungsverfahren. Während sich eine produktbezogene Bewertung ausschließlich auf die von den Schülerinnen und Schülern dokumentierte Lösung bezieht (z. B. eine Skizze des Versuchsaufbaus), werden bei der prozessbezogenen Bewertung insbesondere die Schritte auf dem Weg zur Lösung berücksichtigt. Die Analyse der produktbezogenen Auswertungsverfahren (u.a. *HarmoS-Experimentiertest*, *TIMSS-Experimentiertest*) zeigt, dass diese in der Regel nach rein fachlichen Kriterien bewertet werden. Die von Schreiber (2012, S. 53-54) analysierten prozessbezogenen Auswertungsverfahren gehören zu Experimentiertests*,* die nicht für den Einsatz in Large-Scale Assessments konzipiert wurden (z. B. Neumann, 2004). Das zeigt sich sowohl in den Auswertungsverfahren als auch in der Konzeption der Tests: Die Testaufgaben wurden in Kleingruppen bearbeitet, und die Auswertungsverfahren berücksichtigen die Kommunikation in der Gruppe. Gut (2012, S. 85) berücksichtigt bei seiner qualitativen Analyse von Auswertungsverfahren u.a. folgende Experimentiertests: *HarmoS-Experimentiertest*, *TIMSS-Experimentiertest*, ausgewählte Aufgaben: *NAEP-Assessments*, *APU-Staffeln*, *QuiP-Experimentiertest*. Bei der Analyse unterscheidet Gut (2012, S. 87-89) die Auswertungsverfahren nach den folgenden sechs Wissensarten, die zur korrekten Lösung der Aufgaben benötigt werden (*Korrektheitsideale*):

- Theoretische Korrektheit
- Evidenzielle Korrektheit
- Technisch-praktische Korrektheit
- Heuristische Korrektheit
- Schlusslogische Korrektheit
- Mathematisch-logische Korrektheit

Beispielsweise liegt das Korrektheitsideal *Heuristische Korrektheit* Auswertungsverfahren zugrunde, die das methodische Vorgehen bewerten (ebenda). Darüber hinaus werden nach Gut (2012, S. 89-91) die folgenden kompetenzunspezifischen Maßstäbe angelegt: Vollständigkeit, Darstellung und Konsistenz.

Die von Schreiber (2012, S. 51-55.) und Gut (2012, S. 87-91.) durchgeführten Analysen zeigen, dass jeder Experimentiertest üblicherweise unterschiedliche, testspezifische Bewertungskriterien festlegt, sodass eine Vergleichbarkeit der Testwerte (als Maß für die Lösungsqualität) verschiedener Experimentiertests kaum gegeben ist. Darüber hinaus liegen bisher keine konsensfähigen Qualitätsstandards für die Konzeption von Auswertungsverfahren für Experimentiertests vor.

Die Auswahl bzw. Konzeption eines geeigneten Auswertungsverfahrens ist daher bei der Entwicklung eines Testverfahrens zur Messung experimenteller Kompetenz nach wie vor eine Herausforderung für die Testentwicklerinnen und Testentwickler.

3 MeK-LSA Experimentiertest: Das entwickelte Testverfahren

Aufbauend auf den theoretischen Grundlagen zur Messung experimenteller Kompetenz (Kapitel 2) wurde vom MeK-LSA Projektteam ein Testverfahren (MeK-LSA Experimentiertest) mit vollständig on-screen zu bearbeitenden, experimentellen Aufgabenstellungen entwickelt, das experimentelle Kompetenz von Schülerinnen und Schülern am Ende der Sekundarstufe I messen soll.

In diesem Kapitel werden die Konzeption (Abschnitt 3.1), die Materialien (Abschnitt 3.2) und das Auswertungsverfahren (Abschnitt 3.3) des MeK-LSA-Experimentiertests vorgestellt, um einen Überblick über das Testverfahren zu erhalten. Eine detaillierte Darstellung des MeK-LSA Experimentiertests findet man in Theyßen et al. (2016b).

3.1 MeK-LSA Experimentiertest: Testkonzeption

Grundlage für die Entwicklung des MeK-LSA Experimentiertest*s* war das in Abbildung 3.1 auf Seite 30 dargestellte Aufgabenentwicklungsmodell. Das Modell folgt bzgl. der inneren Struktur experimenteller Kompetenz dem Teilprozessansatz und ist damit anschlussfähig an die Mehrzahl bestehender Modelle experimenteller Kompetenz und die in den Bildungsstandards formulierten Kompetenzerwartungen zum Experimentieren (vgl. Abschnitt 2.1 auf Seite 17). Die acht im Aufgabenentwicklungsmodell beschriebenen Experimentierfähigkeiten basieren auf dem *eXkomp-Modell* und dem *Spinnennetzmodell* (vgl. Abschnitt 2.1.1 auf Seite 18). Diese Experimentierfähigkeiten sind mit dem experimentellen Prozess von der Entwicklung einer experimentellen Grundidee zu einer gegebenen Fragestellung bis zur Interpretation von Messdaten bezüglich dieser Fragestellung verbunden. Der MeK-LSA Experimentiertest legt bewusst einen Schwerpunkt auf Experimentierfähigkeiten, die im Zentrum des Experimentierens im Physikunterricht stehen (Durchführung), und für deren Erwerb es damit Lerngelegenheiten im Unterricht gibt. Es geht nicht um die Erfassung von Experimentierfähigkeit als Teil einer übergeordneten Problemlösefähigkeit, sondern um Experimentierfähigkeiten, die zur Beantwortung einer gegebenen physikalischen Fragestellung benötigt werden. Die zur Verfügung stehende Testzeit soll sich auf die Erfassung dieser Experimentierfähigkeiten konzentrieren. Fähigkeiten wie das eigenständige *Entwickeln von Fragestellungen* oder *Bilden von Hypothesen,* die beim Experimentieren als Teil eines übergeordneten Problemlöseprozesses zentral sind, spielen in der Unterrichtspraxis kaum eine Rolle (vgl. Tesch & Duit, 2004, S. 59; Oetinger, B., 2013, S. 81). Sie sind zudem bei nicht trivialen und unterrichtsnahen Themen stark vom Fachwissen der Schülerinnen und Schüler abhängig.

Jede Aufgabe wird anhand des Aufgabenentwicklungsmodells (vgl. Abbildung 3.1 auf Seite 30) in sechs Teilaufgaben unterteilt, die in einer schultypischen und sachlogischen Reihenfolge durchlaufen werden. Die Teilaufgaben *Versuchsplan entwerfen*, *Versuch aufbauen und testen* und *Messung durchführen und dokumentieren* sind in jeder Aufgabe enthalten. Sie werden ergänzt durch eine weitere Teilaufgabe zur Planung und zwei Teilaufgaben zur Auswertung.

Abbildung 3.1: Aufgabenentwicklungsmodell zu den drei Bereichen des Experimentierens (dunkelgrau) mit acht Experimentierfähigkeiten experimenteller Kompetenz (hellgrau) (vgl. Theyßen et al., 2016b)

Der inhaltliche Rahmen jeder Aufgabe ist im Aufgabenstamm beschrieben. Er enthält eine übergeordnete experimentelle Aufgabenstellung und aufgabenspezifische fachliche Erklärungen (z. B. woran man erkennt, dass zwei physikalische Größen *proportional* sind). Die Informationen aus dem Aufgabenstamm können bei Bedarf während der Aufgabenbearbeitung wieder aufgerufen werden. Abbildung 3.2 auf Seite 31 zeigt beispielhaft den Aufgabenstamm der Aufgabe *Ausdehnung eines Gummibandes*[3]. Diese Aufgabe ist in Anhang A.1 auf den Seiten 177 bis 183 vollständig dargestellt.

Um die Abhängigkeit zwischen Teilaufgaben und das Auftreten von Folgefehlern zu vermeiden (vgl. Schreiber, 2012, S. 136 & Abschnitt 2.2 auf Seite 22) darf die Bearbeitung einer Teilaufgabe nicht davon abhängen, ob die vorherige Teilaufgabe gelöst wurde. Aus diesem Grund werden zu Beginn einer Teilaufgabe die benötigten Zwischenlösungen vorangegangener Teilaufgaben zur Verfügung gestellt. Beispielsweise wird für die Bearbeitung der Teilaufgabe *Versuch aufbauen und testen* der entworfene Versuchsplan (Geräte, Skizze, Vorgehensweise) vorgegeben (vgl. Abbildung 3.3 auf Seite 32). Somit haben die Schülerinnen und Schüler die Möglichkeit, alle Teilaufgaben zu bearbeiten, auch wenn (einzelne) Teilaufgaben nicht oder nicht korrekt bearbeitet werden konnten.

[3] Die Aufgabe *Ausdehnung eines Gummibandes* dient in der gesamten Arbeit als roter Faden, entlang dessen der MeK-LSA Experimentiertest beispielhaft vorgestellt wird und die Validitätsbewertung nachvollzogen werden kann.

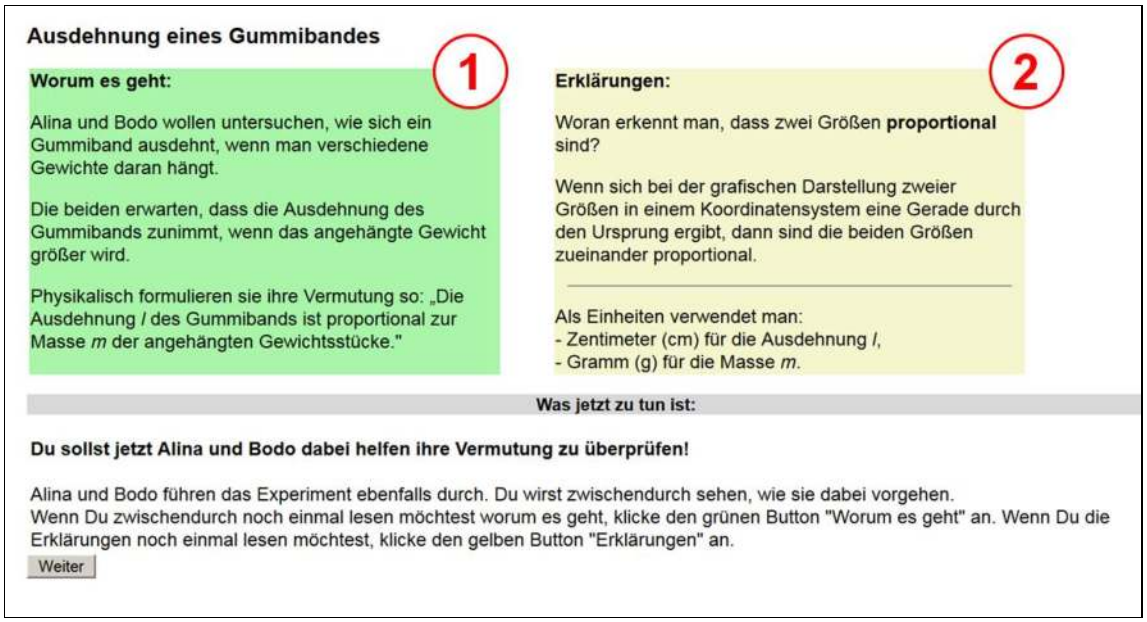


Abbildung 3.2: Aufgabenstamm der Aufgabe Ausdehnung eines Gummibandes; 1: übergeordnete Aufgabenstellung, 2: Fachinformation (Theyßen, Schecker, Dickmann, Eickhorst & Neumann, 2016a)

Um das Zeigen der Zwischenlösungen für die Schülerinnen und Schüler möglichst authentisch zu gestalten, werden die Zwischenlösungen in eine Begleitgeschichte eingebettet: Die Schülerinnen und Schüler werden während der Bearbeitung durch die fiktive Schülerin Alina und den fiktiven Schüler Bodo begleitet. Alina und Bodo stellen zu Beginn eine Vermutung auf bzw. werfen eine Fragestellung auf, aus der sich die experimentelle Aufgabenstellung für die Schülerinnen und Schüler ergibt. Die Vermutung bzw. die Fragestellung von Alina und Bodo soll durch das Bearbeiten der Teilaufgaben überprüft bzw. beantwortet werden. Die Schülerinnen und Schüler sollen Alina und Bodo bei der Bearbeitung des Experiments helfen und sehen, gewissermaßen im Gegenzug, nach jeder Teilaufgabe, wie Alina und Bodo die vorherige Teilaufgabe bearbeitet haben (Zwischenlösung).

Mit dem MeK-LSA Experimentiertest sollen Experimentierfähigkeiten von Schülerinnen und Schülern über deren Anwendung in – wenn auch virtuellen – experimentellen Handlungssituationen erfasst werden (*Zeigen wie;* vgl. Abschnitt 2.2 auf Seite 22). Zu diesem Zweck sind im MeK-LSA Experimentiertest interaktive Elemente (Simulationen, Zeichentools etc.) integriert, die Handlungen mit der Computermaus in einer virtuellen Experimentier- bzw. Zeichenumgebung erfordern. Die interaktiven Elemente werden durch Textfelder ergänzt, in denen Antworten (Text und Zahlenwerte) notiert werden können.

Die Teilaufgabe *Versuchsplan entwerfen* enthält die Teilschritte *Geräteauswahl*, *Anfertigen einer Versuchsskizze* und *Beschreibung der Vorgehensweise*. Der Teilschritt *Geräteauswahl* erfolgt interaktiv. Die Geräte können per Maus ausgewählt werden. Dabei können die Geräte vergrößert dargestellt und verschoben sowie gedreht werden, um sie von verschiedenen Seiten zu betrachten. Die Versuchsskizze wird freihand mit der Maus gezeichnet. Die Beschreibung der Vorgehensweise kann über die Tastatur in ein Textfeld eingegeben werden (vgl. vollständige Darstellung der Aufgabe *Ausdehnung eines Gummibandes* in Anhang A.1 auf den Seiten 177-183).

In der Teilaufgabe *Versuch aufbauen und testen* bauen die Schülerinnen und Schüler auf Basis eines vorgegebenen Versuchsplans (vgl. Abbildung 3.3) eine Versuchsanordnung selbstständig auf. Die Materialien können in der virtuellen Experimentierumgebung mit der Maus bewegt und angeordnet werden. Hängt man Massestücke an das Gummiband, so dehnt sich das Gummiband aus. Die Materialien werden in der virtuellen Umgebung fotorealistisch dargestellt und bilden die Eigenschaften der realen Materialien überwiegend realitätsnah ab.

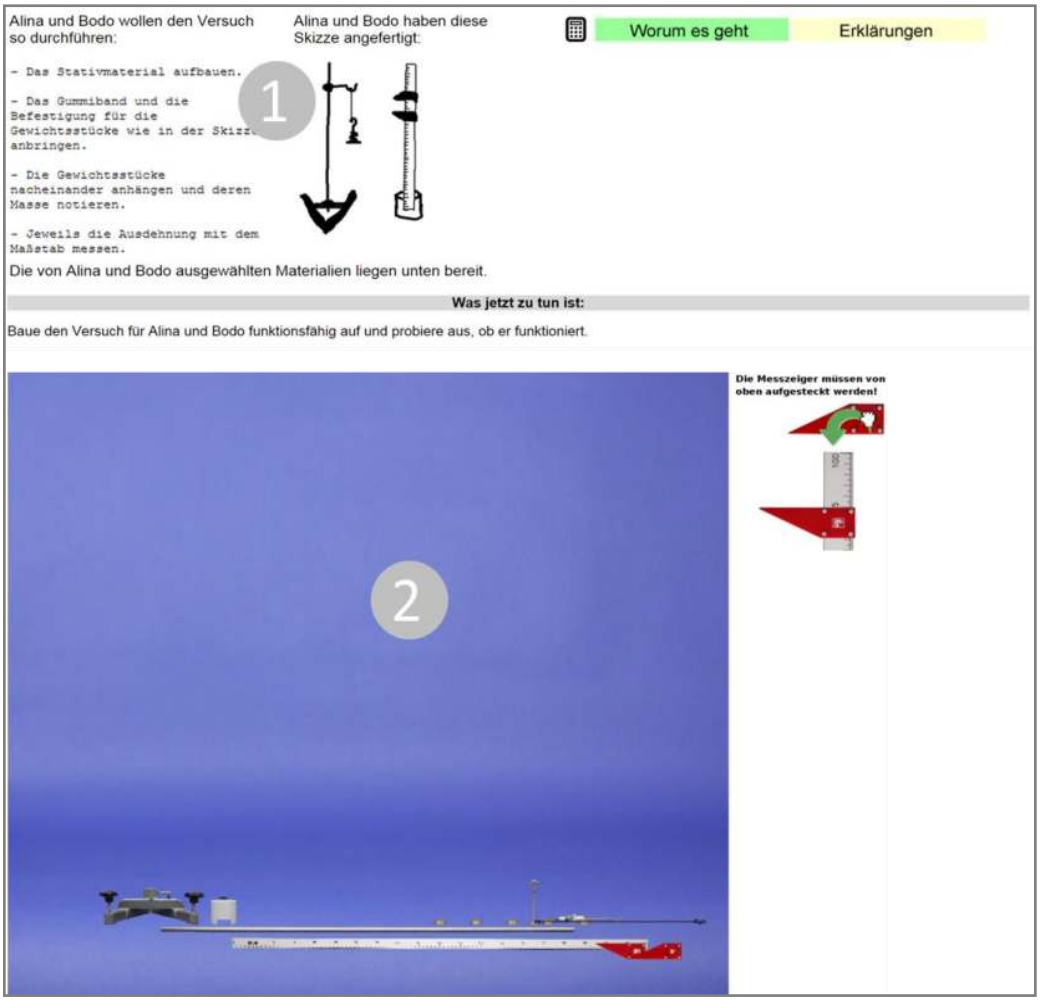


Abbildung 3.3: Beispiel für die Angabe einer Zwischenlösung: Bei der Teilaufgabe Versuch aufbauen und testen *wird den Schülerinnen und Schülern gezeigt, welchen Versuchsplan die fiktiven Personen Alina und Bodo entworfen haben (1). Ihre Aufgabe besteht nun darin, den Versuch für Alina und Bodo in der Experimentierumgebung (2) funktionsfähig aufzubauen und auszuprobieren, ob er funktioniert.*

In der Teilaufgabe *Messung durchführen und dokumentieren* arbeiten die Schülerinnen und Schüler mit einer prinzipiell funktionsfähigen Versuchsanordnung (vgl. Abbildung 3.4 auf Seite 33). Mit dieser Versuchsanordnung können die für die Messung erforderlichen Einstellungen (Justieren der Messzeiger; Anhängen der Massestücke) vorgenommen und die Messwerte abgelesen werden. Weiterführende Änderungen der Versuchsanordnung (z. B. Auseinanderbauen des Stativmaterials, Abhängen des Gummibandes) sind nicht mehr möglich. Folgefehler aus einer möglicherweise nicht funktionsfähigen selbst aufgebauten Versuchsanordnung werden mit der vorgegebenen Versuchsanordnung vermieden.

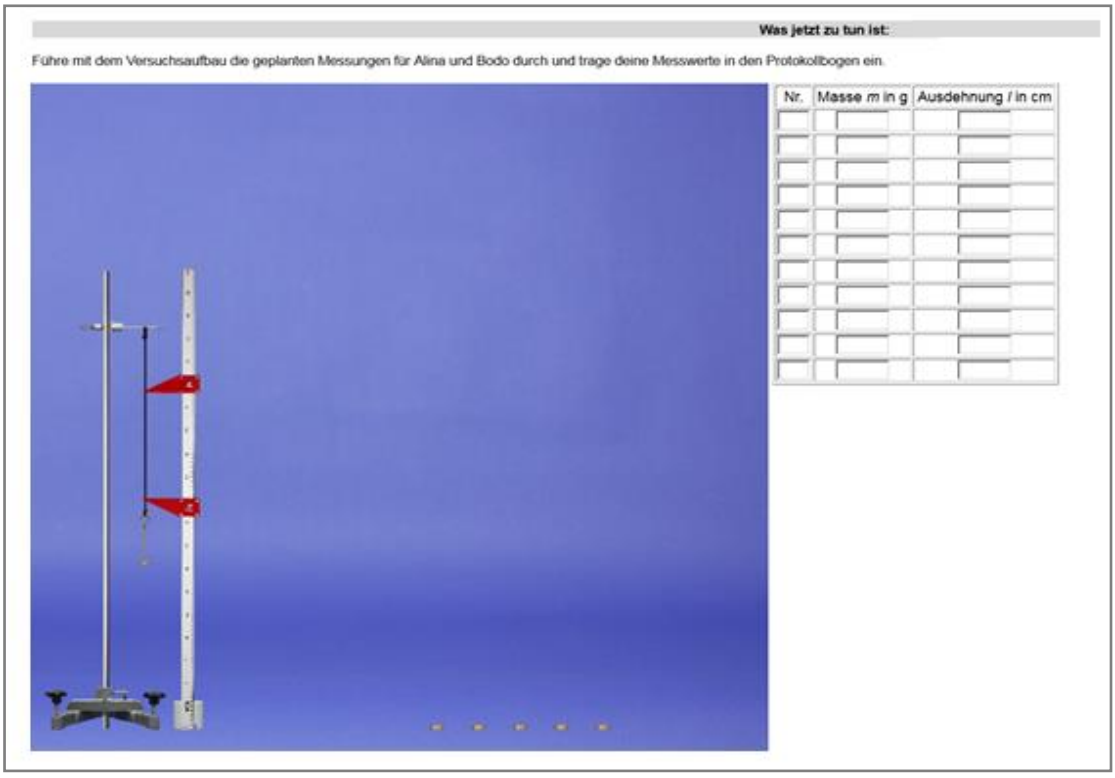


Abbildung 3.4: Teilaufgabe Messung durchführen und dokumentieren *der Aufgabe zur Ausdehnung eines Gummibandes*

Für die Teilaufgabe *Datenauswertung durchführen* steht u.a. ein interaktives Tool zur Erstellung von Messwertediagrammen zur Verfügung (vgl. Anhang A.1 auf Seite 182).

Alle Schülerantworten und virtuellen Handlungen mit der Computermaus (für jeden Mausklick) werden als Zeichenfolge mit Schülercode und Zeitpunkt in einer Datenbank auf einem Server gespeichert (Log-Daten).

3.2 MeK-LSA Experimentiertest: Materialien

Ausgehend von der in Abschnitt 3.1 beschriebenen Konzeption des MeK-LSA Experimentiertests sind insgesamt zwölf Aufgaben zu den physikalischen Inhaltsbereichen Mechanik, geometrische Optik und Elektrizitätslehre entwickelt worden. Die Überführung der Testkonzeption in konkrete Testaufgaben wird ausführlich in Kapitel 4 beschrieben. Zusätzlich zu den Testaufgaben wurde eine Aufgabe aus dem Bereich der Elektrizitätslehre als Trainingsaufgabe konzipiert. Durch die Trainingsaufgabe lernen die Schülerinnen und Schüler das Aufgabenformat kennen (Testbedienung, Erwartungshorizont). Erprobungen der Testaufgaben zeigen, dass eine Bearbeitungszeit von 25 Minuten für eine Aufgabe mit sechs Teilaufgaben angemessen ist. In Abschnitt 4.7 ist eine tabellarische Übersicht der zwölf Testaufgaben und der Trainingsaufgabe zu finden. Neben den zwölf Testaufgaben und der Trainingsaufgabe gehören zu den Materialien des MeK-LSA Experimentiertests ein Manual mit Verhaltens- und Ablaufanweisungen für die Testleiter, um eine objektive Testdurchführung zu gewährleisten, ein Ablaufskript für die Durchführung der Trainingsaufgabe, Protokollbögen zur Dokumentation von (technischen) Problemen und Auffälligkeiten bei der Testdurchführung sowie ein Auswertungsverfahren zur Bewertung der Schülerlösungen.

Das Auswertungsverfahren zur Bewertung der Schülerlösungen wird im folgenden Abschnitt beschrieben.

3.3 MeK-LSA Experimentiertest: Auswertungsverfahren

Die in der Datenbank als Zeichenfolgen gespeicherten Log-Daten der Schülerlösungen (vgl. Abschnitt 3.1) lassen sich über ein Software-Tool auslesen. In dem Tool kann der für jede Teilaufgabe und jeden Schülercode gespeicherte Endzustand aufgerufen und in Form von Screenshots angezeigt werden (vgl. Abbildung 3.5). Für die interaktiven Elemente können zusätzlich auch alle gespeicherten Zwischenzustände in chronologischer Reihenfolge aufgerufen und angezeigt werden (vgl. Abbildung 3.5).

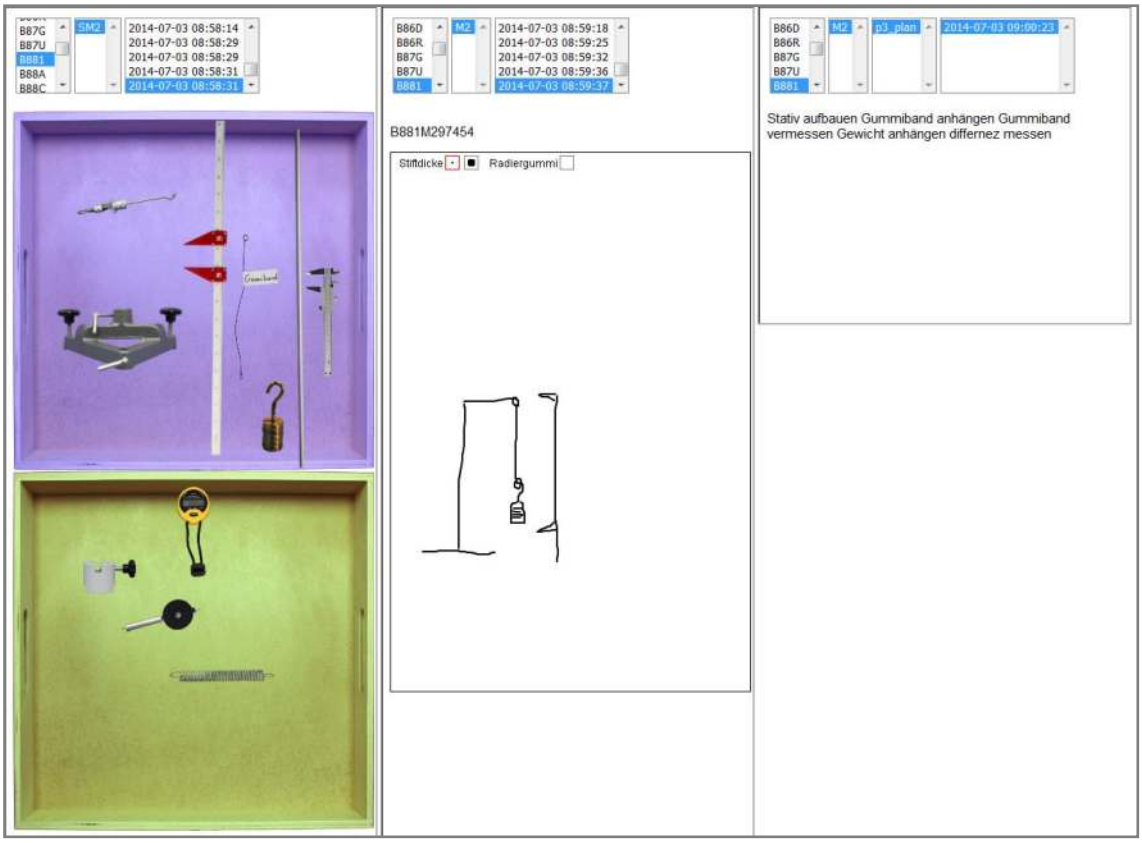


Abbildung 3.5: Software-Tool zur Bewertung der Schülerlösungen bei der Teilaufgabe Versuchsplan entwerfen*; oben: Drop-Down Menü zur Auswahl von Schülercode und Zwischenzustand; links: Geräteauswahl; Mitte: Skizze; rechts: Vorgehensweise.*

Das Software-Tool ermöglicht eine teilautomatisierte und in der Folge zeitökonomische Bewertung der Schülerlösungen (Aufwand für geschulte Rater: ca. drei Minuten für jede Testaufgabe mit sechs Teilaufgaben). Für die Teilaufgabe *Messung durchführen und dokumentieren* konnte darüber hinaus eine automatische Auswertungsroutine entwickelt werden, die insbesondere prüft, ob die von den Schülerinnen und Schülern gemessenen Werte mit den dokumentierten Werten übereinstimmen (vgl. Eickhorst, in Vorbereitung). Die Bewertung der Schülerlösungen erfolgt über die im Software-Tool dargestellten Log-Daten[4].

[4] Für die automatische Auswertungsroutine zur Teilaufgabe *Messung durchführen und dokumentieren* wird unmittelbar auf die Log-Daten in der Datenbank und nicht auf das Software-Tool zurückgegriffen.

In der Regel wird der Endzustand bewertet. Zur Bewertung der Schülerlösungen wurde für den MeK-LSA Experimentiertest ein detailliertes Kodierhandbuch mit Bewertungsmaßstäben für jede Teilaufgabe ausgearbeitet (vgl. Theyßen et al., 2016b). Bei den Teilaufgaben *Versuchsplan entwerfen*, *Versuch aufbauen und testen* und *Messung durchführen und dokumentieren* sind drei Bewertungsstufen vorgesehen (geeignet, teilweise geeignet und ungeeignet). Tabelle 3.1 zeigt den Bewertungsmaßstab für die Teilaufgabe *Versuchsplan entwerfen* der Aufgabe *Ausdehnung eines Gummibandes*. Bei Teilaufgaben zum *Versuchsplan entwerfen* liegt eine ungeeignete Lösung vor, wenn kein experimentbezogener Lösungsansatz erkennbar ist, z. B. weil keine Verwendung von Gewichten angedeutet ist. Eine teilweise geeignete Lösung liegt vor, wenn zwar ein experimentbezogener Lösungsansatz erkennbar ist (z. B. Gummiband hängt senkrecht nach unten und Verwendung von Gewichten ist angedeutet), das Experiment mit diesem Ansatz jedoch nicht durchführbar wäre, z. B. weil die Ausdehnung nicht genau gemessen werden kann. Eine geeignete Lösung liegt dagegen vor, wenn das Experiment mit dem entworfenen Versuchsplan durchführbar wäre (vgl. Abbildung 3.5 auf Seite 34). Die Teilaufgaben *Versuch aufbauen und testen* und *Messung durchführen und dokumentieren* werden auf analoge Weise bewertet.

Tabelle 3.1: Auszug aus dem Kodierhandbuch für die Teilaufgabe Versuchsplan entwerfen *der Aufgabe* Ausdehnung eines Gummibandes *(vereinfachte Darstellung); ODER: mindestens ein Kriterium muss erfüllt sein; UND: beide Kriterien müssen erfüllt sein.*

Aufgabenstellung	**Experimentell überprüfen, ob die Ausdehnung eines Gummibands proportional zur Masse der angehängten Gewichtsstücke ist.**
Teilaufgabe	Versuchsplan entwerfen
vorhandene Bauteile	Gummiband, Gewichte, Feder, Rolle, Stoppuhr (Längen-)Messgeräte: Maßstab, Messschieber Stativmaterial: Stativstange, Haken, Stativfuß (groß), Tonnenfuß (klein)
Kriterien für die Qualität der Bearbeitung	
ungeeignet	ODER: • Gummiband hängt *nicht* senkrecht nach unten. • Verwendung von Gewichten nicht angedeutet.
teilweise geeignet	UND: • Gummiband hängt senkrecht nach unten • Verwendung von Gewichten angedeutet.
geeignet	Zusätzlich: ODER: • Gummiband mit Stativmaterial an einer Seite fixiert • Messgerät mit Stativmaterial fixiert

Die Bewertungskriterien für die Teilaufgaben *Versuchsplan entwerfen*, *Versuch aufbauen und testen* und *Messung durchführen und dokumentieren* sind nicht als Kriterien rein fachlicher Korrektheit zu klassifizieren. Besonders bei der Unterscheidung zwischen ungeeigneten und teilweise geeigneten Lösungen ist das Vorhandensein eines experimentbezogenen Lösungsansatzes entscheidend und nicht die fachliche Richtigkeit des Ansatzes. Die fachliche

Richtigkeit ist erst bei der Unterscheidung von teilweise geeigneten und geeigneten Lösungen entscheidend.

Die Bewertung für die Teilaufgaben zu *Datenauswertung durchführen*[5], in denen die Schülerinnen und Schüler ein Messwertediagramm erstellen müssen, erfolgt ebenfalls dreistufig, jedoch nach einem anderen System. Zunächst werden die folgenden drei Aspekte separat bewertet: Koordinatensystem mit Beschriftung, Skalierung der Achsen, Messpunkte korrekt eingetragen. Ähnliche Bewertungsaspekte finden sich beispielsweise bei Lachmayer (2008), die ein Strukturmodell *Diagrammkompetenz* für den Biologieunterricht entwickelt und überprüft hat. Für die Messwertediagramme im MeK-LSA Experimentiertest werden die Bewertungsaspekte wie folgt zu einem dreistufigen Bewertungsmaßstab verknüpft: keine der Bewertungskategorien ist erfüllt (0 Punkte), eine Kategorie ist erfüllt (1 Punkt), mindestens zwei Kategorien sind erfüllt (2 Punkte).

Alle weiteren Teilaufgaben des MeK-LSA Experimentiertests werden dichotom (geeignet; ungeeignet) bewertet (vgl. Tabelle 3.2).

Tabelle 3.2: Bewertungskriterien für die dichotom ausgewerteten Teilaufgaben (vereinfachte Darstellung)

Teilaufgabe	**Lösung geeignet, wenn …**
Grundidee beschreiben	… die zu messenden Größen und die zu variierende Größe richtig benannt werden.
Messprotokoll vorbereiten	… die Messgrößen erkennbar sind und für jeden Untersuchungsgegenstand mind. ein Wert bzw. ein Wertepaar notiert werden kann.
Datenauswertung planen	… mind. zwei der folgenden drei Kriterien erfüllt sind: Teilaufgabenziel benannt; Bezug zu Messgrößen hergestellt; geeignetes Auswertungsverfahren beschrieben.
Datenauswertung durchführen (Berechnungen ausführen)	… wenn alle Werte richtig berechnet werden.
Schlüsse ziehen	… wenn die Vermutung richtigerweise verifiziert oder falsifiziert wird und bei der Begründung Bezug zu einem funktionalen Zusammenhang oder den Messdaten hergestellt wird.

Um die Objektivität des Kodiervorgangs abzusichern, wurden für jede teilautomatisch ausgewertete Teilaufgabe mindestens 10 % aller Daten von zwei geschulten Kodierern unabhängig voneinander bewertet. Die Urteilerübereinstimmung (Cohens Kappa; vgl. Wirtz & Caspar, 2002, S. 55-67) ist zufriedenstellend bis sehr gut ($.63 < \kappa < 1$, mittleres $\kappa = .84$).

[5] Die Teilaufgabe *Datenauswertung durchführen* umfasst in der Regel entweder das Erstellen eines Messwertediagramms oder das Ausführen von mathematischen Berechnungen (z. B. Addition von Messwerten).

4 MeK-LSA Experimentiertest: Überführung der Testkonzeption in Testaufgaben

Für den MeK-LSA Experimentiertest wurde die Testkonzeption (vgl. Abschnitt 3.1 auf Seite 29) systematisch in konkrete Testaufgaben überführt. Neben normativen Entscheidungen des MeK-LSA Projektteams[6] umfasste die Überführung der Testkonzeption in konkrete Testaufgaben insbesondere kriteriengeleitete Dokumentanalysen und Einschätzungen von Expertinnen und Experten, die nicht operativ am Projekt MeK-LSA beteiligt waren. Das Flussdiagramm in Abbildung 4.1 veranschaulicht den Überführungsprozess der Testkonzeption in konkrete Testaufgaben für den MeK-LSA Experimentiertest. Der Überführungsprozess wird in den folgenden Abschnitten detailliert beschrieben.

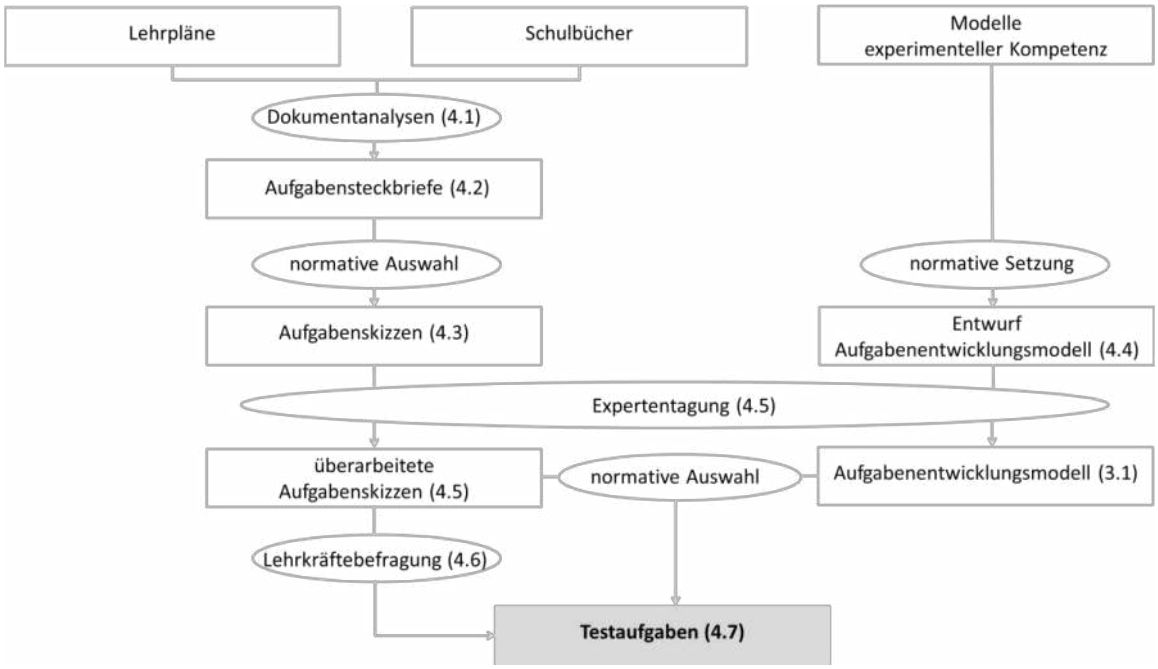


Abbildung 4.1: Flussdiagramm zur Überführung der Testkonzeption in konkrete Testaufgaben für den MeK-LSA Experimentiertest (ovale Komponenten: Prozesse zur Produkterstellung; rechteckige Komponenten: Produkte; Zahlen in Klammern: Abschnittsnummern)

4.1 Dokumentanalysen

Bildungsstandards geben einen verbindlichen Rahmen für Kompetenzerwartungen vor, *„ohne diese mit differenzierten inhaltlichen Vorgaben zu verknüpfen“* (Ropohl, Sumfleth & Walpuski, 2014, S. 7). Sie stellen daher keinen geeigneten Ausgangspunkt für die Bestimmung von Testinhalten für kompetenzorientierte Testverfahren dar. Erst auf der Ebene von Lehrplänen und Schulbüchern findet eine Verknüpfung von Kompetenzerwartungen und Inhalten statt (ebenda). Lehrpläne und Schulbücher werden in Deutschland von Physiklehrkräften häufig zur Unterrichtsvorbereitung genutzt (Härtig, Kauertz & Fischer, 2012, S. 198). Beide Quellen sind folglich ein geeigneter Ausgangspunkt für die Auswahl von Testinhalten für den MeK-LSA Experimentiertest. In einem ersten Schritt wurden durch eine Lehrplananalyse (vgl. Abschnitte 4.1.1 bis 4.1.3) Unterrichtsthemen identifiziert, die in möglichst vielen Bundesländern verbindlich unterrichtet werden sollen. Schulbücher gelten als Vermittler

[6] Im Folgenden auch als Testentwicklungsteam bezeichnet.

zwischen vorgesehenem und tatsächlich umgesetztem Lehrplaninhalt, da diese die Ausgestaltung der Unterrichtsthemen weiter konkretisieren (Vollstädt, Tillmann, Rauin, Höhmann & Tebrügge, 1999; Härtig, 2010; Härtig et al., 2012). Im Anschluss an die Lehrplananalyse ermöglichte eine Schulbuchanalyse (vgl. Abschnitte 4.1.4 bis 4.1.6) daher die Verknüpfung von lehrplan-verbindlichen Unterrichtsthemen mit typischen experimentellen Aufgabenstellungen.

4.1.1 Lehrplananalyse: Datenbasis

Die Lehrplananalyse für das Unterrichtsfach Physik in der Sekundarstufe I wurde im Schuljahr 2011/2012 durchgeführt. Die in diesem Schuljahr in den 16 Bundesländern gültigen Lehrpläne wurden über die Lehrplandatenbank des Sekretariats der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland ermittelt. Ziel der Analyse war es, physikalische Unterrichtsthemen zu identifizieren, die in möglichst vielen Bundesländern bis zum Ende der Sekundarstufe I verbindlich unterrichtet werden. Zur Sekundarstufe I zählen in der Regel die Jahrgangsstufen 5 bis 9 (achtjährige Gymnasien) bzw. 5 bis 10 (Schulen des Sekundarbereichs I und neunjährige Gymnasien). Aufgrund der Vielzahl länder- und schulformspezifischer Lehrpläne wurden zunächst Lehrpläne für das Gymnasium analysiert. Da in neun Bundesländern Physik in den Jahrgangstufen 5 bis 6 oder 5 bis 7 nicht als eigenständiges Fach unterrichtet wird, sind in diesen Bundesländern auch die Lehrpläne für das Fach Naturwissenschaften berücksichtigt worden. Eine Sonderrolle nehmen Berlin und Brandenburg ein. Hier zählen die Jahrgangsstufen 5 und 6 noch zur Grundschule. In beiden Ländern wurden für diese Jahrgangsstufen daher auch die Grundschullehrpläne für das Fach Naturwissenschaften berücksichtigt. In Anhang A.2 auf Seite 184 sind die analysierten Lehrpläne aufgeführt.

4.1.2 Lehrplananalyse: Vorgehensweise

In einem ersten Schritt wurden induktiv Begriffe ermittelt, mit denen in den Lehrplänen physikalische Sachverhalte benannt werden. Als Begriffe werden hier einzelne Nomen, durch Adjektive ergänzte Nomen oder Kombinationen mehrerer Nomen bezeichnet. Aufgrund länderspezifischer Sprachregelungen und Schwerpunktsetzungen werden in den Lehrplänen verschiedene Begriffe verwendet, die sehr ähnliche physikalische Sachverhalte beschreiben oder zu denen sehr ähnliche Experimente vorgeschlagen werden. Um eine bessere Vergleichbarkeit zwischen den Ländern zu gewährleisten, wurden die Begriffe zu Unterrichtsthemen gruppiert. Zwei Kodierer haben unabhängig voneinander die Begriffe (N=166) den Unterrichtsthemen zugeordnet. Für die Übereinstimmung der Kodierer ergeben sich getrennt nach den Inhaltsbereichen Mechanik, Optik, Elektrizitätslehre und Wärmelehre mindestens zufriedenstellende Werte ($.79 < \kappa < .95$; mittleres $\kappa = .86$). Tabelle 4.1 auf Seite 39 zeigt die Zusammenfassung der Begriffe beispielhaft für das Unterrichtsthema *Kraft und Verformung*.

Tabelle 4.1: Begriffe zum Unterrichtsthema Kraft und Verformung

Unterrichtsthema	Begriffe
Kraft und Verformung	Kraft und Verformung
	Kraft und Auslenkung
	Kraft und Formänderung
	Dehnung, Verformung, Verlängerung (Feder / Gummiband)
	Wirkungen von Kräften
	Kraftmessung
	Hookesches Gesetz

Die Lehrpläne wurden im Hinblick auf die Unterrichtsthemen systematisch und kriteriengeleitet durchsucht. Es wurde erfasst,

a) ob das Unterrichtsthema generell vorhanden ist,
b) ob das Unterrichtsthema im Zusammenhang mit dem Planen, Durchführen oder Auswerten von Experimenten vorkommt,
c) bis zu welcher Jahrgangsstufe das Unterrichtsthema genannt wird,
d) ob das Unterrichtsthema verbindlicher oder optionaler Bestandteil des Lehrplans ist.

Für die Entwicklung der Testaufgaben sind die Unterrichtsthemen von Interesse, die in möglichst vielen Bundesländern bis zum Ende der Sekundarstufe I verbindlich unterrichtet werden sollen. Im Hinblick auf die Messung experimenteller Kompetenz sollten die Lehrpläne diese Themen im Idealfall verbindlich im Zusammenhang mit dem Planen, Durchführen oder Auswerten von Experimenten nennen.

4.1.3 Lehrplananalyse: Ergebnisse

Die Ergebnisse der Lehrplananalyse zeigen, dass es Unterrichtsthemen gibt, die bis zum Ende der Sekundarstufe I in mindestens zwölf von 16 Bundesländern verbindlich unterrichtet werden sollen (vgl. Tabelle 4.2 auf Seite 40). Ein Unterrichtsthema, das dieses Kriterium erfüllt, wird im Folgenden als *relevantes Unterrichtsthema* bezeichnet. Dieser Anteil liegt noch deutlich über dem Kriterium der *TIMS-Studie* (national: 60 %, international: 50 %; vgl. Baumert, Klieme, Lehrke & Savelsbergh, 2000, S. 7). Die als relevant identifizierten Unterrichtsthemen lassen sich den Inhaltsbereichen Elektrizitätslehre, Mechanik, geometrische Optik und Wärmelehre zuordnen. Unterrichtsthemen aus dem Inhaltsbereich Akustik spielen bundesweit nur eine untergeordnete Rolle. Eine Übersicht über alle identifizierten Unterrichtsthemen findet sich in Anhang A.3 auf Seite 185. Für die verbindliche Nennung von Unterrichtsthemen im Zusammenhang mit dem Planen, Durchführen oder Auswerten von Experimenten zeigt sich kein eindeutiges Bild. Das ist u.a. darauf zurückzuführen, dass durch die zunehmende Einführung von Kernlehrplänen (z. B. MSW NRW, 2008) nur selten konkrete Vorschläge für die Verknüpfung von Inhalten und Experimenten zu finden sind. Stattdessen werden übergreifende Experimentierfähigkeiten benannt. Aus der Lehrplananalyse lässt sich somit

nicht verlässlich ableiten, welche Experimente typischerweise zu den relevanten Unterrichtsthemen durchgeführt werden sollen. Es bleibt hier weitgehend den Schulen bzw. den Lehrkräften überlassen, an welchen Inhalten experimentelle Kompetenz zu vermitteln ist. Daher wurden in einem zweiten Schritt Schulbücher analysiert.

Tabelle 4.2: relevante Inhaltsbereiche und Unterrichtsthemen (Anzahl verbindlich: Anzahl der Bundesländer in denen das Unterrichtsthema verbindlich bis zum Ende des Sekundarstufe I unterrichtet werden soll; Anzahl verbindlich mit Experiment: Anzahl der Bundesländer in denen das Unterrichtsthema verbindlich im Zusammenhang mit dem Planen, Durchführen oder Auswerten von Experimenten bis zum Ende der Sekundarstufe I unterrichtet werden soll)

Inhaltsbereich	**Unterrichtsthema**	**Anzahl verbindlich I**	**Anzahl verbindlich mit Experiment**
Elektrizitätslehre	elektrischer Widerstand	15	8
	Reihenschaltung	14	8
	Parallelschaltung	14	9
	elektrische Leistung	13	3
	Leiter und Nichtleiter	13	4
Mechanik	Kraft und Verformung[7]	16	10
	Auftrieb in Flüssigkeiten	13	6
	Dichtebestimmung	14	7
	Bewegungen	15	11
	mechanische Arbeit	12	2
Optik	Reflexionsgesetz	16	9
	Brechungsgesetz	16	8
	Abbildungen an Linsen	14	9
Wärmelehre	Wärmeübertragung	12	5
	Aggregatzustandsänderungen	14	6

4.1.4 Schulbuchanalyse: Datenbasis

Um zu identifizieren, welche Experimente typischerweise zu den als relevant identifizierten Unterrichtsthemen (vgl. Abschnitt 4.1.3) durchgeführt werden, wurde eine Schulbuchanalyse durchgeführt. Für die Schulbuchanalyse wurden acht Schulbücher für das Unterrichtsfach Physik in der Sekundarstufe I für das Gymnasium in Nordrhein-Westfalen ausgewählt. Insgesamt wurden fünf Schulbuchreihen von vier Schulbuchverlagen analysiert, da die Verlage *„häufig eigene Wege der Interpretation und der Auslegung von Lehrplanvorgaben“* (Mikelskis, 2006, S. 46) gehen. Tabelle 4.3 auf Seite 41 zeigt die analysierten Schulbücher.

[7] Das Unterrichtsthema *Kraft und Verformung* schließt im Gegensatz zu den in Dickmann & Theyßen (2013, S. 588) berichteten Werten auch das Unterrichtsthema *Hookesches Gesetz* mit ein. Daher weichen die Werte hier leicht ab. An der inhaltlichen Aussage ändert sich nichts.

Tabelle 4.3: analysierte Schulbücher zur Identifikation typischer Experimente

Titel	Fokus Physik		Impulse Physik 1	Impulse Physik 2	Dorn Bader Physik 1	Dorn Bader Physik 2	Kuhn Physik 1	Physik für Gymnasien
Verlag	Cornelsen		Klett		Schroedel		Westermann	Cornelsen
Klassen	5-6	7-9	5-6	7-9	5-6	7-9	7-10	5-10

Die Beschränkung auf Länderausgaben aus Nordrhein-Westfalen erscheint zulässig, da es nach Härtig (2010, S. 54) nur selten zur Überarbeitung einzelner Abschnitte bei unterschiedlichen Länderausgaben kommt. In einer stichprobenartigen Überprüfung weiterer Länderausgaben wurden keine Hinweise gefunden, die gegen das Beibehalten dieser Annahme sprechen.

4.1.5 Schulbuchanalyse: Vorgehensweise

Auf Basis der Registerverzeichnisse der Schulbücher wurde erfasst, ob die als relevant identifizierten Unterrichtsthemen (vgl. Tabelle 4.2 auf Seite 40) in den Schulbüchern vorhanden sind. Darauf aufbauend wurden die Schulbücher nach Experimenten durchsucht, die sich diesen Unterrichtsthemen zuordnen lassen.

4.1.6 Schulbuchanalyse: Ergebnisse

Die Schulbuchanalyse zeigt, dass in den Schulbuchreihen verschiedener Verlage zu den relevanten Unterrichtsthemen in der Regel vergleichbare Experimente vorgeschlagen werden. Vergleichbare Experimente zeichnen sich durch sehr ähnliche experimentelle Anforderungen bei der Durchführung des Experiments aus (z. B. eine Feder und/oder ein Gummiband der Reihe nach mit mehreren Gewichtsstücken belasten und jeweils die Ausdehnung messen). Solche vergleichbaren Experimente werden im Folgenden als *typisch* für das jeweilige Unterrichtsthema bezeichnet. Für das relevante Unterrichtsthema *Kraft und Verformung* findet sich z. B. das folgende typische Experiment (genauer: experimentelle Aufgabenstellung):

> *„Hänge an eine Schraubenfeder der Reihe nach Körper verschiedener Gewichtskraft und miss die Verlängerungen […]. Wiederhole den Versuch mit einer zweiten Feder und mit einem Gummifaden."* (Kuhn, 2008, S. 108)

4.2 Aufgabensteckbriefe

Die Ergebnisse der Lehrplan- und Schulbuchanalyse (vgl. Abschnitte 4.1.1 bis 4.1.6) wurden für die identifizierten relevanten Unterrichtsthemen in je einem Aufgabensteckbrief zusammengestellt. Abbildung 4.2 auf Seite 42 zeigt beispielhaft den Aufgabensteckbrief für das Unterrichtsthema *Kraft und Verformung*. Zusätzlich zur Schulbuch- und Lehrplananalyse wurde für die als typisch identifizierten Experimente bei fünf Lehrmittelherstellern überprüft, ob Schülerexperimentiermaterial zur Durchführung des Experiments angeboten wird. Die Ergebnisse dieser Überprüfung sind auch in den Aufgabensteckbriefen dargestellt. Die Aufgabensteckbriefe bildeten die Grundlage für die Ausarbeitung von Aufgabenskizzen durch das Testentwicklungsteam.

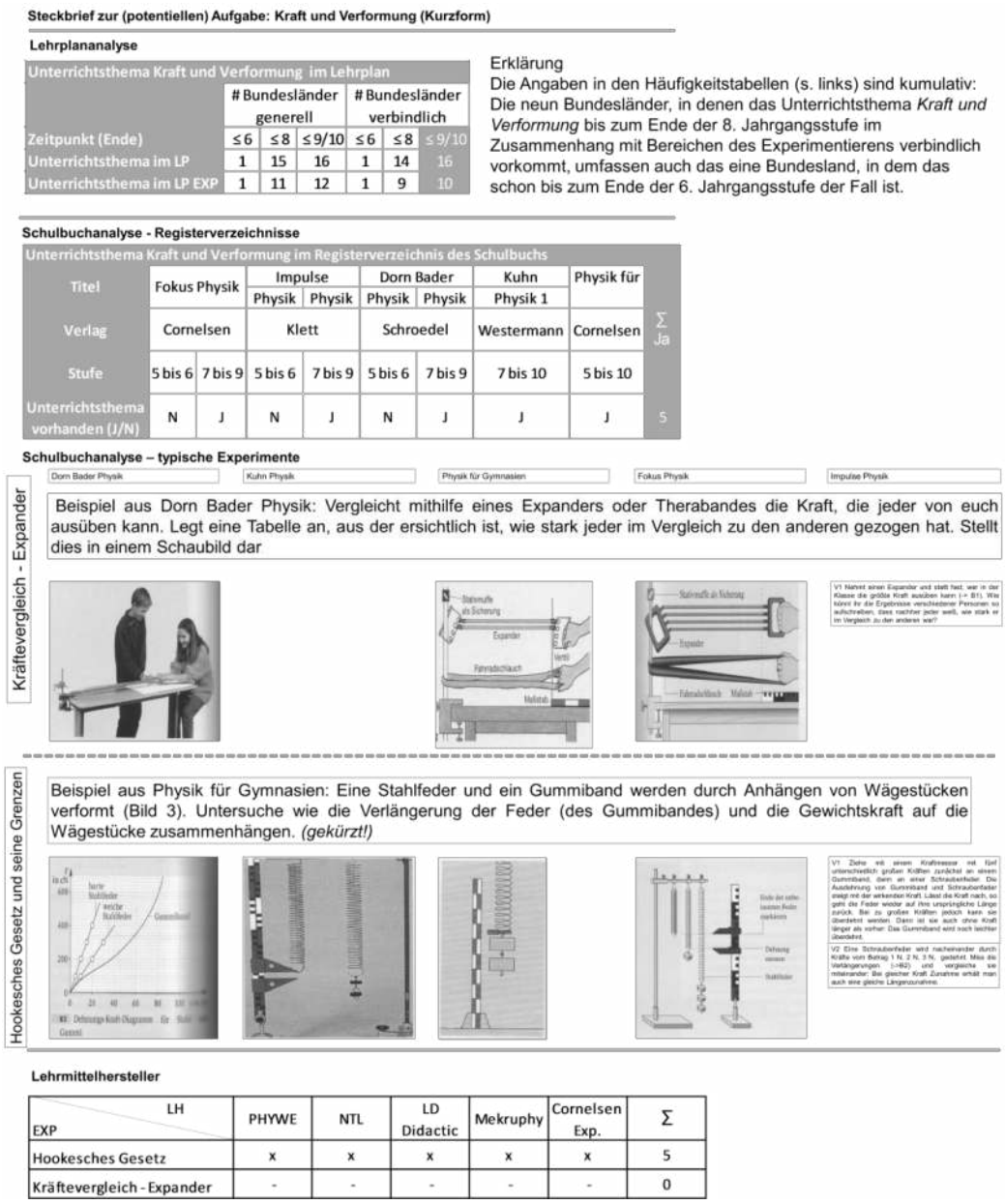


Abbildung 4.2: Aufgabensteckbrief zum relevanten Unterrichtsthema Kraft und Verformung *(Kurzform)*

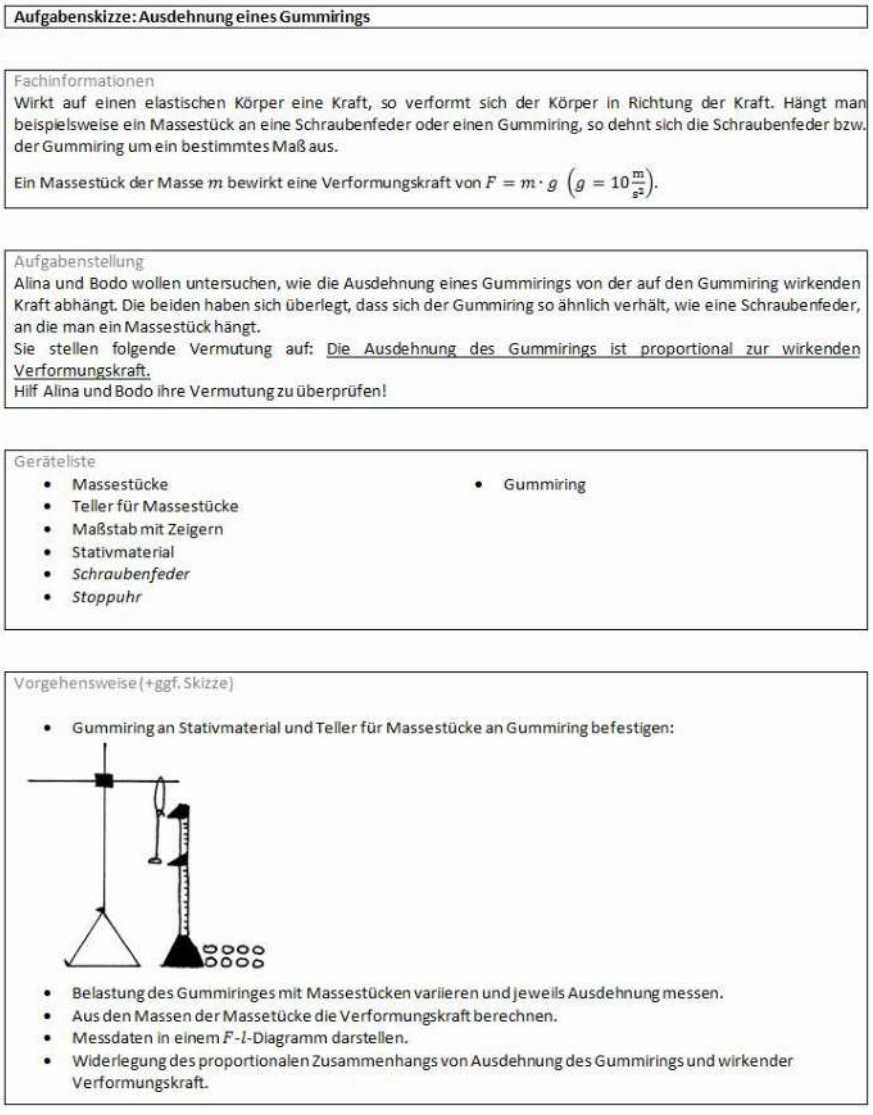
4.3 Aufgabenskizzen

Auf Basis der Aufgabensteckbriefe (vgl. Abschnitt 4.2) wurden Aufgabenskizzen zu den vier Inhaltsbereichen Mechanik, geometrische Optik, Elektrizitätslehre und Wärmelehre durch das Testentwicklungsteam ausgearbeitet. Sie enthalten die zur Bearbeitung benötigten Fachinformationen, die experimentelle Aufgabenstellung, eine Geräteliste sowie die für eine erfolgreiche Bearbeitung erwartete Vorgehensweise (vgl. Beispiel in Abbildung 4.3 auf Seite 43). Wenn ein Aufgabensteckbrief mehrere Experimente bzw. experimentelle Aufgabenstellungen nahelegt, wurden gegebenenfalls auch mehrere Aufgabenskizzen zu einem Steckbrief erstellt (z. B. vier Aufgabenskizzen zum Aufgabensteckbrief *elektrischer Widerstand* aus dem Inhaltsbereich Elektrizitätslehre). Insgesamt wurden 22 Aufgabenskizzen ausgearbeitet. Tabelle 4.4 auf Seite 43 zeigt die Verteilung der Aufgabenskizzen auf die vier Inhaltsbereiche.

Tabelle 4.4: Anzahl der ausgearbeiteten Aufgabenskizzen pro Inhaltsbereich

Inhaltsbereich	Anzahl Aufgabenskizzen
Elektrizitätslehre	7
geometrische Optik	8
Mechanik	4
Wärmelehre	3

Eine Besonderheit liegt für den Inhaltsbereich geometrische Optik vor. Aufgrund der in den Aufgabensteckbriefen aufgeführten typischen Experimente wurden zwei unterschiedliche Arten von Aufgabenskizzen (jeweils vier) ausgearbeitet. Diese zwei Arten unterscheiden sich bezogen auf die zum Experimentieren verwendeten optischen Geräte: *Optikexperimente auf dem Tisch,* in denen Schnittmodelle von Objekten (z. B. Linsen) und Leuchtboxen auf dem Tisch verwendet werden und *Optikexperimente auf der optischen Bank,* in denen Objekte (z. B. Linsen) und Lampen auf der optischen Bank eingesetzt werden.



Aufgabenskizze: Ausdehnung eines Gummirings

Fachinformationen

Wirkt auf einen elastischen Körper eine Kraft, so verformt sich der Körper in Richtung der Kraft. Hängt man beispielsweise ein Massestück an eine Schraubenfeder oder einen Gummiring, so dehnt sich die Schraubenfeder bzw. der Gummiring um ein bestimmtes Maß aus.

Ein Massestück der Masse m bewirkt eine Verformungskraft von $F = m \cdot g$ $\left(g = 10\frac{m}{s^2}\right)$.

Aufgabenstellung

Alina und Bodo wollen untersuchen, wie die Ausdehnung eines Gummirings von der auf den Gummiring wirkenden Kraft abhängt. Die beiden haben sich überlegt, dass sich der Gummiring so ähnlich verhält, wie eine Schraubenfeder, an die man ein Massestück hängt.

Sie stellen folgende Vermutung auf: Die Ausdehnung des Gummirings ist proportional zur wirkenden Verformungskraft.

Hilf Alina und Bodo ihre Vermutung zu überprüfen!

Geräteliste

- Massestücke
- Teller für Massestücke
- Maßstab mit Zeigern
- Stativmaterial
- *Schraubenfeder*
- *Stoppuhr*
- Gummiring

Vorgehensweise (+ggf. Skizze)

- Gummiring an Stativmaterial und Teller für Massestücke an Gummiring befestigen:

- Belastung des Gummiringes mit Massestücken variieren und jeweils Ausdehnung messen.
- Aus den Massen der Massetücke die Verformungskraft berechnen.
- Messdaten in einem F-l-Diagramm darstellen.
- Widerlegung des proportionalen Zusammenhangs von Ausdehnung des Gummirings und wirkender Verformungskraft.

Abbildung 4.3: Aufgabenskizze zur Aufgabe Ausdehnung eines Gummirings

4.4 Aufgabenentwicklungsmodell: Entwurfsfassung

Parallel zur Ausarbeitung der Aufgabenskizzen (vgl. Abschnitt 4.3) hat das Testentwicklungsteam einen Entwurf für ein Aufgabenentwicklungsmodell erarbeitet. Ausgangspunkt der Erarbeitung waren das *exKomp-Modell* und das *Spinnennetzmodell* (vgl. Abschnitt 2.1.1 auf Seite 18). Abbildung 4.4 zeigt die Entwurfsfassung des Aufgabenentwicklungsmodells.

Abbildung 4.4: Entwurfsfassung des Aufgabenentwicklungsmodells für den MeK-LSA Experimentiertest

Im nächsten Schritt (Abschnitt 4.5) wurden Expertinnen und Experten in den Überführungsprozess der Testkonzeption in konkrete Testaufgaben eingebunden (vgl. Exkurs: Expertenbefragungen zu Testinhalten).

4.5 Expertentagung

Exkurs: Expertenbefragungen zu Testinhalten

Ein häufig eingesetztes Verfahren zur Einschätzung der *Relevanz und Repräsentativität* von Testinhalten (vgl. Kapitel 5) ist die Expertenbefragung. Die Grundidee besteht darin, dass Expertinnen und Experten die Inhalte der Testaufgaben hinsichtlich der Relevanz und der Passung zu den beabsichtigten Testzielen einschätzen. Die Ergebnisse der Einschätzung werden dann meist in Form einer quantitativen Zusammenfassung dargestellt (Sireci, 1998, S. 108). Für das Verfahren *Expertenbefragung* sieht Häder (2014, S. 64) die Gefahr einer geringen methodischen Absicherung, da keine einheitlichen Standards für Expertenbefragungen vorliegen. Ergebnisse aus diesen Befragungen werden daher auch nur selten (ebenda) oder zumindest nicht sehr detailliert berichtet. Bezogen auf die Entwicklung von Kompetenztests halten Jenßen, Dunekacke und Blömeke (2015) fest, dass häufig unklar bleibt, *„ob die inhaltliche Überführung des theoretischen Rahmens in Testaufgaben systematisch validiert wurde“* (S. 11) und empfehlen daher zur Validierung von Testinhalten *„systematische bzw. standardisierte Expertenbefragungen mit Nachweis der Expertise aller Beurteilerinnen und Beurteiler“* (S. 24).
Die Durchführung und Auswertung einer Expertenbefragung zur Relevanz und Repräsentativität von Testinhalten auf Ebene einzelner Testaufgaben sollte nach Jenßen et

al. (2015, S. 24-27) u.a. folgende Kriterien berücksichtigen:

- Die Befragung externer Expertinnen und Experten.
- Die Befragung einer heterogenen Expertengruppe zur Vermeidung von Beurteilungsfehlern aufseiten der Expertinnen und Experten und/oder der Testentwickler und Testentwicklerinnen.
- Einen Nachweis für die Expertise aller Expertinnen und Experten, um die Gefahr von Fehleinschätzungen zu minimieren.
- Eine transparente Darstellung der Zieldomäne (z. B. Experimentieren im Physikunterricht der Sekundarstufe I) und des beabsichtigten Verwendungszwecks (z. B. experimentelle Kompetenz im Large-Scale messen), um eine möglichst valide Beurteilung der Aufgaben zu ermöglichen.
- Eine vernetzte quantitative und qualitative Beurteilung der Aufgaben sowie (in der Folge) Auswertung der Befragung.
    - Quantitativ: Einschätzung der Relevanz und Repräsentativität der Testaufgaben mittels Ratingskalen und z. B. Bestimmung mittlerer Experteneinschätzungen.
    - Qualitativ: Einholen und Analyse von Anmerkungen, um Hinweise zur Überarbeitung oder Neuentwicklung von Aufgaben zu erhalten.
- Die Einbindung von Expertinnen und Experten in weitere Phasen der Testentwicklung (z. B. Modellentwicklung, Testzusammenstellung).

4.5.1 Expertentagung: Vorgehensweise

Im Anschluss an die Ausarbeitung der Aufgabenskizzen (vgl. Abschnitt 4.3) und die Erarbeitung eines Aufgabenentwicklungsmodells (Entwurfsfassung; vgl. Abschnitt 4.4) wurde eine Tagung mit elf Expertinnen und Experten durchgeführt. Ziel der Tagung war es,

1. anhand der Aufgabenskizzen die Eignung der Aufgabenthemen, der Aufgabenformulierungen und der Aufgabenanforderungen und
2. anhand prototypischer Aufgaben die Passung der Aufgaben zum Aufgabenentwicklungsmodell

zu diskutieren.

Keine bzw. keiner der elf Expertinnen und Experten war operativ am Projekt MeK-LSA beteiligt. Die Expertengruppe setzte sich zum einen aus erfahrenen Physiklehrkräften mit vertieftem physikdidaktischem Hintergrundwissen zusammen (z. B. durch Promotion in Physikdidaktik oder Tätigkeit als Referendarausbilder), zum anderen aus Physikdidaktikerinnen und Physikdidaktikern, die sich in ihren Forschungsarbeiten mit dem Experimentieren im Physikunterricht beschäftigen, zum Teil aus sehr unterschiedlichen Perspektiven.

Im Vorfeld der Tagung wurde den Expertinnen und Experten der Entwurf des Aufgabenentwicklungsmodells (vgl. Abbildung 4.4 auf Seite 44) zusammen mit ausgewählten Aufgabenskizzen (z. B. Abbildung 4.3 auf Seite 43) zur Verfügung gestellt. Die Tagung war in Form einer Gruppendiskussion organisiert (vgl. Häder, 2014), die durch

Kleingruppenarbeitsphasen ergänzt wurde. Eine Konsensfindung der Expertinnen und Experten war nicht beabsichtigt. Ziel war es vielmehr, Anregungen zur Verbesserung der Aufgaben und der Testkonzeption zu erhalten. Das Aufgabenentwicklungsmodell und die Aufgabenskizzen wurden intensiv und kritisch-konstruktiv diskutiert. Die Diskussionen wurden durch das MeK-LSA Projektteam dokumentiert und in einer zweitägigen Tagungsnachlese ausgewertet. Die zentralen Schlussfolgerungen aus der Expertentagung wurden den teilnehmenden Expertinnen und Experten schriftlich zurückgemeldet.

4.5.2 Expertentagung: Schlussfolgerungen zu Aufgabenskizzen und Aufgabenentwicklungsmodell

Die Rückmeldungen der Expertinnen und Experten haben zur Überarbeitung der Aufgabenskizzen und zur Überarbeitung des Aufgabenentwicklungsmodells beigetragen. Die zentrale Schlussfolgerung im Hinblick auf die Überarbeitung der Aufgabenskizzen lässt sich wie folgt zusammenfassen:

> *„Alle Aufgabentexte werden umfassend überarbeitet. Im Zweifel wird der Verständlichkeit gegenüber der Fachsprache Vorrang gegeben“* (Rückmeldung des MeK-LSA Projektteams an die Expertinnen und Experten).

Abbildung 4.5 zeigt beispielhaft die überarbeitete Version der Aufgabenskizze *Ausdehnung eines Gummirings*.

Aufgabenskizze: Ausdehnung eines Gummirings

Fachinformation

Kraft misst man in der Einheit Newton (N).

Die Kraft F, mit der ein Massestück an einer Spiralfeder oder einem Gummiband zieht, kann man mit folgender Formel berechnen: Man multipliziert die Masse m des Massestücks mit g=10m/s² ($F = m \cdot g$). Beispiel: Ein Massestück der Masse m=0,1kg zieht mit einer Kraft von 1 N an einem Gummiband ($F = 0{,}1\ kg \cdot 10 m/s^2 = 1N$).

Proportional bedeutet: Wenn man in einem Diagramm die x-Achse für die Kraft wählt und die y-Achse für die Ausdehnung, dann lässt sich durch die eingezeichneten Messpunkte eine Gerade legen. Diese Gerade geht durch den Ursprung des Koordinatensystems.

Aufgabenstellung

Alina und Bodo wollen untersuchen, wie sich ein Gummiring ausdehnt, wenn man verschieden stark daran zieht.

Die beiden erwarten: Der Gummiring verhält sich wie eine Spiralfeder.

Sie vermuten daher: Die Ausdehnung eines Gummirings ist proportional zu der Kraft, mit der man daran zieht.

Hilf Alina und Bodo, ihre Vermutung zu überprüfen!

Vorgehensweise (+ggf. Skizze)

- Gummiring an Stativmaterial und Teller für Massestücke an Gummiring befestigen.
- Belastung des Gummiringes mit Massestücken variieren und jeweils Ausdehnung messen.
- Aus den Massen der Massetücke die Kraft berechnen.
- Messdaten in einem F-l- Diagramm darstellen.
- Widerlegung des proportionalen Zusammenhangs von Ausdehnung des Gummirings und wirkender Kraft.

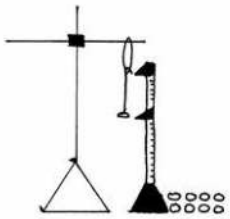


Abbildung 4.5: Aufgabenskizze zur Aufgabe Ausdehnung eines Gummirings *(überarbeitete Version auf Basis der Expertentagung)*

Im Vergleich zur ursprünglichen Aufgabenskizze (Abbildung 4.3 auf Seite 43) sind die Fachinformation und die Aufgabenstellung umformuliert worden (z. B. Berechnung der Kraft wird anhand eines Beispiels erklärt), mit der Absicht, die Verständlichkeit der Aufgabenstellungen für die Schülerinnen und Schüler zu erhöhen.[8]

Neben der Überarbeitung der Aufgabenskizzen haben die Expertenrückmeldungen auch zur Überarbeitung des Aufgabenentwicklungsmodells im Bereich der Planung beigetragen. Abbildung 4.6 auf Seite 48 zeigt die auf Basis der Expertenrückmeldungen im Bereich der Planung vorgenommenen Änderungen. Die Expertinnen und Experten kritisierten, dass die Fähigkeiten im Bereich der Planung unterschiedlich fein aufgelöst werden, wobei die Komponente *Vorgehensweise planen* im Vergleich zur Komponente *Geräte auswählen* sehr umfangreich ist. Die Expertinnen und Experten schlugen vor, die Komponente *Vorgehensweise planen* feiner aufzulösen und zum Beispiel als eine Komponente im Modell konkret danach zu fragen, welche Größen bei vorgegebener Fragestellung gemessen und welche variiert werden müssen (*Grundidee*). Die Bedeutung der Komponente *Geräte auswählen* wurde kritisch diskutiert. Einerseits könnte aus Sicht einzelner Expertinnen und Experten auf die Komponente vollständig verzichtet werden, andererseits passt der Ablauf des Modells mit Geräteauswahl aus Sicht anderer Expertinnen und Experten gut zur tatsächlichen Vorgehensweise im Unterricht. Ausgehend von diesen Expertenrückmeldungen ist der Bereich der Planung im Rahmen der Tagungsnachlese neu strukturiert worden (vgl. Abbildung 4.6 auf Seite 48). Das Verständnis der Aufgabenstellung wird implizit über die Komponente *Grundidee skizzieren* erfasst. Die Komponente *Geräteauswahl* wird mit den weiteren Aspekten der Komponente *Vorgehensweise planen* zur Komponente *Versuchsplan entwerfen* (Geräteauswahl, Skizze, Vorgehensweise) zusammengefasst. Die Komponente *Versuchsplan entwerfen* operationalisiert die Umsetzung der Grundidee, während die Komponente *Messprotokoll vorbereiten* einen vorgegebenen Versuchsplan operationalisiert. Die drei Komponenten im neu strukturierten Bereich der Planung ergeben folglich drei Ebenen der schrittweisen Konkretisierung der Planung. Der Auflösungsgrad der drei Komponenten unterscheidet sich auch im neu strukturierten Modell, wobei die Komponente *Versuchsplan entwerfen* am umfangreichsten ist (Geräteauswahl, Skizze, Vorgehensweise). Der unterschiedliche Auflösungsgrad lässt sich aber – im Gegensatz zur alten Struktur – durch die drei Konkretisierungsebenen, welche unterschiedlich umfangreiche Planungsschritte erfordern, inhaltlich sinnvoll begründen. Die Komponente *Versuchsplan entwerfen* als zentrale handlungsbezogene Planungskomponente wird in jeder Aufgabe berücksichtigt und aufgabenspezifisch durch die Komponenten *Grundidee skizzieren* oder *Messprotokoll vorbereiten* ergänzt. Eine Überarbeitung der Komponenten in den Bereichen Durchführung und Auswertung war aus Sicht der Expertinnen und Experten nicht erforderlich.

[8]Die genaue Formulierung der Aufgabenstellung und die Fachinformation zur Aufgabe *Ausdehnung eines Gummirings* haben sich im weiteren Projektverlauf weiterentwickelt (vgl. Abbildung 3.2 auf Seite 31).

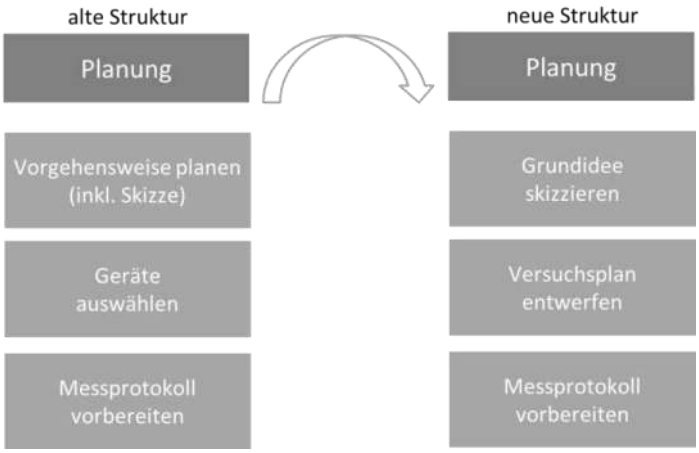


Abbildung 4.6: veränderte Struktur im Bereich der Planung (alt: links; neu: rechts)

4.6 Lehrkräftebefragung

Die auf Basis der Expertentagung (vgl. Abschnitt 4.5) überarbeiteten Aufgabenskizzen[9] bilden die Grundlage für eine Lehrkräftebefragung zur Relevanz und Repräsentativität der Testinhalte für das Experimentieren im Physikunterricht der Sekundarstufe I.

4.6.1 Lehrkräftebefragung: Vorgehensweise

Ziel der Befragung war,

1. die Überprüfung der Bekanntheit der in den Aufgabenskizzen beschriebenen Experimente für die Schülerinnen und Schüler aus Sicht der Lehrkräfte und
2. die Überprüfung der Passung zwischen den Anforderungen der Aufgabenskizzen und den experimentellen Anforderungen, die üblicherweise an Schülerinnen und Schüler im Physikunterricht der Sekundarstufe I gestellt werden.

Insgesamt wurden 53 Physiklehrkräfte (gymnasialer Bildungsgang: 43) aus neun Bundesländern befragt, die Unterrichtserfahrung in den für die Befragung relevanten Inhaltsbereichen haben. Die Befragung wurde als dezentrale Online-Befragung realisiert. Die Lehrkräfte waren über die Befragung hinaus nicht persönlich am Forschungsprojekt beteiligt und können daher als externe Expertinnen und Experten bezeichnet werden. Um Transparenz im Hinblick auf die Zieldomäne (*Experimentieren im Physikunterricht der Sekundarstufe I*) und den beabsichtigten Verwendungszweck (*Messung experimenteller Kompetenz im Large-Scale*) zu gewährleisten, erhielten die Lehrkräfte umfangreiche Projektinformationen. Um den zeitlichen Aufwand für die Lehrkräfte zu begrenzen, wurden sie nach dem Zufallsprinzip in zwei Gruppen eingeteilt. Den Lehrkräften jeder Gruppe wurden elf Aufgabenskizzen vorgelegt. Die Aufgabenskizzen enthielten die experimentelle Aufgabenstellung, die zur Bearbeitung benötigten Fachinformationen sowie die für eine erfolgreiche Bearbeitung erwartete Vorgehensweise. Abbildung 4.5 auf Seite 46 zeigt ein Beispiel für eine Aufgabenskizze. Darüber hinaus wurde den Lehrkräften vorgegeben, sich bei Ihrer Einschätzung auf *typische Schülerinnen und Schüler* (ohne nähere Erläuterung) der 9. Klasse an ihrer Schule zu beziehen.

[9] Im Folgenden zur besseren Lesbarkeit als Aufgabenskizzen bezeichnet. Gemeint sind weiterhin die überarbeiteten Aufgabenskizzen.

Auf Basis dieser Informationen schätzten die Lehrkräfte für jede Aufgabenskizze fünf Fragen auf einer vierstufigen Rating-Skala (1 ≙ sehr unwahrscheinlich bis 4 ≙ sehr wahrscheinlich) ein (vgl. Abbildung 4.7).

Die ersten beiden Fragen beziehen sich auf die Bekanntheit der Experimente für die Schülerinnen und Schüler aus Lehrkraftperspektive. Die letzten drei Fragen zielen auf die Erfüllbarkeit der experimentellen Anforderungen in den Bereichen Planung, Durchführung und Auswertung ab. Um konkrete Hinweise auf mögliche Schülerschwierigkeiten und Anhaltspunkte zur weiteren Überarbeitung der Aufgabenskizzen zu erhalten, wurde zusätzlich danach gefragt, woran es liegt, dass Schülerinnen oder Schüler das Experiment nicht oder eher nicht durchführen können. Dabei wurden bei jeder Aufgabenskizze als Antwortalternativen *mangelndes Fachwissen* und *mangelnde experimentelle Erfahrung* vorgegeben. Ergänzt wurde diese Auflistung durch aufgabenspezifische *Knackpunkte* und ein offenes Feld für sonstige Anmerkungen. Ein *Knackpunkt* ist in diesem Zusammenhang ein entscheidender Aspekt einer Aufgabe, von dem die erfolgreiche Bearbeitung der Aufgabe abhängt (z. B. Schülerinnen und Schüler können mit dem Begriff *proportional* trotz Erklärung in der Fachinformation nicht umgehen).

	sehr wahrscheinlich	wahrscheinlich	unwahrscheinlich	sehr unwahrscheinlich	kann ich nicht einschätzen
Wie wahrscheinlich haben Schülerinnen und Schüler dieses oder ein sehr ähnliches Experiment selbst durchgeführt?	○	○	○	○	○
Wie wahrscheinlich haben Schülerinnen und Schüler dieses oder ein sehr ähnliches Experiment gesehen (ohne es selbst durchzuführen)?	○	○	○	○	○
Stellen Sie sich vor, Schülerinnen und Schüler sollen das Experiment am Ende der 9. Klasse als Realexperiment durchführen. Wie wahrscheinlich können Schülerinnen und Schüler dieses Experiment...					
...planen:	○	○	○	○	○
...durchführen:	○	○	○	○	○
...auswerten:	○	○	○	○	○

Abbildung 4.7: Fragen, die von den Lehrkräften zu jeder vorgelegten Aufgabenskizze zu beantworten waren (erste und zweite Frage: Bekanntheit der Aufgaben für Schülerinnen und Schüler aus Lehrkraftperspektive; dritte bis fünfte Frage: Erfüllbarkeit der Anforderungen in den Bereichen Planung, Durchführung und Auswertung)

Die in der Befragung eingesetzte Rating-Skala kann als quasi-intervallskaliert angenommen werden, sodass eine Auswertung der Aufgabenskizzen anhand der mittleren Experteneinschätzung erfolgte. Für jede Aufgabenskizze und jede der fünf oben beschriebenen Fragen wird die mittlere Experteneinschätzung bestimmt.

4.6.2 Lehrkräftebefragung: Ergebnisse zu den Aufgabenskizzen

Die aus Sicht des Testentwicklungsteams wichtigste Lehrkräfteeinschätzung für die Auswahl von Aufgaben für den MeK-LSA Experimentiertest ist die Einschätzung zur Erfüllbarkeit der experimentellen Anforderungen im Bereich der Durchführung, da die Teilaufgaben *Versuch aufbauen und testen* und *Messung durchführen und dokumentieren* im Zentrum des Testverfahrens stehen. Bei einer eher positiven Einschätzung bezüglich der Frage zur Durchführbarkeit des Experiments (*Wie wahrscheinlich können Schülerinnen und Schüler dieses Experiment durchführen?*; im Folgenden als Kurzform: *Durchführbarkeit?*) wird angenommen,

dass die Schülerinnen und Schüler im Unterricht Lerngelegenheiten hatten, die es ihnen ermöglichen, die Aufgaben erfolgreich zu bewältigen. Abbildung 4.8 zeigt die mittlere Lehrkräfteeinschätzung für die Frage zur Durchführbarkeit des Experiments getrennt nach Aufgabenskizzen.

Die mittlere Lehrkräfteeinschätzung liegt für alle Experimente bezüglich der Durchführbarkeit deutlich über dem neutralen Wert von 2,5 auf der vierstufigen Rating-Skala (mindestens: 2,7; maximal: 3,4). Dreizehn Experimente liegen sogar bei mindestens 3,0, davon jeweils drei Experimente aus den Inhaltsbereichen geometrische Optik, Mechanik und Wärmelehre und vier Experimente aus dem Inhaltsbereich Elektrizitätslehre. Betrachtet man jeweils die untere und obere Grenze des 95 %-Konfidenzintervalls, so liegt die untere Grenze des Konfidenzintervalls für die Experimente *Reflexion am Wölbspiegel* und *Lochkamera* unter einem Wert von 2,5. Die obere Grenze des Konfidenzintervalls liegt für alle Experimente mindestens bei einem Wert von 3,0. Die vollständigen Ergebnisse der Lehrkräftebefragung sind in Anhang A.4 auf Seite 186 aufgeführt und werden in Kapitel 10 wieder aufgegriffen.

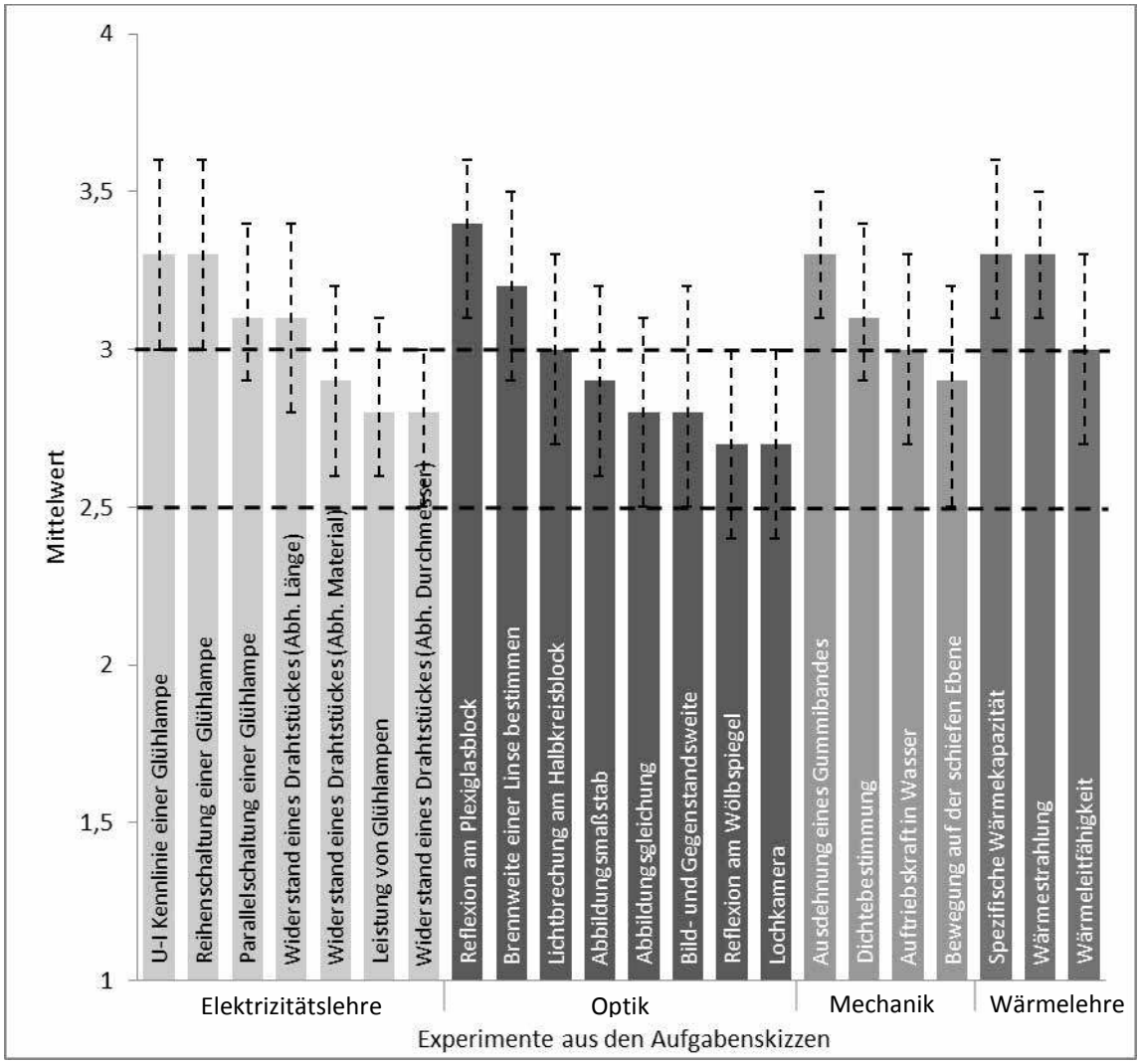


Abbildung 4.8: Ergebnisse der Lehrkräftebefragung zur Frage Durchführbarkeit*? (vertikale Achse: Mittelwert der Einschätzungen auf der vierstufigen Rating-Skala mit 1 ≙ sehr unwahrscheinlich bis 4 ≙ sehr wahrscheinlich; senkrecht gestrichelte Linien an den Balken = 95 % Konfidenzintervalle)*

4.7 Auswahl der Testaufgaben für den MeK-LSA Experimentiertest

Im letzten Schritt des Überführungsprozesses der Testkonzeption in konkrete Testaufgaben (vgl. Abbildung 4.1 auf Seite 37) wurden aus den 22 Aufgabenskizzen zwölf Testaufgaben und eine Trainingsaufgabe für den MeK-LSA Experimentiertest ausgewählt. Für die Auswahl der Aufgaben sind die folgenden Kriterien durch das Testentwicklungsteam zugrunde gelegt worden:

a) Eine mittlere Lehrkräfteeinschätzung bezüglich der Durchführbarkeit der Experimente aus den Aufgabenskizzen, die deutlich über einem neutralen Wert von 2,5 auf der vierstufigen Rating-Skala (Skala von 1 bis 4) liegt.
b) Die Berücksichtigung von drei Inhaltsbereichen, um bei insgesamt zwölf auszuwählenden Testaufgaben eine gute Balancierung zwischen der Abdeckungsbreite relevanter Inhaltsbereiche (Ermöglichung von Untersuchungen zur generischen Verfügbarkeit von Experimentierfähigkeiten) und einer möglichst großen Aufgabenanzahl pro Inhaltsbereich (Erhöhung der Messgenauigkeit innerhalb der Inhaltsbereiche) zu erreichen.
c) Die gleichmäßige Verteilung der Testaufgaben über die Inhaltsbereiche, um mögliche Effekte von Inhaltsbereichen statistisch besser untersuchen zu können.

Kriterium a) ist für alle 22 Aufgabenskizzen erfüllt (vgl. Abschnitt 4.6.2). Die 13 Aufgabenskizzen mit den höchsten mittleren Lehrkräfteeinschätzungen bezüglich der Durchführbarkeit der Experimente (mindestens 3,0 auf der vierstufigen Rating-Skala) stammen aus vier Inhaltsbereichen und decken diese Inhaltsbereiche nahezu gleichmäßig ab (vgl. Abschnitt 4.6.2). Die Auswahl dieser 13 Aufgabenskizzen (zwölf Testaufgaben und eine Trainingsaufgabe) erfüllt allerdings nicht Kriterium b).

Damit Kriterium b) erfüllt ist, wurden die Inhaltsbereiche Elektrizitätslehre, geometrische Optik und Mechanik durch das Testentwicklungsteam als Inhaltsbereiche für die Testaufgaben ausgewählt. Der Inhaltsbereich Wärmelehre wurde im Vergleich zu diesen drei Inhaltsbereichen, auch auf Basis der Erkenntnisse aus der Entwicklungsphase (z. B. Dokumentanalysen), als weniger relevant für das Experimentieren im Physikunterricht der Sekundarstufe I eingeschätzt. Diese Einschätzung wird in Teil II dieser Arbeit (vgl. Kapitel 10) aus dem Blickwinkel der Validität geprüft. Tabelle 4.5 auf Seite 52 zeigt eine Übersicht der zwölf Testaufgaben und der Trainingsaufgabe, die für den MeK-LSA Experimentiertest ausgewählt wurden.

Tabelle 4.5: Übersicht der zwölf Testaufgaben und der Trainingsaufgabe, die für den MeK-LSA Experimentiertest ausgewählt wurden (In Klammern: mittlere Lehrkräfteeinschätzung zur Frage Durchführbarkeit? *auf der vierstufigen Rating-Skala mit 1 ≙ sehr unwahrscheinlich bis 4 ≙ sehr wahrscheinlich; Grau-Abstufungen der Zellen zeigen Inhaltsbereiche (von oben nach unten): Elektrizitätslehre, geometrische Optik, Mechanik)*

	Thema der Aufgabe, (mittlere Lehrkräfteeinschätzung zur Durchführbarkeit)	**Aufgabenstellung bzw. zu überprüfende Vermutung**
Elektrizitätslehre	U-I-Kennlinie einer Glühlampe (3,3)	„Die Stromstärke in der Glühlampe ist proportional zur Spannung an der Glühlampe."
	Leistung von Glühlampen (2,8)	„Welche maximalen Leistungen haben die drei Glühlampen?"
	Parallelschaltung von Glühlampen (3,1)	„Die Summe der Spannungen an den Glühlampen ist gleich der Gesamtspannung an der Spannungsquelle."
	Reihenschaltung von Glühlampen (3,3)	"Die Gesamtstromstärke ist gleich der Summe der Stromstärken in den Glühlampen."
	Trainingsaufgabe: Widerstand eines Drahtstücks (3,1)	„Der Widerstand eines Drahtes ist proportional zu seiner Länge."
Optik	Lichtbrechung am Halbkreisblock (3,0)	„Der Brechungswinkel ist proportional zum Einfallswinkel."
	Reflexion an einem Plexiglasblock (3,4)	„Auch wenn ein Lichtstrahl auf eine nicht-verspiegelte Oberfläche trifft, wird er reflektiert und es gilt das Reflexionsgesetz: Einfallswinkel gleich Ausfallswinkel."
	Totalreflexion (Nachentwicklung)	„Je größer der Brechungsindex des Materials, desto größer ist auch der Grenzwinkel der Totalreflexion."
	Brennweitenbestimmung einer Linse (3,2)	„Je größer die Linsendicke ist, desto kleiner ist die Brennweite."
Mechanik	Dichtebestimmung (3,1)	„Welche Dichten haben die drei Zylinder?“
	Ausdehnung eines Gummibandes[10] (3,3)	"Die Ausdehnung des Gummibands ist proportional zur Masse der angehängten Gewichtsstücke."
	Auftriebskraft in Wasser (3,0)	„Je größer - bei gleichem Volumen - der Durchmesser des Zylinders ist, desto größer ist die Auftriebskraft auf den Zylinder."
	Fahrzeit auf der schiefen Ebene (Nachentwicklung)	"Die Fahrzeit des Autos ist proportional zum Kehrwert des Neigungswinkels der Rampe."

Für den Inhaltsbereich geometrische Optik wurden in Passung zu den Einschätzungen der befragten Lehrkräfte ausschließlich Aufgaben des Typs *Optikexperimente auf dem Tisch* (vgl. Abschnitt 4.3) ausgewählt. Die Aufgabenskizze *Reflexion am Wölbspiegel* wurde jedoch nicht in eine Testaufgabe überführt, da die Lotbestimmung an einer gewölbten Fläche sowohl vom

[10] Abweichend zur Lehrkräftebefragung wird in der Testaufgabe ein Gummiband anstelle eines Gummirings verwendet. Aus diesem Grund wurde die Aufgabenbezeichnung angepasst. An den wesentlichen experimentellen Anforderungen ändert sich nichts.

Testentwicklungsteam als auch von einzelnen Lehrkräften als zu schwierig angesehen wird. Aus diesem Grund wurde eine Aufgabe zur *Totalreflexion* nachentwickelt, für die keine Lehrkräfteeinschätzung vorliegt. Die Aufgabe zur *Totalreflexion* passt zum Aufgabentyp *Optikexperimente auf dem Tisch.* Als Grundlage konnte die Simulation zur Aufgabe *Brechung am Halbkreisblock* mit veränderter Startkonfiguration verwendet werden. Die Aufgabenstellung zur Testaufgabe *Fahrzeit auf der schiefen Ebene* aus dem Inhaltsbereich Mechanik (vgl. Tabelle 4.5 auf Seite 52) weicht deutlich von der Aufgabenstellung in der Aufgabenskizze *Bewegung auf der schiefen Ebene* (*Experimentell nachweisen, dass die Bewegung eines Spielzeugautos auf der schiefen Ebene nicht gleichförmig ist*) ab. Die Einschätzung der Lehrkräfte zur Aufgabenskizze *Bewegung auf der schiefen Ebene* ist folglich nicht auf die Testaufgabe zur *Fahrzeit auf der schiefen Ebene* übertragbar, sodass diese Testaufgabe als Nachentwicklung zu werten ist.

Die Aufgabenauswahl (vgl. Tabelle 4.5 auf Seite 52) des Testentwicklungsteams passt, bis auf die zwei oben diskutierten Ausnahmen (*Totalreflexion* und *Fahrzeit auf der schiefen Ebene*), gut zu den Dokumentanalysen und den Einschätzungen der Lehrkräfte. In Teil II dieser Arbeit (vgl. Kapitel 10 und 11) wird die Aufgabenauswahl aus dem Blickwinkel der Validität geprüft.

4.7.1 Konstruktionsanleitung für die Aufgaben

Die zwölf Testaufgaben und die Trainingsaufgabe (vgl. Tabelle 4.5 auf Seite 52) wurden nach einem einheitlichen Schema konstruiert. Zur Umsetzung der Testkonzeption (z. B. konsekutives Aufgabenformat mit Zwischenlösungen; vgl. Abschnitt 3.1 auf Seite 29) wurden die Aufgabenskizzen in Testaufgaben mit Teilaufgaben überführt. Die Testaufgaben bilden jeweils sechs der acht im Aufgabenentwicklungsmodell beschriebenen Fähigkeiten ab (vgl. Abbildung 3.1 auf Seite 30). Die Fähigkeiten *Versuchsplan entwerfen*, *Versuch aufbauen und testen* und *Messung durchführen und dokumentieren* sind in jeder Testaufgabe als Teilaufgabe enthalten. Die anderen Fähigkeiten (z. B. *Grundidee beschreiben*) sind in mindestens der Hälfte aller Testaufgaben als Teilaufgaben berücksichtigt. Der konkreten Umsetzung der Aufgabenskizzen in Aufgaben und Teilaufgaben dient ein auf PHP-Seiten (PHP: Hypertext Preprocessor) basierendes Framework, das aufgabenübergreifend ein einheitliches Layout für jede der acht Teilaufgaben (z. B. *Versuch aufbauen und testen*) vorgibt. Innerhalb dieses Layouts könnten auch weitere Aufgabenstellungen aus anderen Inhaltsbereichen umgesetzt werden.

Teil II: Studien zur Validität des MeK-LSA Experimentiertests

In Teil I der Dissertation ist neben der Begründung und der Vorstellung des MeK-LSA Experimentiertests insbesondere die systematische Überführung der Testkonzeption in konkrete Testaufgaben beschrieben worden (vgl. Kapitel 4 auf den Seiten 37-53). Diese auf empirische Daten gestützte Überführung bildet den ersten Schritt im Prozess der Testentwicklung bis zur Zuweisung von Testwerten. Darauf aufbauend wird in Teil II der Dissertation über die Methodik und die Ergebnisse umfangreicher Validierungsstudien berichtet, die den gesamten Prozess der Testentwicklung bis zur Zuweisung der Testwerte begleitet haben. Auf diese Weise wird der MeK-LSA Experimentiertest aus dem Blickwinkel der Validität bewertet. In Kapitel 5 wird zunächst das der Arbeit zugrundeliegende Verständnis von Validität vorgestellt. Daran anknüpfend wird ein Bezugsrahmen für die Validitätsbewertung des MeK-LSA Experimentiertest*s* beschrieben, der die notwendigen Validierungsschritte systematisiert und gleichzeitig einen konsistenten Rahmen für den Bewertungsprozess schafft. Der Bezugsrahmen bildet das zentrale Element zur weiteren Strukturierung dieser Arbeit (vgl. Kapitel 6). Ausgehend vom Bezugsrahmen wird theorie- und evidenzbasiert bewertet, bis zu welchem Grad die Testwerte des MeK-LSA Experimentiertest*s* valide als Ausdruck von Experimentierfähigkeiten aufgefasst werden können.

5 Theoretischen Grundlagen zur Validität von Testverfahren

> *"Maybe it's just that jargon tends to be sloppy, but I think that the validation of inferences using tests rather than just tests is a very important idea"* (Rubin, 1988, S. 242)

5.1 Verständnis von Validität

Das Verständnis von Validität hat sich im Laufe der Zeit (ca. seit den 1970er Jahren) grundlegend verändert. Nach Stobart (2009) besteht eine der bedeutsamsten Veränderungen darin,

> *„[to] move away from treating validity as a measurement of a fixed property towards seeing it as an argument about the appropriateness of the inferences drawn from the results and the consequences of these inferences"*(S. 162).

Eine Aussage zur Validität eines Tests (z. B. *Der Physikkompetenztest ist valide*), ist ohne die Berücksichtigung der Testwertinterpretation und des jeweiligen Verwendungszwecks nicht zulässig, da unklar bleibt, worauf sich die Aussage zur Validität genau bezieht. Zunächst muss eine Beschreibung der Testwertinterpretation (z. B. *Die Testwerte sind Indikatoren für die Physikkompetenz von Schülerinnen und Schülern am Ende der Sekundarstufe I*) und des Verwendungszwecks (z. B. *Die Testwerte sind die Entscheidungsgrundlage für die Teilnahme an einer überregionalen Physikolympiade*) vorliegen. Die Frage nach der Validität eines Tests kann daher nicht allgemeingültig, sondern immer nur vor dem Hintergrund der Testwertinterpretation für den jeweiligen Verwendungszweck beantwortet werden (vgl. Nitko & Brookhart, 2007, S. 38).

In den Teststandards der Fachverbände AERA, APA und NCME (2014) wird Validität wie folgt definiert:

> *„Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests“(*AERA et al., 2014, S. 11).

Validität ist demnach als Eigenschaft von Testwertinterpretationen zu verstehen. Die Grundidee besteht darin, theorie- und evidenzbasiert zu beurteilen, bis zu welchem Grad die Testwertinterpretation für den jeweiligen Verwendungszweck valide ist. Die Validierung von Testwertinterpretationen ist dabei als Argumentationsprozess aufzufassen. In diesem Prozess werden Argumente konstruiert und beurteilt, die für bzw. gegen die Testwertinterpretation und die Relevanz dieser Interpretation für den beabsichtigten Verwendungszweck sprechen (AERA et al., 2014, S. 11). Bei der Konstruktion der Argumente und zu ihrer Beurteilung werden verschiedene Validitätsaspekte (*Sources of Validity Evidence*: z. B. Messick, 1995, S. 745; Tabelle 5.2 auf Seite 59; AERA et al., 2014, S. 13-21) berücksichtigt. Diese verschiedenen Validitätsaspekte stellen keine disjunkten Arten von Validität dar, sondern sind Bausteine auf dem Weg zu einer kohärenten, schlüssigen und ganzheitlichen Validitätsargumentation. Nach Leuders (2014) zeigt sich Validität allerdings *„nicht im additiven Vorliegen einzelner Eigenschaften“* (S. 12) von Testwertinterpretationen. Erst durch das Zusammenspiel aller für die Testwertinterpretation relevanten Bausteine (Validitätsaspekte) kann beurteilt werden, bis zu welchem Grad die Testwertinterpretation für den jeweiligen Verwendungszweck valide ist. Welche Validitätsaspekte bzw. in welchem Umfang einzelne Validitätsaspekte in der Validitätsargumentation berücksichtigt werden müssen, ist von dem jeweiligen Verwendungszweck abhängig (vgl. Reynolds, Livingston & Willson, 2010, S. 128). Die Herausforderung bei der Validierung von Testwertinterpretationen besteht im Grunde darin, eine Argumentation zu finden, die die Testwertinterpretation für ihren Verwendungszweck schlüssig belegt und begründet:

> *„A sound validity argument integrates various strands of evidence into a coherent account of the degree to which existing evidence and theory support the intended interpretations of test scores for specific uses“* (AERA et al., 2014, S. 21).

Wesentliche Schritte auf dem Weg zu dem hier skizzierten ganzheitlichen Validitätskonzept sind die Arbeiten von Cronbach (1971, 1988), Messick (u.a. 1989, 1995) und Kane (u.a. 1992, 2006, 2013). Auf der einen Seite wird die Idee dieses ganzheitlichen Validitätskonzepts *„mittlerweile von vielen Vereinigungen der pädagogisch-psychologischen Forschung als Grundlage akzeptiert“* (Gärtner & Pant, 2011a, S. 11) und aktiv vertreten (vgl. AERA et al., 2014). Auf der anderen Seite wird dieses Konzept nach wie vor durch eine kritische Diskussion begleitet (vgl. u.a. Borsboom, Mellenbergh & van Heerden, 2004; Borsboom & Markus, 2013; Lissitz & Samuelsen, 2007). Nach Newton (2012a, S. 1) lässt sich diese kritische Diskussion entlang von zwei Leitfragen strukturieren (vgl. Tabelle 5.1 auf Seite 57). Zum einen stellt sich die Frage, wie das Konzept der Validität zu definieren ist. Zum anderen muss geklärt werden, wie Validität konkret nachgewiesen werden kann. Die erste Leitfrage, wie das Konzept Validität definiert werden kann, wird auf einer theoretischen Ebene diskutiert. Eine ausführliche Darstellung dieser Diskussion ist beispielsweise bei Newton (2012a, 2012b) oder Newton und Shaw (2014) zu finden.

Tabelle 5.1: Leitfragen zur Diskussion des Konzepts der Validität (in Anlehnung an Newton, 2012a, S. 1)

Leitfragen	
Wie kann das Konzept *Validität* definiert werden?	Wie kann *Validität* nachgewiesen werden?
diskutiert wird …	
… wofür es genutzt wird, … wer es benutzt, … worauf es sich bezieht, … wo die Grenzen verlaufen, … wie es mit anderen Konzepten zusammenhängt, …	… welche Validitätsaspekte relevant sind, … in welchem Umfang die Validitätsaspekte berücksichtigt werden sollten, … wie die Validitätsaspekte für eine schlüssige und kohärente Argumentation strukturiert werden müssen, …

Zusammenfassend lässt sich sagen, dass es trotz anhaltender Diskussion als weitestgehend anerkannt gilt, wenn man …

- *„Validität als fortwährenden Prozess der argumentativen und empirischen Verteidigung miteinander verbundener Validitätsaspekte“* (Leuders, 2014, S. 11) auffasst,
- wobei Validität als Eigenschaft von Testwertinterpretationen und nicht als inhärente Eigenschaft eines Tests zu verstehen ist (vgl. u.a. AERA et al., 2014, S. 11; Blömeke, 2013, S. 11; Kane, 2013, S. 3; Messick 1995, S. 741; Cronbach, 1971, S. 447).

Die vorliegende Arbeit bezieht sich auf dieses ganzheitliche Verständnis von Validität.

Ein anderes Bild zeigt sich bei einem Blick auf die Diskussion zur zweiten Leitfrage, wie Validität nachgewiesen werden kann. Diese Diskussion findet primär auf der Ebene der Forschungspraxis statt. Eine schlüssige und kohärente Validitätsargumentation erfordert die Integration verschiedener Validitätsaspekte aus unterschiedlichen Quellen (Kane, 2006, S. 23). Dieser Integrationsschritt wird in der Praxis aber nur selten transparent und vollständig dargestellt. Ein möglicher Grund ist, dass konkrete Hinweise zur Durchführung dieses Schrittes fehlen. So bleibt unklar, in welchem Umfang verschiedene Validitätsaspekte berücksichtigt werden müssen, wie die Validitätsaspekte für eine schlüssige Argumentation strukturiert werden müssen oder wann überhaupt ausreichend Evidenz zur Stützung der beabsichtigten Testwertinterpretation vorliegt (vgl. Kane, 2013, S. 8). Gerade der letzte Aspekt führt dazu, dass eine adäquate Untersuchung der Validität von Testwertinterpretationen einer *„Sisyphos-Aufgabe gleicht“* (Rossa, 2012, S. 74), da auch mit größtmöglichem Aufwand der Validierungssprozess niemals endgültig abgeschlossen werden kann (vgl. auch Anastasi, 1986; Cronbach, 1988). Um die Anwendbarkeit des ganzheitlichen Validitätskonzepts in der Praxis zu erleichtern, wurden in den letzten Jahren verschiedene Ansätze entwickelt und erprobt (u.a. Kane, 1992, 2006, 2013; Crooks, Kane & Cohen, 1996; Mislevy, Steinberg & Almond, 2003; Sprachtestforschung: Weir, 2005; Chapelle, Enright & Jamieson, 2008; Aryadoust, 2013). Diese Ansätze ermöglichen je nach Testwertinterpretation eine begründete Auswahl von

Validitätsaspekten und sind hilfreich bei der Strukturierung der Validitätsargumentation. Ausgangspunkt für die in der vorliegenden Arbeit geführte Validitätsargumentation bilden begründet ausgewählte Validitätsaspekte aus Messicks Validitätskonzept (1995, S. 745), die im Sinne des Validitätskonzepts von Kane (*argument-based-approach* u.a. 1992, 2006, 2013) in eine schlüssige und kohärente Validitätsargumentation überführt werden. Die Grundidee beider Ansätze wird in den Abschnitten 5.2 und 5.3 skizziert.

5.2 Validitätskonzept von Messick

Als Mitbegründer des ganzheitlichen Validitätskonzepts definiert Messick (1989) Validität als

> *„an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions on test scores or other modes of assessment“* (S. 13).

In dieser Definition bündelt Messick eine Vielzahl von Grundannahmen über das Konzept der Validität.

Grundannahmen im Validitätskonzept von Messick

Neben der Grundannahme, dass Testwertinterpretationen und nicht der Test an sich validiert werden[11] (vgl. Abschnitt 5.1), nimmt das Konzept der Konstruktvalidität eine zentrale Stellung ein. Konstruktvalidität wird nicht länger als Ergänzung von Inhalts- und Kriteriumsvalidität aufgefasst, sondern stellt den Bezugsrahmen für sämtliche Betrachtungen zur Validität dar (Hartig, Frey & Jude, 2012, S. 153). Aspekte der Inhalts- und Kriteriumsvalidität werden im Konzept der Konstruktvalidität integriert (Messick, 1989, S. 17). Darüber hinaus erweitert Messick (1980, 1989) das Validitätskonzept um die Berücksichtigung sozialer Konsequenzen und Wertvorstellungen.

Bei der Validierung von Testwertinterpretationen soll nicht ausschließlich nach bestätigender Evidenz gesucht werden, sondern immer auch nach Evidenz, die die Testwertinterpretationen herausfordern bzw. bedrohen könnte (Newton & Shaw, 2014, S. 113). Die Hauptbedrohungen für die Validität von Testwertinterpretationen sind die Unterrepräsentation des Konstrukts und Konstrukt-irrelevante Varianz (vgl. Messick, 1996, S. 5). Unter der Unterrepräsentation des Konstrukts versteht man, dass die Inhalte und Anforderungen in einem Test zu eng gefasst werden und wichtige Bereiche des interessierenden Konstrukts nicht abgebildet werden (Messick, 1995, S. 742). Unter Konstrukt-irrelevanter Varianz versteht man im Test erzeugte Varianz, die nicht im Hinblick auf das interessierende Konstrukt interpretiert werden kann (ebenda). Das wäre beispielsweise der Fall, wenn Performanz in einem Test, der mathematische Kompetenz im Bereich der Arithmetik erfassen soll, nicht primär auf arithmetische Anforderungen der Testaufgaben, sondern überwiegend auf Lese- bzw. Verstehensanforderungen zurückzuführen ist.

[11] Messick (1989) verweist für diese Annahme auf Cronbach (1971).

Die Frage, ob eine bestimmte Testwertinterpretation valide ist, kann nicht einfach mit Ja oder Nein beantwortet werden. Validität kann unterschiedliche Ausprägungen annehmen und ist daher als graduelles Merkmal zu verstehen (Messick, 1989, S. 13). Unterschiedliche Beurteilerinnen und Beurteiler können zum einen zu unterschiedlichen Testwertinterpretationen auf Basis der gleichen Evidenz kommen, und zum anderen kann neue Evidenz etablierte Testwertinterpretationen jederzeit herausfordern bzw. bedrohen (Newton & Shaw, 2014, S. 114). Validierung ist folglich auch ein niemals endender Prozess. Die Herausforderung bei der Validierung besteht nach Messick (1989, S. 13) darin, eine angemessene Argumentation zu finden, die sowohl der aktuellen Verwendung der Testwerte, als auch der aktuellen Forschung zur Weiterentwicklung des Verständnisses über die Testwertinterpretationen gerecht wird. Eine ausführliche Zusammenfassung der zentralen Grundannahmen von Messicks Validitätskonzept findet sich bei Newton und Shaw (2014, S. 112-113).

Validitätsaspekte nach Messick

Messick räumt ein, dass seine Idealvorstellung von Validierung in der Praxis nicht vollständig umsetzbar ist (1989, S. 50). Er hält in der Praxis eine Auswahl von Evidenz bzw. Evidenzquellen für legitim, betont aber gleichzeitig, dass die Auswahl nicht beliebig sein darf (ebenda). Die Auswahl sollte sich vielmehr an der Frage orientieren, welche Evidenz benötigt wird, um die Testwertinterpretationen zu stützen oder herauszufordern (ebenda). Im Zuge der Weiterentwicklung seines Konzepts listet Messick sechs Validitätsaspekte auf (vgl. Tabelle 5.2).

> *„The intent of these distinctions is to provide a means of addressing functional aspects of validity that help to disentangle some of the complexities inherent in appraising the appropriateness, meaningfulness and usefulness of score inferences“* (Messick, 1995, S. 744).

Im Folgenden werden die sechs Validitätsaspekte beschrieben, die *„eine systematische […] Validitätsanalyse [ermöglichen], die dabei hilft, die für jeden Schritt spezifischen Validitätsbedrohungen zu erkennen und zu bewältigen“* (Leuders, 2014, S. 12).

*Tabelle 5.2: Validitätsaspekte nach Messick (1995, S. 745); *Übersetzung nach Leuders (2014, S.11 -12)*

	Beschreibung*
Inhaltlicher Aspekt	Curriculare und theoretische Absicherung des modellierten Bereichs
Kognitiver Aspekt	Passung der kognitiven Prozesse bei der Kompetenzerfassung zum postulierten theoretischen Kompetenzmodell
Struktureller Aspekt	Passung von theoretischem Kompetenzmodell und gewähltem psychometrischen Messmodell
Generalisierbarkeit	Angemessenheit einer über die Aufgaben- und Personengruppe hinausgehenden Interpretation
Externer Aspekt	Angemessenheit mit Blick auf konvergente, diskriminante und prädiktive Zusammenhänge mit anderen Konstrukten
Konsequentieller Aspekt	Angemessenheit der Nutzung im pädagogischen oder bildungspolitischen Kontext

Inhaltlicher Aspekt

Der inhaltliche Aspekt befasst sich mit der Relevanz und Repräsentativität der Testinhalte (Messick, 1995, S. 745). Der Begriff Testinhalt bezieht sich neben dem Aufgabenthema auch auf die Aufgabenformulierungen, die Aufgabenanforderungen, das Aufgabenformat, die Testadministration und das Bewertungsschema der Aufgaben (AERA et al, 2014, S. 14). Um die Relevanz der Testinhalte beurteilen zu können, muss in einem ersten Schritt der inhaltliche Rahmen des modellierten Bereichs festgelegt werden (Leuders, 2014, S. 12), wobei genau beschrieben werden sollte, was zu diesem Bereich gehört und was nicht (Newton & Shaw, 2014, S. 114). Bei der Beschreibung der Struktur und der Grenzen des modellierten Bereichs sollte – soweit möglich – auf bereits bestehende Theorien zurückgegriffen werden (Messick, 1995, S. 745). Neben der Relevanz der Testinhalte muss auch deren Repräsentativität für den modellierten Bereich überprüft werden. Die Repräsentativität einer Testaufgabe zeigt sich insbesondere in der Passung der gestellten Aufgabenanforderungen und -inhalte zu den Anforderungen und Inhalten, die eine vergleichbare Realsituation außerhalb des Testkontexts kennzeichnen (ebenda). Evidenz für die Relevanz und Repräsentativität von Testinhalten kann durch eine systematische Literaturrecherche zu theoretischen Beschreibungen des modellierten Bereichs und systematische Analysen von Lehrplänen und Schulbüchern sowie qualitative Expertenbefragungen gewonnen werden (u.a. Gärtner & Pant, 2011b, S. 87; Messick, 1995, S. 745; Wilhelm & Kunina, 2009, S. 319). Leutner, Hartig & Jude (2008, S. 183) betonen die zentrale Stellung des inhaltlichen Aspekts für die Kompetenzdiagnostik, da Kompetenzkonstrukte immer auch kontext- und situationsspezifisch definiert werden.

Kognitiver Aspekt

Der kognitive Aspekt beschäftigt sich mit der Passung der kognitiven Prozesse bei der Kompetenzerfassung zum postulierten theoretischen Kompetenzmodell (Leuders, 2014, S. 11). In Erweiterung des inhaltlichen Aspekts wird untersucht, inwieweit die beabsichtigten kognitiven Prozesse bei den Testteilnehmenden während der Bearbeitung der Testaufgaben ablaufen (vgl. Messick, 1995, S. 745; Newton & Shaw, 2014, S. 115; AERA et al., 2014, S. 15). Soll der Test beispielsweise arithmetische Fähigkeiten von Schülerinnen und Schülern am Ende der Sekundarstufe I erfassen, muss bestimmt werden, ob die Schülerinnen und Schüler während der Lösung der Testaufgaben tatsächlich arithmetische Überlegungen anstellen. Evidenz für den kognitiven Aspekt kann einerseits explorativ durch die Erfassung der kognitiven Prozesse in testähnlichen Situationen mit Think-Aloud- (z. B. Ericsson & Simon, 1993) und Stimulated-Recall-Methoden (z. B. Gass & Mackey, 2000) oder durch die Aufzeichnung von Blickbewegungen mit Eye-Tracking-Verfahren (z. B. van Gog & Jarodzka, 2013) gewonnen werden (vgl. Messick, 1995, S. 745; Leuders, 2014, S. 15). Evidenz kann andererseits auch durch *„die systematische Untersuchung der kognitiven Prozesse in Abhängigkeit von Eigenschaften der Items“* (Hartig et al., 2012, S. 161) generiert werden. Ein Ziel dieser systematischen Untersuchung ist es, konstrukt-relevante und schwierigkeitserzeugende Merkmale von Testaufgaben zu identifizieren und auf deren Basis die empirisch ermittelten Itemschwierigkeiten vorherzusagen (ebenda, S. 162). Voraussetzung für diese Vorgehensweise ist, dass theoretisch fundierte Annahmen über die zur Bearbeitung

der Testaufgaben erforderlichen kognitiven Prozesse in Form von kognitiven Modellen der Aufgabenbearbeitung vorliegen (vgl. Rupp & Mislevy, 2007). Da häufig (noch) keine theoretisch fundierten kognitiven Modelle vorhanden sind, ist nach Borsboom und Mellenbergh (2007, S. 100) auch eine Kombination eines explorativen und eines modellbasierten Ansatzes vorstellbar (Beispiel: Gierl, Leighton & Hunka, 2007).

Struktureller Aspekt

Der strukturelle Aspekt befasst sich mit der Passung zwischen dem theoretischen Kompetenzmodell und dem gewählten psychometrischen Messmodell[12] (vgl. Leuders, 2014, S. 11). Wird beispielsweise angenommen, dass sich experimentelle Kompetenz nach drei Bereichen (Planung, Durchführung und Auswertung) strukturieren lässt, so muss sich diese Struktur auch in den Testdaten bzw. im psychometrischen Messmodell wiederfinden lassen. Evidenz kann durch empirische und theoretische Analysen gewonnen werden. Empirisch können verschiedene Modelle (eindimensionale vs. mehrdimensionale) im Hinblick auf den Passungsgrad zu den Testdaten verglichen werden (Leuders, 2014, S. 18). Theoretisch kann die Qualität der Skalen oder der Zusammenhang zwischen den (latenten) Variablen mit dem Wissen über die kognitiven Prozesse während der Aufgabenbearbeitung in Beziehung gesetzt und begründet werden (ebenda).

Generalisierbarkeit

Der Aspekt der Generalisierbarkeit befasst sich mit der Frage, bis zu welchem Grad Testwertinterpretationen über Zielgruppen, Messzeitpunkte, Aufgaben und Verfahren zur Aufgabenbewertung (z. B. verschiedene Rater) hinweg verallgemeinerbar sind (Messick, 1995, S. 745-746; Leuders, 2014, S. 18). Obwohl die Struktur und die Grenzen des modellierten Bereichs bereits unter dem Blickwinkel des inhaltlichen Aspekts beschrieben werden, können Annahmen zur Generalisierbarkeit nicht ohne weitere Prüfung beibehalten werden (Newton & Shaw, 2014, S. 115). Aufgrund begrenzter Testzeit kann beispielsweise immer nur eine Auswahl von Aufgaben aus dem modellierten Bereich eingesetzt werden. Um die Generalisierbarkeit der Testwertinterpretationen zu erhöhen, schlägt Messick (1996, S. 11) insbesondere für Interpretationen auf Gruppenebene die Verwendung eines Multi-Matrix-Designs vor. So erhöht sich bei gleichbleibender Testzeit die Anzahl der eingesetzten Aufgaben in der Gruppe. In einem qualitativ-inhaltlichen Schritt kann die Generalisierbarkeit von Testwertinterpretationen z. B. durch die Standardisierung von Aufgabeneigenschaften und Administrationsbedingungen verbessert werden (im Kontext von Performance-Assessments: Kane, Crooks & Cohen, 1999, S. 10). Allerdings bleibt bei dieser Vorgehensweise die Generalisierbarkeit auf die standardisierten Eigenschaften und Bedingungen begrenzt. Empirische Aussagen zur Generalisierbarkeit der Testwertinterpretationen sind durch die Bestimmung des Anteils konstrukt-irrelevanter

[12] Die Begriffe orientieren sich in Anlehnung an Leuders (2014) an der probabilistischen Testtheorie und nicht an den im Original von Messick (1989, 1995) verwendeten Begriffen.

Varianz, z. B. durch den Einfluss verschiedener Rater oder Messzeitpunkte, im Rahmen der Generalisierbarkeitstheorie (z. B. Brennan, 2011) möglich.

Externer Aspekt

Der externe Aspekt befasst sich mit der Frage, welcher theoretische und empirische Zusammenhang zwischen dem erfassten Konstrukt und vorhandenen Theorien sowie anderen Konstrukten besteht (Leuders, 2014, S. 19). Vermutet man, dass zwei Testverfahren dasselbe Konstrukt erfassen, so spricht man von der Untersuchung konvergenter Validität. Vermutet man, dass zwei Testverfahren unterschiedliche Konstrukte erfassen, so spricht man von der Untersuchung diskriminanter Validität. Zur Untersuchung der Zusammenhänge werden im Rahmen der klassischen Testtheorie in der Regel Korrelationen zwischen den Testwerten berechnet. Hinreichend hohe Korrelationen zwischen den Testwerten sind ein Indikator für konvergente Validität. Keine oder niedrige Korrelationen sind dagegen Indikatoren für diskriminante Validität. Welche Korrelationswerte als hoch oder niedrig bezeichnet werden können, sollte – falls möglich – durch theoretisch begründete Mindest- bzw. Höchstwerte a priori festgelegt werden (Hartig et al., 2012, S. 158-159). Gleichzeitige Evidenz für konvergente und diskriminante Zusammenhänge kann beispielsweise durch korrelative Multitrait-Multimethod-Vergleiche (vgl. Campbell & Fiske, 1959) gewonnen werden.

Konsequentieller Aspekt

Der konsequentielle Aspekt befasst sich mit der Frage, welche beabsichtigten und unbeabsichtigten Konsequenzen sich aus den Testwertinterpretationen und deren Verwendungszweck ergeben (Messick, 1995, S. 746). Sollen zugewiesene Testwerte zur Selektion im Schulsystem genutzt werden (*High-Stakes Einsatz*), so muss sichergestellt werden, dass sich keine negativen sozialen Konsequenzen, aufgrund fehlerhafter Testwertinterpretationen oder einer unfairen Verwendung der Testwertinterpretationen, ergeben. Leuders (2014, S. 19) sieht in Anlehnung an Pellegrino et al. (2001, S. 2) eine zu breit angenommene Zielsetzung eines Testverfahrens als Bedrohung für die konsequentielle Validität von Testwertinterpretationen.

Grenzen für die Anwendung in der Praxis

Alle sechs Validitätsaspekte sind wichtige Bezugspunkte für die Untersuchung von Fragen zur Validität. Aus praktischer Perspektive ist die Berücksichtigung aller sechs Validitätsaspekte ein Anspruch dem – wenn überhaupt – nur breit angelegte Forschungsprogramme gerecht werden können (Wolming & Wikström, 2010). Kleinere Forschungsprojekte müssen sich aus pragmatischen Gründen auf ausgewählte Validitätsaspekte beschränken. Eine Beschränkung auf für die Testwertinterpretationen relevante Validitätsaspekte ist auch für Messick (1989, S. 50) durchaus legitim. Allerdings finden sich in seinen Arbeiten keine konkreten Hinweise, nach welchen Kriterien die relevanten Validitätsaspekte auszuwählen sind. Eine Möglichkeit, den Prozess der Validierung im Hinblick auf die Testwertinterpretationen zu systematisieren und gleichzeitig nur ausgewählte Validitätsaspekte zu berücksichtigen, bietet das Validitätskonzept von Kane (vgl. Abschnitt 5.3).

5.3 Argument-based-approach von Kane

Das Validitätskonzept *einer theoretisch und empirisch fundierten argumentativen Stützung einer Testwertinterpretation*[13] (*argument-based approach*) wurde verstärkt wieder von Kane (u.a. 1992, 2006, 2013) in die Validitätsdebatte eingebracht.

Grundidee des argument-based-approach

Der *argument-based-approach* umfasst zwei Argumente (vgl. Kane, 2013, S. 9-10):

1. ein Interpretations-Nutzungs-Argument (*interpretation use argument*; INA)
2. ein Validitätsargument (*validity argument*)

Das INA bildet den Bezugsrahmen für die Validierung beabsichtigter Testwertinterpretationen und Verwendungszwecke[14]. Im INA wird ein Netz aus Aussagen und Annahmen geknüpft, das die Testwertinterpretationen stützt. Das Validitätsargument evaluiert das INA entlang der folgenden Fragen (vgl. Kane, 2012, S. 13):

1. Sind die Aussagen und Annahmen im INA sowie die Maßnahmen zu deren Stützung detailliert genug beschrieben, um die Argumentation für die Testwertinterpretationen nachvollziehen zu können? (*Clarity of the argument*)
2. Ist das INA kohärent, d. h. ermöglicht es eine schlüssige Argumentation? (*Coherence of the argument*)
3. Sind die Aussagen und Annahme im INA a priori plausibel oder durch Evidenz gestützt? (*Plausibility of assumptions*)

Die Testwertinterpretationen sind bis zu dem Grad valide, zu dem das INA die drei beschriebenen Kriterien erfüllt (Kane, 2013, S. 14). Ein endgültiger Beweis der Validität der Testwertinterpretationen ist nicht möglich (Kane, 1992, S. 527).

Merkmale des argument-based-approach

Bei jeder Anwendung des *argument-based-approach* muss das INA im Hinblick auf die beabsichtigten Testwertinterpretationen ausgearbeitet werden. Die Validierung von Testwertinterpretationen kann daher auch mit diesem Ansatz nicht als *„standardisiertes Routineverfahren“* (Schmiemann & Lücken, 2014, S. 108) erfolgen. Der Ansatz kann allerdings dabei helfen *„[to] organize our thinking about important questions and identify priorities“* (Shepard, 1993, S. 432). Ein Vorteil, den das INA als Bezugsrahmen für Testwertinterpretationen gegenüber anderen Bezugsrahmen bietet, ist die Flexibilität seiner Ausgestaltungsmöglichkeiten. Normalerweise bilden entweder operational definierte Merkmale oder ein theoretisches Konstrukt den Bezugsrahmen für Testwertinterpretationen (Hartig et al., 2012, S. 147; vgl. auch Abbildung 5.1 auf Seite 64). Operational definierte Merkmale werden im Kern durch die Testinhalte definiert (ebenda).

[13] Übersetzung in Anlehnung an Blömeke (2013, S. 9)

[14] Im Folgenden wird die Kurzform Testwertinterpretationen verwendet. Gemeint sind weiterhin auch Verwendungszwecke.

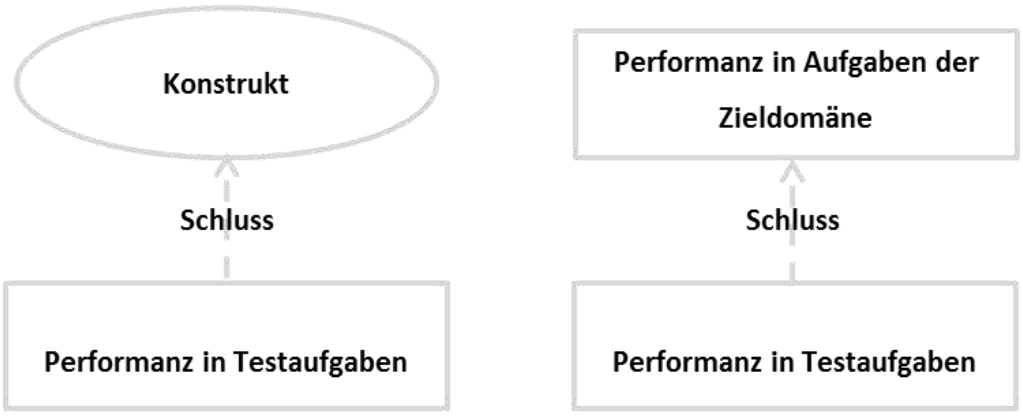


Abbildung 5.1: theoretisches Konstrukt (links) und operationale Merkmalsdefinition (rechts) als Bezugsrahmen für Testwertinterpretationen in Anlehnung an Chapelle et al. (2008, S. 3-4)

Ein Konstrukt ist nach Wilhelm und Kunina *„ein nicht direkt beobachtbarer Sachverhalt innerhalb einer wissenschaftlichen Theorie“* (2009, S. 315). Aufgrund zum Teil nur wenig formal entwickelter Theorien ist der Übergang zwischen theoretischem Konstrukt und operationaler Merkmalsdefinition als Bezugsrahmen für Testwertinterpretationen häufig fließend (Hartig et al., 2012, S. 148). Für den Bereich *Sprachkompetenz* nennen Chapelle et al. (2008, S. 5) beispielsweise drei Aspekte, die in einem Bezugsrahmen für Testwertinterpretationen berücksichtigt werden sollten. Übertragen auf den Bereich *experimentelle Kompetenz* sollte der Bezugsrahmen für Testwertinterpretationen …

1. ein theoretisches Konstrukt *experimentelle Kompetenz* sein, das Experimentierfähigkeiten auf einem breiten Anforderungs- und Inhaltsspektrum abdeckt.
2. beabsichtigte kognitive Prozesse und Strategien enthalten.
3. relevante und repräsentative Inhalte und Anforderungen berücksichtigen.

Während die ersten beiden Aspekte ihren Ursprung im *Konstrukt-Ansatz* haben, lässt sich der dritte Aspekt primär dem *Operationalen-Ansatz* zuordnen (vgl. Abbildung 5.1). Das INA ermöglicht eine Kombination beider Ansätze. Ist das INA einmal ausgearbeitet, liefert es einen verbindlichen Bezugsrahmen, der vorgibt, welche Kriterien zur Validierung der Testwertinterpretationen evaluiert werden müssen (Kane, 2013, S. 9). Die Anwendung des *argument-based-approach* erfordert menschliche Urteile und folgt keinem automatischen Ablauf. Urteilsfehler sind daher nicht auszuschließen.

Kane (2013, S. 18-19) beschreibt fünf potentielle Fehlerquellen:

- *begging-the-question fallacy*: Gemeint ist damit, dass Aussagen und Annahmen im INA als plausibel angenommen werden, die weder a priori plausibel sind noch ausreichend durch Evidenz gestützt werden.
- *straw man fallacy*: Gemeint ist damit, dass die INAs ambitionierter und umfassender sind als für die beabsichtigten Testwertinterpretationen erforderlich ist.
- *reification fallacy*: Gemeint ist damit, dass aus einer beobachteten Regelmäßigkeit (z. B. konsistente Performanz bei bestimmten Aufgaben) auf die Existenz eines Konstrukts (z. B. analytische Fähigkeiten) als Ursache für diese Regelmäßigkeit geschlossen wird, ohne dass ausreichend Evidenz für diesen Schluss oder zusätzliche Annahmen zum Konstrukt vorliegen (z. B. Generalisierbarkeit über Kontexte und Messzeitpunkte).
- *gilding the lily*: Gemeint ist damit, dass zusätzliche Evidenz für bereits umfassend gestützte Aussagen und Annahmen gesammelt wird, um zu überdecken das andere Teile des INAs nicht gestützt werden können.
- *fallacy of statistical surrogation*: Gemeint ist damit, dass zur Stützung von Annahmen statistische Konzepte (z. B. Korrelation) verwendet werden, für deren Stützung ein anspruchsvolleres Konzept (z. B. Kausalität) notwendig wäre.

Kane (z. B. 2006, S. 25) unterscheidet im Hinblick auf den Nachweis von Validität zwei Phasen bei der Entwicklung eines Testverfahrens: Die Entwicklungsphase (*development stage*) und die Bewertungsphase (*appraisal stage*). In der Entwicklungsphase stehen die Testentwicklerinnen und Testentwickler vor der Herausforderung, ein Testverfahren und ein INA zu entwickeln, die die beabsichtigten Testwertinterpretationen stützen (ebenda). In dieser Phase wird zur Validierung der Testwertinterpretationen ein Plädoyer für die Testwertinterpretationen gehalten, wobei Bestätigungsfehler (*confirmation bias*) kaum zu vermeiden sind (Kane, 2012, S. 4). In der Bewertungsphase wird eine kritischere Sicht auf Validität eingenommen. In dieser Phase wird zur Validierung der Testwertinterpretationen eine möglichst objektive Evaluation des INA vorgenommen und Evidenz für und gegen die Testwertinterpretationen gleichermaßen berücksichtigt (ebenda).

Zusammenfassend lässt sich sagen, dass der *argument-based-approach* durch das INA zusätzlich zur Strukturierung des Validierungsprozesses eine Fokussierung auf die für die beabsichtigte Testwertinterpretationen zu erbringenden Evidenznachweise ermöglicht (Newton & Shaw, 2014, S. 153). Durch die Evaluation des INA (*Validitätsargumen*t) kann trotz der Fokussierung auf ausgewählte Validitätsaspekte eine schlüssige Argumentation für die Stützung der Testwertinterpretationen erfolgen, mit klarem Ausgangs- und Endpunkt.

6 Interpretations-Nutzungs-Argument für den MeK-LSA Experimentiertest

In diesem Kapitel wird der Bezugsrahmen für die Validitätsbewertung des MeK-LSA Experimentiertests beschrieben, der die notwendigen Validierungsschritte benennt, systematisiert und gleichzeitig einen konsistenten Rahmen für den Bewertungsprozess schafft. Die Formulierung des Bezugsrahmens erfolgt in Anlehnung an den *argument-based-approach* von Kane (vgl. Abschnitt 5.3 auf Seite 63) als Interpretations-Nutzungs-Argument (INA).

Die Zieldomäne *Experimentieren im Physikunterricht der Sekundarstufe I* bildet den Ausgangspunkt für das INA. Um die Zieldomäne im MeK-LSA Experimentiertest abzubilden, sind drei aufeinander aufbauende Schritte notwendig (vgl. Abbildung 6.1).

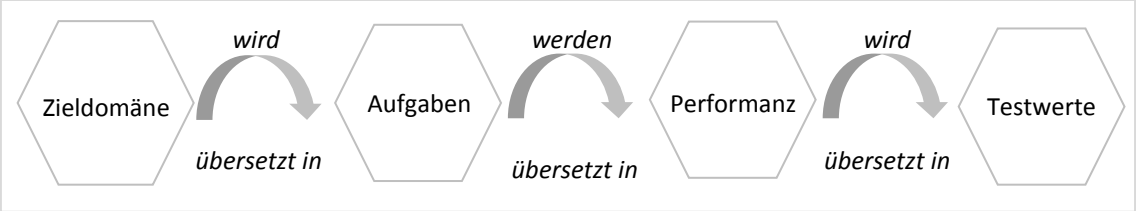


Abbildung 6.1: Erforderliche Schritte von der Zieldomäne Experimentieren im Physikunterricht der Sekundarstufe I *bis zur Zuweisung von Testwerten*

Jeder der drei Schritte (genauer: die Ergebnisse der Schritte) muss Qualitätsanforderungen erfüllen, um mit dem MeK-LSA Experimentiertest experimentelle Kompetenz in der Zieldomäne valide messen zu können. Die von den drei Schritten zu erfüllenden Qualitätsanforderungen werden im INA durch die folgenden übergeordneten Aussagen beschrieben:

1. Die Aufgaben umfassen relevante und repräsentative Inhalte und Anforderungen aus der Zieldomäne (Teil I des INA).
2. Die beobachtete Performanz passt zur beabsichtigten Performanz (Teil II des INA).
3. Die beobachtete Performanz wird in geeigneter Art und Weise in Testwerte überführt (Teil III des INA).

Innerhalb des INA erfolgt eine Konkretisierung der übergeordneten Aussagen in Form von Annahmen (vgl. Abschnitte 6.1 bis 6.3). Diese Annahmen werden theorie- und evidenzbasiert geprüft (vgl. Prüfung der Annahmen aus den INA Teilen I bis III, Kapitel 9 bis 16). Die sich aus dieser Prüfung ergebenden Nachweise für das Beibehalten der Annahmen stützen die übergeordneten Aussagen. Nach der Prüfung aller Annahmen kann daher für den MeK-LSA Experimentiertest bewertet werden, bis zu welchem Grad die folgende Testwertinterpretation beibehalten werden kann:

Die Testwerte können valide als Ausdruck von Experimentierfähigkeiten aufgefasst werden.

6.1 Die Aufgaben umfassen relevante und repräsentative Inhalte und Anforderungen aus der Zieldomäne (Teil I des INA)

In Anlehnung an den inhaltlichen Aspekt von Messick (vgl. Abschnitt 5.2 auf Seite 58) muss der erste Schritt (von der Zieldomäne zu den Aufgaben) die folgende Qualitätsanforderung (übergeordnete Aussage) erfüllen: *Die Aufgaben umfassen relevante und repräsentative Inhalte und Anforderungen aus der Zieldomäne*. Diese übergeordnete Aussage lässt sich durch eine zu prüfende Annahme zu den Aufgabeninhalten (I.I) und eine zu prüfende Annahme zu den Anforderungen der Aufgaben (I.II) konkretisieren (vgl. Abbildung 6.2).

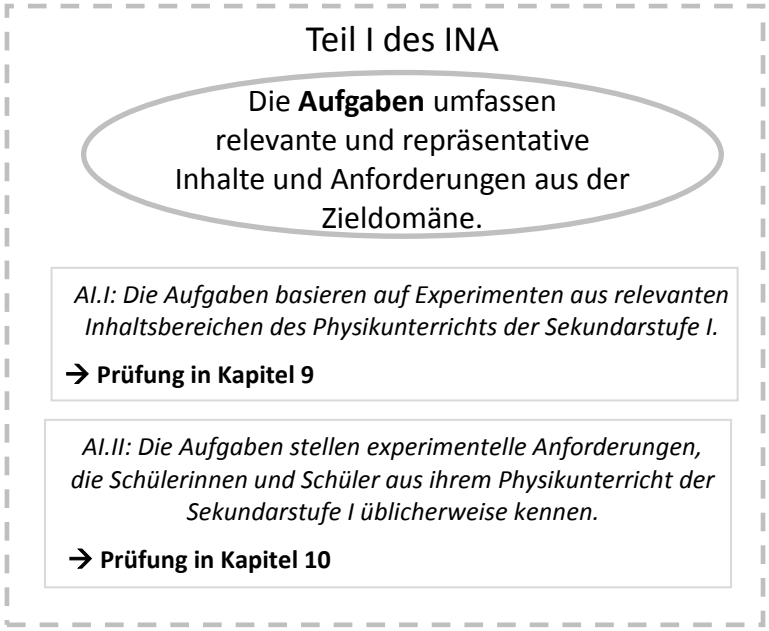


Abbildung 6.2: Teil I des INA mit den Annahmen I.I und I.II zur ersten übergeordneten Aussage (ovale Komponente; Prüfung der Annahmen in Kapitel 9 und 10)

Die Zieldomäne *Experimentieren im Physikunterricht der Sekundarstufe I* kann u.a. aufgrund begrenzter Testzeit nicht vollständig in den Aufgaben des MeK-LSA Experimentiertest*s* abgebildet werden. Daher kommt der Definition, Eingrenzung und Auswahl von Aufgaben eine zentrale Bedeutung bei der Testentwicklung zu (Hartig & Jude, 2007, S. 26). Relevante Inhalte und Anforderungen werden durch die Beschreibung der Struktur und der Grenzen der Zieldomäne bestimmt (Messick, 1995, S. 745). Die Repräsentativität zeigt sich insbesondere in der Passung der Aufgabeninhalte und -anforderungen zu den tatsächlich gestellten Inhalten und Anforderungen in der Praxis (ebenda).

Begründung von Annahme I.I

Bereits während der Entwicklung des MeK-LSA Experimentiertest*s* wurde die Testkonzeption systematisch in konkrete Testaufgaben überführt (vgl. Kapitel 4 auf den Seiten 37-53). Da sich die Auswahl relevanter Inhaltsbereiche für die Testaufgaben auf umfangreiche Lehrplan- und Schulbuchanalysen stützt (vgl. Abschnitt 4.1 auf Seite 37), ist folgende Annahme (I.I) plausibel: *Die Aufgaben basieren auf Experimenten aus relevanten Inhaltsbereichen des Physikunterrichts der Sekundarstufe I* (vgl. Abbildung 6.2). Bedrohungen für das Beibehalten von Annahme I.I

bestehen darin, dass bei der abschließenden Auswahl der Testaufgaben auch normative Entscheidungen des Testentwicklungsteams eine Rolle gespielt haben (vgl. Abschnitt 4.7 auf Seite 51). Um zu prüfen, ob Annahme I.I dennoch beibehalten werden kann, werden in Kapitel 9 Ergebnisse einer Lehrkräftebefragung zur Bekanntheit der Experimente aus den Aufgabenskizzen (vgl. Abschnitt 4.6 auf Seite 48 und Studie A in Tabelle 7.1 auf Seite 79), Ergebnisse einer Schülerbefragung zur Bekanntheit der Experimente aus den Testaufgaben (vgl. Studie B in Tabelle 7.1 auf Seite 79) sowie Ergebnisse einer externen Lehrkräftebefragung zum Einsatz von Experimenten im eigenen Physikunterricht (Karaböcek & Erb, 2015) zusammengeführt.

Begründung von Annahme I.II

Einen curricularen Rahmen für relevante Anforderungen der Zieldomäne bilden die Bildungsstandards für das Unterrichtsfach Physik. Kompetenzerwartungen zum Experimentieren werden dort innerhalb des Kompetenzbereichs *Erkenntnisgewinnung* beschrieben (KMK, 2005c). Am Ende der Sekundarstufe I sollen die Schülerinnen und Schüler im Physikunterricht in der Regel folgende Kompetenzerwartungen erfüllen können (KMK, 2005c, S. 11):

- Aufstellen von Hypothesen an einfachen Beispielen [E6].
- Einfache Experimente nach Anleitung durchführen und auswerten können [E7].
- Einfache Experimente planen, sie durchführen und die Ergebnisse dokumentieren können [E8].
- Gewonnene Daten auswerten können, ggf. auch durch einfache Mathematisierungen [E9].
- Die Gültigkeit empirischer Ergebnisse und deren Verallgemeinerung beurteilen [E10].

Eine Konkretisierung relevanter Anforderungen erfolgt über Kompetenzmodelle, mit denen die in den Standards formulierten Kompetenzerwartungen ausdifferenziert werden (vgl. auch Klieme et al., 2003, S 71). Grundlage für die Entwicklung des MeK-LSA Experimentiertests war ein Aufgabenentwicklungsmodell (vgl. Abbildung 3.1 auf Seite 30), das auf dem *eXkomp-Modell* (vgl. Abbildung 2.1 auf Seite 19) und dem S*pinnennetzmodell* (vgl. Abbildung 2.2 auf Seite 20) basiert. Beide Modelle (*eXkomp-Modell*, *Spinnennetzmodell*) sind anschlussfähig an nationale und internationale Standardbeschreibungen zu Kompetenzerwartungen beim Experimentieren und weisen eine hohe Passung zu anderen im Teilprozessansatz (Dreischritt: Ideen- bzw. Hypothesenfindung, Durchführung, Auswertung) verorteten Kompetenzmodellen auf (vgl. Abschnitt 2.1.1 auf Seite 18). Da auch Lehrkräfte die in den Modelle beschriebenen Fähigkeiten als relevant für die Unterrichtspraxis einschätzen (vgl. *eXkomp-Modell*) bzw. die eigene Unterrichtspraxis mit in die Entwicklung eingebracht haben (vgl. *Spinnennetzmodell*), ist es plausibel anzunehmen, dass die Modelle Fähigkeiten abbilden, für die es im naturwissenschaftlichen Unterricht Lerngelegenheiten gibt (vgl. Abschnitt 2.1.1). Die im Aufgabenentwicklungsmodell beschriebenen Fähigkeiten können aufgrund der Bezugsmodelle (*eXkomp-Modell* und S*pinnennetzmodell*) als relevant und repräsentativ für die Zieldomäne

angenommen werden. Das Aufgabenentwicklungsmodell bildet damit eine plausible Grundlage für Testaufgaben, die relevante und repräsentative Anforderungen der Zieldomäne abdecken sollen. Aus den im Aufgabenentwicklungsmodell beschriebenen Fähigkeiten kann jedoch nicht darauf geschlossen werden, auf welchem Anforderungsniveau die Fähigkeiten getestet werden sollen.

Die für die Aufgaben des MeK-LSA Experimentiertest*s* ausgewählten Experimente (vgl. Abschnitt 4.7 auf Seite 51) wurden, mit Ausnahme von zwei nachentwickelten Experimenten, von Lehrkräften auf Basis von Aufgabenskizzen als für Schülerinnen und Schüler wahrscheinlich durchführbar eingeschätzt (vgl. Abschnitt 4.6.2 auf Seite 49). Das ist ein Argument für die folgende Annahme (I.II): *Die Aufgaben stellen experimentelle Anforderungen, die Schülerinnen und Schüler aus ihrem Physikunterricht der Sekundarstufe I üblicherweise kennen* (vgl. Abbildung 6.2 auf Seite 68). Bedrohungen für das Beibehalten von Annahme I.II bestehen zum einen darin, dass bisher nicht berücksichtigt wurde, inwieweit die Experimente aus den Aufgabenskizzen auch die Anforderungen in den Bereichen Planung und Auswertung erfüllen. Zum anderen ist bisher nicht geprüft worden, inwieweit die experimentellen Anforderungen des MeK-LSA Experimentiertest*s* aus Sicht von Schülerinnen und Schülern Bestandteil des eigenen Physikunterrichts sind. Zur weiteren Prüfung von Annahme I.II werden in Kapitel 10 Ergebnisse der Lehrkräftebefragung zur Erfüllbarkeit der experimentellen Anforderungen (vgl. Abschnitt 4.6 auf Seite 48 und Studie A in Tabelle 7.1 auf Seite 79) und die Ergebnisse einer Schülerbefragung zu experimentellen Anforderungen im eigenen Physikunterricht (vgl. Studie C in Tabelle 7.1 auf Seite 79) zusammengeführt.

6.2 Die beobachtete Performanz passt zur beabsichtigten Performanz (Teil II des INA)

Unter der Voraussetzung, dass die Aufgaben relevante und repräsentative Inhalte und Anforderungen aus der Zieldomäne abdecken (vgl. Abschnitt 6.1.), muss der nächste Schritt (von den Aufgaben zur Performanz) die folgende Qualitätsanforderung (übergeordnete Aussage) erfüllen: *Die während der Bearbeitung der Testaufgaben beobachteten Schülerhandlungen (beobachtete Performanz) passen zu experimentellem Handeln in repräsentativen Situationen (experimentelle Performanz)*. Diese übergeordnete Aussage lässt sich durch vier zu prüfende Annahmen (II.I bis II.IV) konkretisieren (vgl. Abbildung 6.3 auf Seite 71).

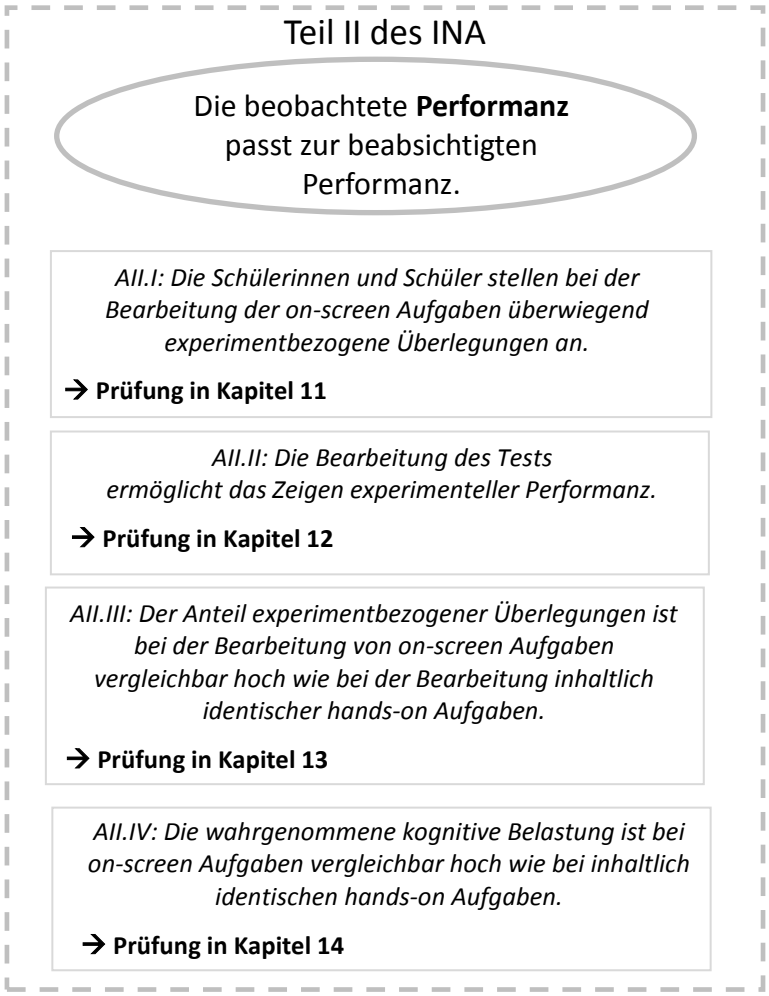


Abbildung 6.3: Teil II des INA mit den Annahmen II.I bis II.IV zur zweiten übergeordneten Aussage (ovale Komponente; Prüfung der Annahmen in Kapitel 11 bis 14)

Begründung von Annahme II.I

In Anlehnung an den kognitiven Aspekt von Messick (vgl. Abschnitt 5.2 auf Seite 58) besteht eine Grundvoraussetzung darin, dass die beobachteten Schülerhandlungen überwiegend auf experimentbezogenen Überlegungen basieren. Bislang liegen für Experimentiertests keine abgesicherten Theorien zu während der Aufgabenbearbeitung ablaufenden Überlegungen (kognitiven Prozessen) vor. Aus diesem Grund kann a priori nicht verlässlich bestimmt werden, welche Aufgabenmerkmale experimentelle Performanz auf kognitiver Ebene beeinflussen. Auch bei Berücksichtigung relevanter und repräsentativer Inhalte und Anforderungen aus der Zieldomäne (vgl. Abschnitt 6.1) muss daher folgende Annahme (II.I) geprüft werden*: Die Schülerinnen und Schüler stellen bei der Bearbeitung der on-screen Aufgaben überwiegend experimentbezogene Überlegungen an* (vgl. Abbildung 6.3). Die Prüfung von Annahme II.I erfolgt in Kapitel 11 auf Basis von Think-Aloud-Daten (vgl. Studie C in Tabelle 7.1 auf Seite 79).

Begründung von Annahme II.II

Der MeK-LSA Experimentiertest basiert auf vorstrukturierten, unabhängig voneinander zu bearbeitenden on-screen (Teil-)Aufgaben mit interaktiven Simulationen (vgl. Abschnitt 3.1 auf Seite 29). Die Konzeption des Tests erfolgte in dieser Weise aufgrund vorliegender Ergebnisse zur Eignung von Testverfahren zur Messung experimenteller Kompetenz und zur Vorstrukturierung von Aufgaben (vgl. Abschnitt 2.2 auf Seite 22). Auch wenn die Aufgaben experimentelle Anforderungen stellen, die Schülerinnen und Schüler üblicherweise aus ihrem Physikunterricht kennen (vgl. Annahme I.II in Abbildung 6.2 auf Seite 68), wird ihnen das Aufgabenformat nicht bekannt sein. Um das Aufgabenformat kennenzulernen, bearbeiten die Schülerinnen und Schüler daher zu Beginn des Tests eine Trainingsaufgabe. Die Konzeption des Tests (v. a. konsekutive Struktur mit Zwischenlösungen) und die Trainingsaufgabe zum Kennenlernen des Aufgabenformats sollen sicherstellen, dass die folgende Annahme (II.II) beibehalten werden kann: *Die Bearbeitung des Tests ermöglicht das Zeigen experimenteller Performanz* (vgl. Abbildung 6.3 auf Seite 71). Annahme II.II wäre infrage gestellt, wenn die konkrete Umsetzung der Testkonzeption (z. B. Einbettung der Zwischenlösungen in eine Rahmengeschichte mit Alina und Bodo) oder der Trainingsaufgabe von Schülerinnen und Schüler nicht in der intendierten Weise wahrgenommen würde. Zur Prüfung von Annahme II.II in Kapitel 12 wird zum einen auf die Ergebnisse einer Schülerbefragung zur wahrgenommenen Bearbeitung des Tests (vgl. Studie C in Tabelle 7.1 auf Seite 79) zurückgegriffen. Zum anderen werden die im Rahmen der Abschlussarbeiten von Jaschinski (2013) und Eckloff (2014) durchgeführten Studien (vgl. Abschnitt 7.1) berücksichtigt, in denen die oben genannten Aspekte empirisch untersucht wurden.

Begründung von Annahme II.III

Im Physikunterricht werden beim Experimentieren in der Regel hands-on Aufgaben mit Realexperimenten eingesetzt. Hands-on Aufgaben mit Realexperimenten gelten daher als Referenzmaßstab, an dem sich alternative Formate (z. B. on-screen Aufgaben mit interaktiven Simulationen) im Hinblick auf Validitätsaspekte messen lassen müssen. Da der MeK-LSA Experimentiertest vollständig on-screen zu bearbeitende Aufgaben mit interaktiven Simulationen einsetzt, muss auf kognitiver Ebene folgende Annahme (II.III) überprüft werden: *Der Anteil experimentbezogener Überlegungen ist bei der Bearbeitung von on-screen Aufgaben vergleichbar hoch wie bei der Bearbeitung inhaltlich identischer hands-on Aufgaben* (vgl. Abbildung 6.3 auf Seite 71). Die Prüfung von Annahme II.III erfolgt in Kapitel 13 auf Basis von Think-Aloud-Daten (vgl. Studie C in Tabelle 7.1 auf Seite 79).

Begründung von Annahme II.IV

Damit experimentelle Aufgaben erfolgreich bearbeitet werden können, muss im Arbeitsgedächtnis genügend Kapazität für experimentbezogene Überlegungen zur Verfügung stehen. Die Theorie der kognitiven Belastung (*Cognitive-Load-Theory*; Übersicht: Sweller, Ayres & Kalyuga, 2011) geht von der Grundannahme aus, dass das Arbeitsgedächtnis nur eine begrenzte Verarbeitungskapazität hat und die dort enthaltenen Informationen nur für kurze

Zeit speichern kann. Unter gewissen Umständen (z. B. Schülerinnen und Schüler mit niedrigem Vorwissen) kann die eigentliche Kapazität des Arbeitsgedächtnisses überschritten werden (Haslam & Hamilton, 2010, S. 1716). Das kann passieren, wenn der Umfang und die Komplexität der zu verarbeitenden Informationen die Kapazität des Arbeitsgedächtnisses übersteigen (*cognitive overload*) (ebenda). Sowohl bei on-screen Aufgaben, als auch bei hands-on Aufgaben können formatspezifische Eigenschaften, z. B. durch die virtuellen Handlungen in den Simulationen oder durch das manuelle Handling im Realexperiment, zu einer nicht experimentbezogenen kognitiven Belastung des Arbeitsgedächtnisses führen. Für die eigentliche Bearbeitung der experimentellen Aufgabe könnte in der Folge in beiden Formaten ein unterschiedlicher Anteil an Denkkapazität zur Verfügung stehen. Ein unterschiedlicher Anteil an Denkkapazität beeinflusst möglicherweise das Zeigen experimenteller Performanz. Es ist daher folgende Annahme (II.IV) zu prüfen: *Die wahrgenommene kognitive Belastung ist bei on-screen Aufgaben vergleichbar hoch wie bei inhaltlich identischen hands-on Aufgaben* (vgl. Abbildung 6.3 auf Seite 71). Die Prüfung von Annahme II.IV erfolgt in Kapitel 14 auf Basis von Selbsteinschätzungsfragebögen zur wahrgenommenen kognitiven Belastung (vgl. Studien D & C in Tabelle 7.1 auf Seite 79).

6.3 Die beobachtete Performanz wird in geeigneter Art und Weise in Testwerte überführt (Teil III des INA)

Unter der Voraussetzung, dass die beobachtete Performanz zur beabsichtigten Performanz passt (vgl. Abschnitt 6.2), muss der nächste Schritt (von der Performanz zu den Testwerten) die folgende Qualitätsanforderung (übergeordnete Aussage) erfüllen: *Die beobachtete Performanz wird in geeigneter Art und Weise in Testwerte überführt*. Diese übergeordnete Aussage lässt sich durch zwei zu prüfende Annahmen (III.I und III.II) konkretisieren (vgl. Abbildung 6.4).

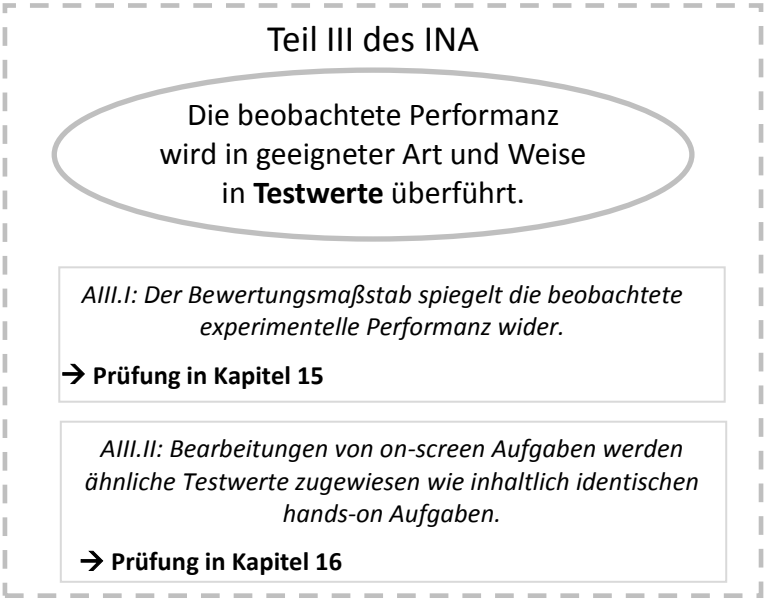


Abbildung *6.4: Teil III des INA mit den Annahmen III.I und III.II zur dritten übergeordneten Aussage (ovale Komponente; Prüfung der Annahmen in Kapitel 15 bis 16)*

Begründung von Annahme III.I

Die Zuweisung von Testwerten zur beobachteten Performanz erfolgt über einen Bewertungsmaßstab. Der Bewertungsmaßstab für den MeK-LSA Experimentiertest ist so konzipiert, dass die Schülerlösungen auf Basis der tatsächlichen Schülerhandlungen bewertet werden können (vgl. Abschnitt 3.3 auf Seite 34). Die Entwicklung des Bewertungsmaßstabs basiert allerdings primär auf normativen Setzungen des Testentwicklungsteams. Es ist daher folgende Annahme (III.I) zu prüfen: *Der Bewertungsmaßstab spiegelt die beobachtete experimentelle Performanz wider* (vgl. Abbildung 6.4 auf Seite 73). Die Prüfung von Annahme III.I erfolgt in Kapitel 15 durch einen Vergleich von Think-Aloud-Daten und der Zuweisung von Testwerten zu Schülerlösungen (vgl. Studie C in Tabelle 7.1 auf Seite 79).

Begründung von Annahme III.II

Auch wenn auf kognitiver Ebene gezeigt werden kann, dass der Anteil experimentbezogener Überlegungen bei on-screen und hands-on Aufgaben vergleichbar ist (vgl. Annahme II.III in Abbildung 6.3 auf Seite 71), bleibt zu untersuchen, ob Personen bei der Bearbeitung von on-screen gestellten Aufgaben ähnliche Leistungen erzielen wie bei der Bearbeitung inhaltlich gleicher hands-on Aufgaben (gemessen an den jeweiligen Testwerten). Daher ist folgende Annahme (III.II) zu prüfen: *Bearbeitungen von on-screen Aufgaben werden ähnliche Testwerte zugewiesen wie inhaltlich identischen hands-on Aufgaben* (vgl. Abbildung 6.4 auf Seite 73). Zur Prüfung von Annahme III.II werden in Kapitel 16 Zusammenhänge zwischen den Testwerten in beiden Formaten untersucht (vgl. Studien D & C in Tabelle 7.1 auf Seite 79).

6.4 Übersicht des INA

Eine Übersicht des INA ist in Abbildung 6.5 auf Seite 75 dargestellt. Die Übersicht zeigt die in den Abschnitten 6.1 bis 6.3 beschriebenen Qualitätsanforderungen (übergeordneten Aussagen) und Annahmen und veranschaulicht die Schritte von der Zieldomäne *Experimentieren im Physikunterricht der Sekundarstufe I* bis zur Zuweisung von Testwerten. Erst wenn die Prüfung aller Annahmen (vgl. Abschnitte 6.1 bis 6.3) abgeschlossen ist, kann bewertet werden, bis zu welchem Grad die Testwerte des MeK-LSA Experimentiertests valide als Ausdruck von Experimentierfähigkeiten interpretierbar sind.

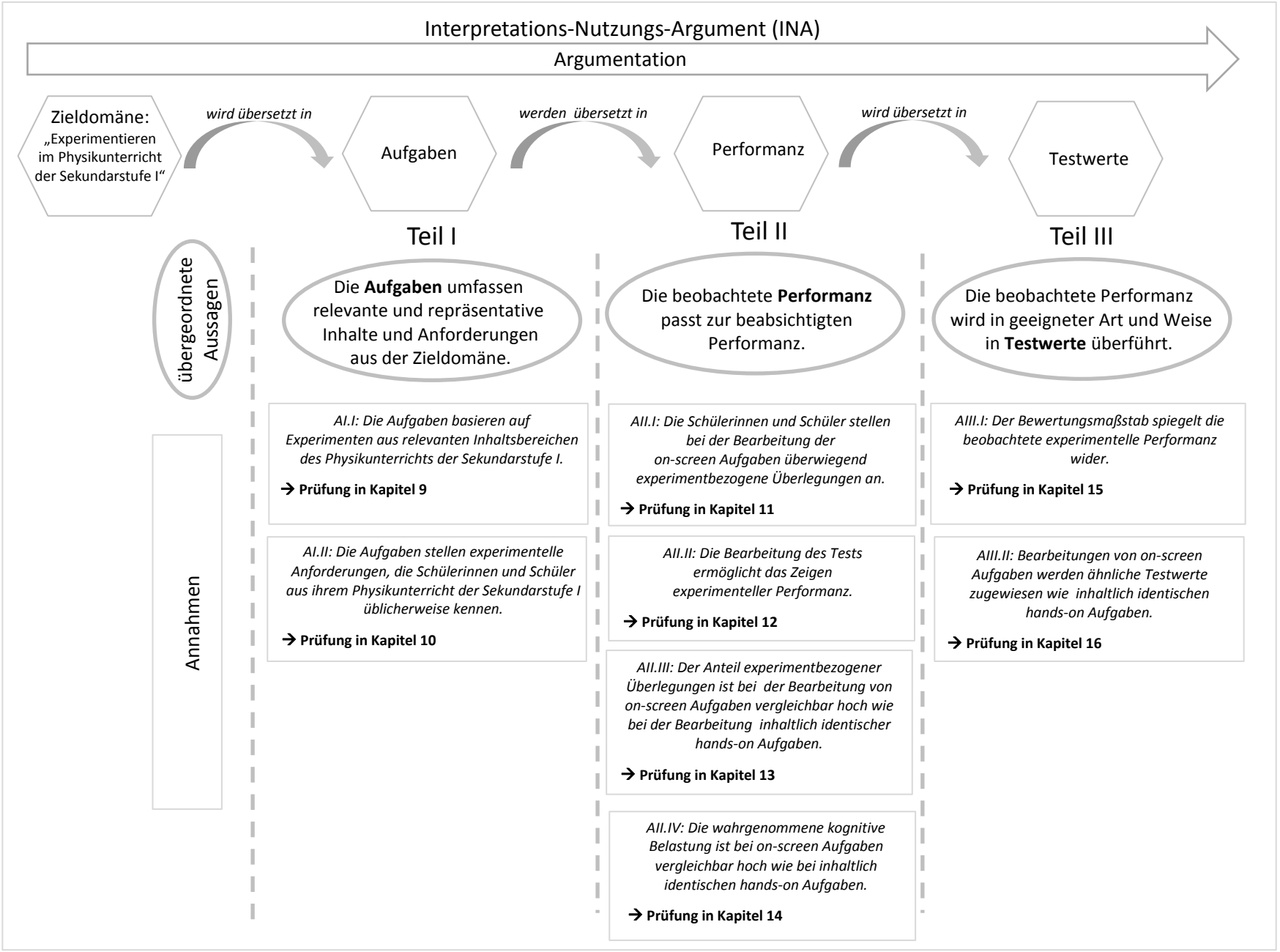


Abbildung 6.5: Übersicht über das INA (Die Argumentation wird entlang der Schritte zu den (übergeordneten) Aussagen geführt; nach der Prüfung aller Annahmen erfolgt eine Validitätsbewertung des MeK-LSA Experimentiertests)

7 Prüfung der Annahmen aus dem INA – eine Orientierungshilfe

Die im INA beschriebenen Annahmen (vgl. Abbildung 6.5 auf Seite 75) werden in den Kapiteln 9 bis 16 geprüft. Evidenz zur Prüfung der Annahmen wurde in fünf Studien gewonnen, die im Rahmen des Projekts MeK-LSA durchgeführt wurden. Darüber hinaus sind im Projekt MeK-LSA zwölf Abschlussarbeiten (Bachelor-, Master- und Staatsexamensarbeiten) entstanden. Fünf dieser Arbeiten liefern aufgrund der untersuchten Fragestellungen einen Beitrag zur Prüfung von Annahmen aus dem INA. In Abschnitt 7.1 wird der Bezug zwischen den im Projekt MeK-LSA durchgeführten Studien, den Abschlussarbeiten und dem INA hergestellt. Um den bereits sehr hohen Aufwand einzugrenzen, konnten in den meisten Studien nicht alle Aufgaben und Teilaufgaben berücksichtigt werden. In Abschnitt 7.2 wird die Auswahl der untersuchten Aufgaben und Teilaufgaben begründet. In zwei Studien wurden hands-on Aufgaben mit Realexperimenten als Referenzmaßstab für die on-screen Aufgaben mit interaktiven Simulationen eingesetzt. Um vergleichbare Testbedingungen herzustellen, wurde für diese Studien das on-screen Format in ein vergleichbares hands-on Format überführt (vgl. Abschnitt 7.3).

7.1 Durchgeführte Studien und Abschlussarbeiten – Bezug zum INA

Zur Prüfung der Annahmen aus dem INA sind im Projekt MeK-LSA fünf Studien (A bis E) durchgeführt worden, die sich bezogen auf die Stichprobe, die Erhebungsmethode und die erfassten Daten unterscheiden (vgl. Tabelle 7.1 auf Seite 79). Zusätzlich zu diesen fünf Studien tragen fünf der zwölf im Projekt MeK-LSA durchgeführten Abschlussarbeiten aufgrund der untersuchten Fragestellungen maßgeblich zur Prüfung von Annahmen aus dem INA bei. Die Arbeiten von Jaschinski (2013) und Eckloff (2014) beinhalten eigene Datenerhebungen, die in enger Absprache mit dem Projekt MeK-LSA erfolgten (Studien F und G; vgl. Tabelle 7.1). Die Arbeiten von Jansen (2014), Türck (2014) und Zirwes (2014) waren in Studien aus dem Projekt MeK-LSA integriert und beinhalten keine eigenen Datenerhebungen. Jansen (2014) hat Schülerleistungen bei der Erstellung von Messwertediagrammen im on-screen und hands-on Format verglichen und damit Vorarbeiten zur Entwicklung des Bewertungsmaßstabes für die Diagrammaufgaben geleistet sowie qualitative Hinweise zur Vergleichbarkeit beider Formate geliefert (Beitrag zu Studie E). Türck (2014) und Zirwes (2014) gingen der Frage nach, womit sich Schülerinnen und Schüler bei der Bearbeitung einer hands-on Experimentieraufgabe bzw. von on-screen Experimentieraufgaben beschäftigen. Beide Arbeiten haben einen Beitrag zur Ausarbeitung und Erprobung eines Kategoriensystems für die Analyse kognitiver Prozesse geleistet (Beitrag zu Studie C). Mit Ausnahme der Masterarbeit von Eckloff (2014), die an der Universität Bremen betreut wurde, wurden alle weiteren in dieser Arbeit verwendeten Abschlussarbeiten an der Universität Duisburg-Essen vom Autor dieser Dissertation inhaltlich beraten und begleitet.

Um bei der Prüfung der Annahmen aus dem INA (vgl. Abbildung 6.5 auf Seite 75) eine mehrfache Beschreibung der im Projekt MeK-LSA durchgeführten Studien (A bis E) zu vermeiden, werden diese in Kapitel 8 im Sinne eines Studien-Glossars beschrieben. Die Beschreibung konzentriert sich auf die Aspekte der Studien, die nicht zu eng mit einer

konkreten Annahme aus dem INA verwoben sind. Das sind z. B. der Ablauf, die erhobenen Daten, die untersuchten (Teil-)aufgaben und die Stichprobe. Eine Sonderrolle nimmt Studie A ein. Der Ablauf und die Ergebnisse dieser Studie wurden bereits in der Entwicklungsphase (Teil I; vgl. Abschnitt 4.6 auf Seite 48) beschrieben bzw. berücksichtigt. In Kapitel 8 wird diese Studie ein zweites Mal aufgeführt, da sie auch bei der Prüfung von Annahmen des INA – aus dem Blickwinkel der Validität – eine Rolle spielt (vgl. Tabelle 7.1 auf Seite 79). Die Studien F und G aus den Abschlussarbeiten (vgl. Tabelle 7.1) und alle konkret mit einer Annahme aus dem INA verbundenen Aspekte einer Studie (z. B. Kategoriensysteme, Ergebnisse) werden in den jeweiligen Kapiteln (9-16) bezogen auf die zu prüfenden Annahmen beschrieben und diskutiert.

Tabelle 7.1: Bezug der im Projekt MeK-LSA durchgeführten Studien und Abschlussarbeiten zum INA (Kurzform; Studien A bis E: Im Projekt MeK-LSA durchgeführte Studien; Studien F und G: im Rahmen von Abschlussarbeiten durchgeführte Studien; INA-Bezug in rechter Spalte)

	Studie	Stichprobe	Erhebungsmethode		erfasste Daten		INA-Bezug: Prüfung von Annahme…
Projekt MeK-LSA	A	53 erfahrene Lehrkräfte	Lehrkräftebefragung		Bekanntheit der Experimente aus Aufgabenskizzen; Erfüllbarkeit der experimentellen Anforderungen aus Aufgabenskizzen		I.I & I.II
	B	1194 Schülerinnen und Schüler	Schülerbefragung		Bekanntheit der Experimente aus Testaufgaben		I.I
	C	106 Schülerinnen und Schüler	Bearbeitung von MeK-LSA Testaufgaben (Auswahl)	Think-Aloud; Schülerbefragung	Qualität der Aufgabenbearbeitungen	Art und Regelbasiertheit der Überlegungen; wahrgenommene Bearbeitung des Tests; experimentelle Anforderungen im Unterricht	I.II & II.I & II.II & II.III & III.I
	D	42 Biologiestudierende		Selbsteinschätzungs-fragebogen		wahrgenommene kognitive Belastung	II.IV & III.II
	E	19 Schülerinnen und Schüler					II.IV & III.II
Abschlussarbeiten	F	13 Schülerinnen und Schüler		Think-Aloud; Schülerbefragung	Einfluss einer Trainingsaufgabe auf die Testbearbeitung		II.II
	G	10 Schülerinnen und Schüler		Think-Aloud; nicht konsekutive Testaufgabe	Art und Regelbasiertheit der Überlegungen		II.II

7.2 Auswahl zu analysierender Aufgaben und Teilaufgaben

In den Studien C, D und E (vgl. Tabelle 7.1 auf Seite 79) konnten aufgrund der erforderlichen zeit- und ressourcenintensiven Datenerhebungs- und -auswertungsmethoden nicht alle Aufgaben und Teilaufgaben berücksichtigt werden. Um im Hinblick auf die Überprüfung des INA trotzdem aussagekräftige Daten zu erhalten, kommt der Auswahl der in den Studien untersuchten und analysierten Aufgaben und Teilaufgaben eine bedeutende Rolle zu. Zum einen sollen die ausgewählten (Teil-)Aufgaben wesentliche Aspekte des Testverfahrens repräsentieren, zum anderen sollen (Teil-)Aufgaben berücksichtigt werden, bei denen die zu prüfenden Annahmen einem besonders kritischen Test unterzogen werden können. In Anlehnung an Methoden zur Fallauswahl aus der qualitativen Sozialforschung (z. B. Schreier, 2010, S. 238-251) wurde ein qualitativer (Teil-)Aufgabenumfangsplan erstellt. Dieser legt Kelle und Kluge (2010, S. 50) folgend relevante Merkmale und Merkmalsausprägungen der (Teil-)Aufgaben fest. In den Festlegungsprozess fließen Vorwissen über das Forschungsfeld *Messung experimenteller Kompetenz* und Wissen über die zu prüfenden Annahmen ein.

Für die Messung experimenteller Kompetenz zeigt sich nach Schreiber et al (2014, S. 163), dass die Leistungen von Schülerinnen und Schülern bei der Bearbeitung experimenteller Aufgabenstellungen u. a. von Aufgabeninhalten (z. B. Shavelson et al. 1993) und dem erforderlichem aufgabenspezifischen Fachwissen (z. B. Gott & Duggan, 2002) abhängen. Vor dem Hintergrund der zu prüfenden Annahmen aus dem INA (vgl. Abbildung 6.5 auf Seite 75) sind in jedem Fall diese beiden Einflussfaktoren zu berücksichtigen. Aus jedem der drei Inhaltsbereiche (Elektrizitätslehre, geometrische Optik, Mechanik) wurde daher mindestens eine Aufgabe berücksichtigt[15].

Weiterhin ist zu klären, welchen Einfluss das Testformat auf die Schülerleistungen hat (z. B. Shavelson et al. 1999). Zur Prüfung der Annahmen aus dem INA sind daher insbesondere die Teilaufgaben zu analysieren, in denen die Schülerinnen und Schüler virtuelle, aber mit dem Realexperiment vergleichbare, Handlungen zur Lösung der Aufgabe ausführen müssen. Die Analyse beschränkt sich daher in der Regel auf die Teilaufgaben *Versuchsplan entwerfen*, *Versuch aufbauen und testen* und *Messung durchführen und dokumentieren.* Um den Einfluss des Testformats untersuchen zu können, wurden zwischen hands-on und on-screen Aufgaben weitgehend vergleichbare Testbedingungen hergestellt (vgl. Abschnitt 7.3). Vergleichbare Bedingungen in beiden Formaten sind eine wichtige Voraussetzung für die Äquivalenz der Formate (vgl. International Test Commission, 2001). Bei der Aufgabe *Ausdehnung eines Gummibandes* zeigten sich jedoch in der Oberflächenstruktur Unterschiede bzgl. der Handlungsmöglichkeiten im on-screen- und hands-on-Format. Im Sinne einer Worst-Case Abschätzung wird daher insbesondere diese Aufgabe zum Vergleich der Bearbeitung von hands-on- und on-screen-Teilaufgaben herangezogen. Ein möglicher Einfluss auf die kognitiven Prozesse sollte sich dort besonders deutlich zeigen. Unterschiede in der Oberflächenstruktur sind auch bei den Teilaufgaben zum *Datenauswertung durchführen* zu erkennen, die die

[15] Die Aufgabe *Ausdehnung eines Gummibandes* wurde mit Ausnahme der Studie E (Diagrammstudie) sowie der Studien F und G aus den Abschlussarbeiten, in allen Studien berücksichtigt.

Anfertigung eines Diagramms erfordern. Die Teilaufgabe *Datenauswertung durchführen* wird daher ergänzend analysiert.

Die anderen im MeK-LSA Experimentiertest enthaltenen Teilaufgaben (*Grundidee skizzieren*, *Messprotokoll vorbereiten*[16], *Datenauswertung planen* und *Schlüsse ziehen*) sind sowohl bei den on-screen, als auch bei den hands-on Aufgaben rein schriftlich zu bearbeiten. Ein Unterschied besteht nur in der Art der Dateneingabe (on-screen: eintippen bzw. klicken; hands-on: mit einem Stift auf Papier schreiben bzw. ankreuzen). Die Übertragung von rein schriftlichen Aufgaben in rein schriftliche on-screen Aufgaben wird in der Literatur unter dem Begriff *Äquivalenzproblematik* diskutiert (z. B. Jurecka & Hartig, 2007, S. 41-43). Die Frage nach der Äquivalenz ist für Leistungstests einerseits noch nicht abschließend beantwortet (ebenda). Andererseits konnten Mead und Drasgow (1993) in einer Metaanalyse zumindest für Power-Tests zeigen, dass Äquivalenz angenommen werden kann. Da der MeK-LSA Experimentiertest als Power- und nicht als Speed-Test konzipiert ist, ist die Äquivalenz für die rein schriftlichen Teilaufgaben wahrscheinlich.

7.3 Übersetzung des on-screen Formats in ein vergleichbares hands-on Format

Bei der Übersetzung des on-screen Formats in ein vergleichbares hands-on Format bezieht sich die Vergleichbarkeit der Testbedingungen auf die Aufgabenstellungen, die Funktionalität und das Aussehen der Geräte und Materialien sowie den Ablauf des Tests (z. B. Abfolge der Teilaufgaben). Um einen mit den on-screen Aufgaben vergleichbaren Ablauf des Tests zu ermöglichen, wurde für die hands-on Aufgaben ein Stationenlauf realisiert (vgl. Abbildung 7.1). Jede Station im hands-on Format entspricht einer Teilaufgabe im on-screen Format. Das schließt auch das Bereitstellen von Zwischenlösungen vorheriger Teilaufgaben mit ein.

Abbildung 7.1: Station zur Teilaufgabe Versuch aufbauen und testen der Aufgabe Ausdehnung eines Gummibandes *(oben: Materialien; unten: Informationen aus dem Aufgabenstamm und Aufgabenstellung mit Zwischenlösung)*

[16] mit Einschränkung

8 Studien zur Prüfung von Annahmen aus dem INA (Glossar)

In Kapitel 8 werden für die Studien A bis E (vgl. Tabelle 7.1 auf Seite 79) die erhobenen Daten, die untersuchte Stichprobe, die untersuchten (Teil-)Aufgaben und der generelle Ablauf beschrieben, um eine mehrfache Beschreibung dieser Studienaspekte bei der Prüfung der Annahmen aus dem INA zu vermeiden. Dieses Kapitel ist explizit als Studien-Glossar zu verstehen, d. h. die Studien werden nacheinander aufgelistet und entlang der oben genannten Merkmale beschrieben. Die konkreten Fragestellungen, die mit den Studien beantwortet werden, und eine mögliche Verknüpfung zwischen einzelnen Studien ergeben sich aus dem INA und werden in den Kapiteln 9 bis 16 zur Prüfung der Annahmen aus dem INA aufgegriffen.

8.1 Studie A: Lehrkräftebefragung

In Studie A wurden 53 Physiklehrkräfte aus neun Bundesländern, bereits während der Überführung der Testkonzeption in Testaufgaben (vgl. Kapitel 4 auf den Seiten 37-53), zur Bekanntheit und zu den Anforderungen der Experimente aus den Aufgabenskizzen befragt (vgl. Abschnitt 4.6 auf Seite 48). Die Lehrkräfte wurden gebeten, sich bei ihrer Einschätzung auf *typische Schülerinnen und Schüler* (ohne nähere Erläuterung) der 9. Klasse an ihrer Schule zu beziehen. Auf Basis dieser Informationen schätzten die Lehrkräfte für jede Aufgabenskizze fünf Fragen auf einer vierstufigen Rating-Skala (1 ≙ sehr unwahrscheinlich bis 4 ≙ sehr wahrscheinlich) ein (vgl. Abbildung 4.7 auf Seite 49). Die ersten beiden Fragen beziehen sich auf die Bekanntheit der Experimente für die Schülerinnen und Schüler aus Lehrkraftperspektive. Die letzten drei Fragen zielen auf die Erfüllbarkeit der experimentellen Anforderungen in den Bereichen Planung, Durchführung und Auswertung eines Experiments ab.

8.2 Studie B: Large-Scale

Nach Abschluss der Studien A, C, D, und E wurde der MeK-LSA Experimentiertest in vier Bundesländern (Bremen, Niedersachsen, Nordrhein-Westfalen und Schleswig-Holstein) mit 1194 Schülerinnen und Schülern in einer Large-Scale Studie erprobt. Die Teilnahme erfolgte im Klassenverband (56 Klassen) während der regulären Schulzeit. In dieser Studie wurden alle zwölf ausgearbeiteten Testaufgaben eingesetzt. Eine Kurzbeschreibung der übergeordneten Aufgabenstellungen der Testaufgaben findet sich in Tabelle 4.5 auf Seite 52. Jede Aufgabe wurde von ca. 400 Schülerinnen und Schülern bearbeitet. Nach der Bearbeitung jeder Aufgabe wurde eine Befragung zur Bekanntheit der Experimente durchgeführt. Eine ausführliche Beschreibung des Ablaufs und der Ergebnisse der Large-Scale Studie erfolgt bei Eickhorst (in Vorbereitung). Für die vorliegende Arbeit ist unter Validierungsaspekten nur die Befragung zur Bekanntheit der Experimente von Interesse.

8.3 Studie C: Aufgabenbearbeitungsprozesse

In Studie C sind u. a. die bei der Bearbeitung der Aufgaben auf Seiten der Schülerinnen und Schüler ablaufenden Überlegungen erfasst worden. Zur Erfassung der bei der Bearbeitung der Aufgaben ablaufenden Überlegungen wurden die Methoden begleitendes und nachträgliches

Think-Aloud eingesetzt (vgl. Abschnitt 8.3.1: Zur Methode der Analyse kognitiver Prozesse). Nach der Bearbeitung einer Trainingsaufgabe zum Kennenlernen des Aufgabenformats und einer Trainingsübung zum begleitenden Think-Aloud bearbeiteten die Schülerinnen und Schüler in Einzelarbeit in der Regel zwei Aufgaben mit jeweils acht Teilaufgaben. Die Schülerinnen und Schüler wurden dazu aufgefordert, während der Bearbeitung der Teilaufgaben alle Überlegungen laut auszusprechen, die ihnen durch den Kopf gehen. Eine Zeitbegrenzung für die Bearbeitung einzelner Teilaufgaben gab es nicht. Jede Schülerin und jeder Schüler wurde von einem geschulten Testleiter begleitet und unmittelbar nach der Bearbeitung jeder Teilaufgabe auf Basis einer *Checkliste* dazu aufgefordert, ihre bzw. seine Überlegungen zu zentralen Schritten der Aufgabenbearbeitung zu äußern (z. B. Vorgehensweise bei der Geräteauswahl, Anzahl der dokumentierten Messwertepaare). Im Anschluss an das nachträgliche Think-Aloud sind die Schülerinnen und Schüler zu ihrer Wahrnehmung der Testsituation und zu experimentellen Anforderungen aus dem Unterricht befragt worden. Die Bearbeitungen und alle Äußerungen wurden über Bildschirmaufzeichnungen und Audiomitschnitte dokumentiert.

An Studie C haben insgesamt 106 Schülerinnen und Schüler der Sekundarstufe I aus drei Bundesländern (Bremen, Niedersachsen & Nordrhein-Westfalen) teilgenommen. Die Teilnahme erfolgte auf freiwilliger Basis (Einzelschüler) außerhalb des regulären Unterrichts. Dabei wurden die folgenden vier on-screen Aufgaben des MeK-LSA Experimentiertests untersucht: *Ausdehnung eines Gummibandes*, *Reihenschaltung von Glühlampen*, *Leistung von Glühlampen* und *Brechung am Halbkreisblock*. Die Aufgabe *Ausdehnung eines Gummibandes* wurde zusätzlich in ein mit dem on-screen Format vergleichbares hands-on Format übertragen. Im hands-on Format kommen Realexperimente anstelle von interaktiven Simulationen zum Einsatz (vgl. Abschnitt 7.3 auf Seite 81 zur Gestaltung des hands-on Formats). Die Auswahl der eingesetzten Aufgaben wurde in Abschnitt 7.2 auf Seite 80 begründet.

Aufgrund der hohen Anzahl auswertbarer Bearbeitungen von Teilaufgaben aus dem begleitenden Think-Aloud (vgl. Abschnitt 11.1.2 auf Seite 104) wurde auf eine Analyse der Daten aus dem nachträglichen Think-Aloud verzichtet. Die Überlegungen aus dem begleitenden Think-Aloud versprechen eine zuverlässigere Annäherung an die tatsächlich abgelaufenen Überlegungen (vgl. Abschnitt 8.3.1: Zur Methode der Analyse kognitiver Prozesse).

8.3.1 Zur Methode der Analyse kognitiver Prozesse

Zur Analyse der Überlegungen eines Individuums bei der Bearbeitung von Aufgaben bietet sich die Methode Think-Aloud an (z. B. Ericsson & Simon, 1993). Schülerinnen und Schüler werden z. B. beim Aufbauen eines Versuchs dazu aufgefordert, alle Gedanken laut auszusprechen, um einen Zugang zu ihren Überlegungen während des Aufbauens zu erhalten. Theoretisch gerahmt wird die Methode durch ein Modell menschlicher Informationsverarbeitung (z. B. Woolfolk, 2008, S. 309-313), das zwischen sensorischem Register, Kurzeitgedächtnis (Arbeitsgedächtnis) und Langzeitgedächtnis unterscheidet. Innerhalb dieses Modells nehmen

Ericsson und Simon (1993, S. 11) an, dass im Kurzzeitgedächtnis gespeicherte Informationen für weitere Verarbeitungsprozesse (z. B. Verbalisierungen) unmittelbar zur Verfügung stehen, während im Langzeitgedächtnis gespeicherte Informationen erst wieder ins Kurzzeitgedächtnis überführt werden müssen. Aufbauend auf diesen Annahmen unterscheiden Ericsson und Simon (1993, S. 17) drei Verbalisierungslevel, die beschreiben, was verbalisiert wird (vgl. Tabelle 8.1).

Tabelle 8.1: Verbalisierungslevel nach Ericsson und Simon (1993, S. 17; eigene Darstellung mit Beispielen für Verbalisierungsaufforderungen in den Levels)

Verbalisierungslevel		
Level 1	Level 2	Level 3
Talk Aloud	Think Aloud	Verbalization Procedures that involve mediating processes before verbalization
z. B. *Sprich alles laut aus…*		z. B. *Sprich nur das laut aus…*
…was du während des Aufbaus leise zu dir selbst sagst.	*...was dir während des Aufbaus durch den Kopf geht, ohne deine Überlegungen zu erklären oder zu strukturieren.*	*…was dir durch den Kopf geht, wenn du Änderungen am Aufbau vornimmst.*

Beim *Talk-Aloud* liegen die Informationen im Kurzzeitgedächtnis bereits in verbal enkodierter Form vor und können direkt laut ausgesprochen werden, während beim Think-Aloud nonverbale Informationen aus dem Kurzzeitgedächtnis zunächst in verbale Informationen überführt werden müssen (Konrad, 2010, S. 479). Der Überführungsprozess in verbale Informationen findet beim Think-Aloud ohne weitere Reflexionsprozesse statt. Im Gegensatz dazu laufen bei Level 3-Verbalisierungen vor dem Verbalisieren noch weitere Verarbeitungsprozesse ab. Für die Annäherung an die tatsächlich während einer Handlung (z. B. Aufbauen eines Versuchs) ablaufenden Überlegungen eignen sich Level 3-Verbalisierungen nicht, da diese die an der Handlung beteiligten Prozesse verlangsamen und verändern (z. B. Bannert, 2007, S. 135). Eine Verlangsamung der an den Handlungen beteiligten Prozesse ist auch beim Think-Aloud zu beobachten. Ericsson und Simon (1993, S. 16) nehmen allerdings an, dass sich die Überlegungen durch Level 1- und Level 2-Verbalisierungen nicht verändern. Folgt man dieser Annahme, so lassen sich Verbalisierungen auf diesen beiden Leveln als beobachtbare Indikatoren für die tatsächlich ablaufenden Überlegungen betrachten. Nach Buber (2007) konnten auch *„einschlägige empirische Untersuchungen keine fundamentalen Unterschiede gegenüber Entscheidungsprozessen ohne lautes Denken finden“* (S. 562).

Bei der Think-Aloud-Methode bestehen allerdings auch Einschränkungen, die bei der Interpretation von Think-Aloud-Protokollen berücksichtigt werden müssen. Zum Beispiel können keine Aussagen über die Vollständigkeit der Think-Aloud-Protokolle gemacht werden, da davon ausgegangen werden muss, dass automatisierte und routinierte Prozesse unbewusst ablaufen und somit nicht verbalisiert werden (Sandmann, 2014, S. 188). Auch Überlegungen

höherer Ordnung (z. B. Informationsverarbeitungsstrategien) können aufgrund ihrer Komplexität nicht immer verbalisiert werden (Konrad, 2010, S. 486). Letztendlich kann auch die Verbalisierungsfähigkeit der Probandinnen und Probanden Auswirkungen auf die Vollständigkeit der Think-Aloud-Protokolle haben (Sandmann, 2014, S. 188).

Eine weitere Annäherungsmöglichkeit an die Überlegungen bietet das nachträgliche Think-Aloud (auch *Stimulated-Recall*; z. B. Gass & Mackey, 2000) an. Zum Beispiel werden die Schülerinnen und Schüler bei dieser Methode erst unmittelbar nach dem Aufbauen des Versuchs zum Verbalisieren ihrer während des Aufbauens abgelaufenen Überlegungen aufgefordert. Um Vergessens- und Inferenzprozesse zu minimieren, wird den Schülerinnen und Schülern Stimulus-Material ihrer Bearbeitung zur Verfügung gestellt. Stimulus-Material kann z. B. das Abspielen der aufgezeichneten Handlungen (z. B. Video des Aufbauens) oder das bearbeitete Aufgabenmaterial (z. B. der fertige Aufbau) sein. Mit dieser Methode gelingt es, Informationen über die Überlegungen von Schülerinnen und Schülern zu erhalten, ohne den eigentlichen Bearbeitungsprozess zu stören (vgl. Bannert, 2007, S. 131). Die Methode bietet sich insbesondere für Schülerinnen und Schüler an, die ihre Überlegungen beim begleitenden Think-Aloud nicht in ausreichendem Maße verbalisieren können. Durch das Bereitstellen von Stimulus-Material besteht allerdings die Gefahr, dass die Schülerinnen und Schüler nicht nur die tatsächlich abgelaufenen Überlegungen wiedergeben. Die Schülerinnen und Schüler geben beim nachträglichen Think-Aloud beispielsweise auch Überlegungen wieder, die sie erst durch das Stimulus-Material entwickeln oder die sie nachträglich entwickeln konnten, weil sie wussten, zu welchem Ergebnis ihre Handlungen geführt haben.

Ist man an den während der Handlung ablaufenden Überlegungen interessiert, sollte – soweit umsetzbar – das begleitende Think-Aloud eingesetzt werden, da dabei die tatsächlichen Überlegungen am wenigsten verfälscht sind. Da Studien zur Untersuchung von Überlegungen, die während einer Handlung ablaufen, mit hohem Aufwand verbunden sind und es keine a priori Erfolgsgarantie für das begleitende Think-Aloud gibt, ist die Methode des nachträglichen Think-Aloud eine geeignete Ergänzung. Allerdings kann auch durch Kombination von begleitendem und nachträglichem Think-Aloud nicht garantiert werden, dass die Äußerungen der Schülerinnen und Schüler die Überlegungen vollständig abbilden.

8.4 Studie D: Testformatvergleich

In Studie D wurde untersucht, ob Personen bei der Bearbeitung von on-screen gestellten Aufgaben ähnliche Leistungen erzielen und eine ähnliche kognitive Belastung wahrnehmen wie bei der Bearbeitung inhaltlich gleicher hands-on Aufgaben. Die Studie wurde an der RWTH-Aachen nach dem ersten Drittel des physikalischen Praktikums für Biologie- und Biotechnologiestudierende[17] durchgeführt. Insgesamt haben 42 Biologiestudierende teilgenommen. Die Teilnahme erfolgte während der regulären Praktikumszeit innerhalb mehrerer Praktikumskohorten (6 bis 8 Studierende pro Kohorte). Die Studie beschränkt sich auf die Teilaufgabentypen *Versuchsplan entwerfen*, *Versuch aufbauen und testen* und *Messung*

[17] Im Folgenden kurz als Biologiestudierende bezeichnet. Gemeint sind weiterhin Biologie- und Biotechnologiestudierende.

durchführen und dokumentieren. Die drei Teilaufgabentypen wurden für fünf Aufgaben aus drei Inhaltsbereichen (Mechanik, Elektrizitätslehre, geometrische Optik) sowohl im on-screen Format als auch in einem vergleichbar gestalteten hands-on Format eingesetzt (vgl. Abschnitt 7.3 auf Seite 81 zur Gestaltung des hands-on Formats). Die Auswahl der Aufgaben und Teilaufgaben(typen) wurde in Abschnitt 7.2 auf Seite 80 begründet. Vor der Bearbeitung eines Teilaufgabentyps (z. B. Versuch aufbauen und testen) erhielten die Studierenden, anhand einer Trainingsaufgabe zum jeweiligen Teilaufgabentyp, eine kurze Einführung in die Bedienung der interaktiven Simulationen. Abbildung 8.1 zeigt den Ablaufplan der Studie exemplarisch für einen der drei Teilaufgabentypen (z. B. Versuch aufbauen und testen).

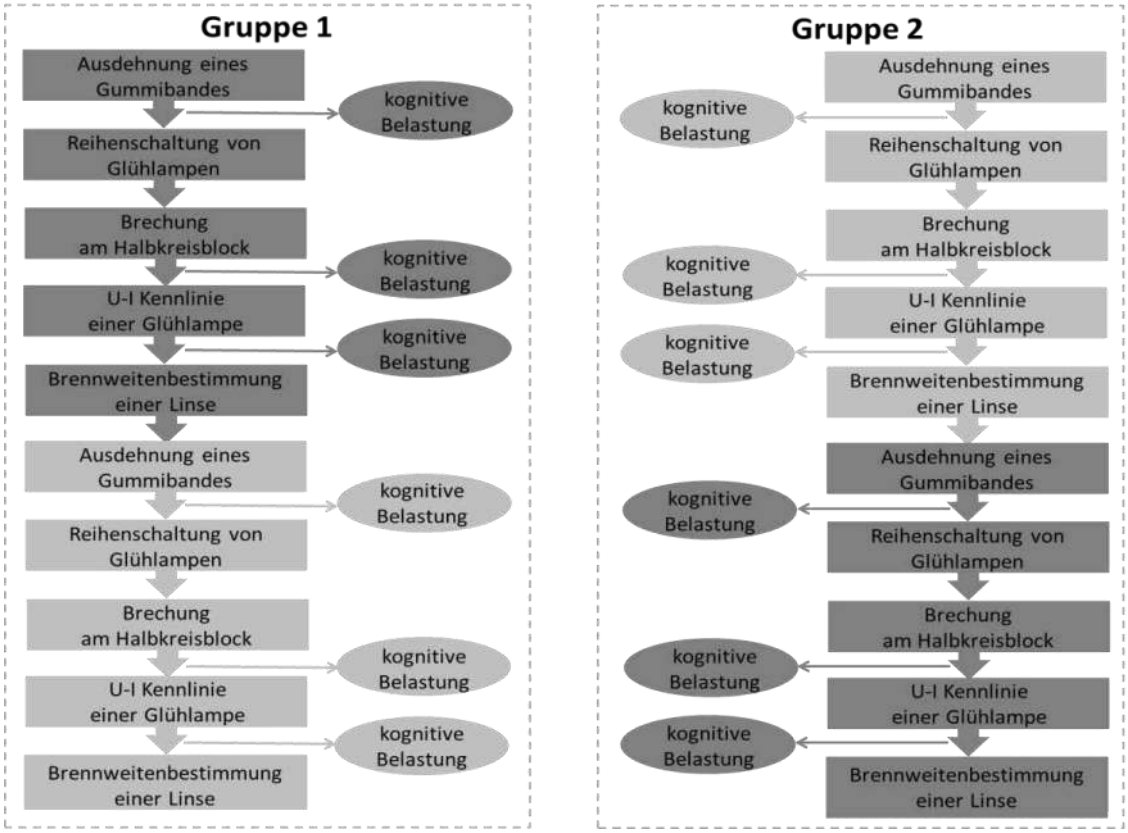


Abbildung 8.1: exemplarischer Ablauf von Studie D (dunkelgrau: hands-on Format; hellgrau: on-screen Format; Rechtecke: Bearbeitungen zu einem Teilaufgabentyp (z. B. Versuch aufbauen und testen); Ovale ≙ Einschätzung der wahrgenommenen kognitiven Belastung)

Für jeden Teilaufgabentyp bearbeiteten die Studierenden in Einzelarbeit fünf Aufgaben in einem Format (z. B. on-screen) und anschließend die gleichen Aufgaben im anderen Format (z. B. hands-on). Nach der Bearbeitung einer Teilaufgabe (z. B. Versuch aufbauen und testen für die Aufgabe *Ausdehnung eines Gummibandes*) wurden keine Zwischenlösungen zur Verfügung gestellt, um Lerneffekte für die Bearbeitung der gleichen Teilaufgabe im nachfolgenden Format zu minimieren. Um einen möglichen Einfluss der Formatreihenfolge kontrollieren zu können, wurde diese abgewechselt (vgl. Abbildung 8.1: Gruppe 1 und Gruppe 2). Die Bearbeitungszeit war auf sechs Minuten pro Teilaufgabe begrenzt. Zusätzlich zur Bearbeitung der Teilaufgaben schätzten die Studierenden in jedem Format die wahrgenommene kognitive Belastung nach drei Teilaufgaben ein (vgl. Abbildung 8.1). Die wahrgenommene kognitive Belastung wurde durch etablierte Selbsteinschätzungsskalen

erfasst (vgl. Abschnitt 8.4.1: Zur Methode der Messung kognitiver Belastung). Um den zeitlichen Aufwand für die Studierenden zu begrenzen, durchliefen alle Studierenden den in Abbildung 8.1 auf Seite 87 dargestellten Ablauf jeweils für zwei Teilaufgabentypen (z. B. *Versuchsplan entwerfen* und *Versuch aufbauen und testen*). Die Bearbeitungen wurden durch Bildschirmaufzeichnungen (on-screen) bzw. Videomitschnitte (hands-on) dokumentiert.

8.4.1 Zur Methode der Messung kognitiver Belastung

In der Forschung zur kognitiven Belastung haben sich nach Sweller et al. (2011, S. 71-85) zwei Verfahren zur Messung der kognitiven Belastung etabliert:

1. Einsatz von Sekundäraufgaben (*dual-task approach*).
2. Selbsteinschätzung der wahrgenommenen kognitiven Belastung.

Beim *dual-task approach* wird eine Sekundäraufgabe eingesetzt, die sich deutlich von der Primäraufgabe (z. B. einer experimentellen Aufgabenstellung) unterscheidet. Häufig geht es bei der Sekundäraufgabe um die Bestimmung der Reaktionszeit, bis die Änderung einer visuellen oder auditiven Komponente erkannt wird (z. B. Brunken, Plass & Leutner, 2003; Marcus, Cooper & Sweller, 1996). Die Reaktionszeit wird dann als Maß für die kognitive Belastung verwendet. Ein Vorteil des *dual-task approach* ist, dass die kognitive Belastung unmittelbar während der Aufgabenbearbeitung gemessen werden kann (Sweller et al., 2011, S. 85). Anderseits ist die Implementierung dieses Ansatzes mit hohem Aufwand verbunden (ebenda). Weiter verbreitet als der *dual-task approach* sind einfach umzusetzende Selbsteinschätzungsskalen zur wahrgenommenen kognitiven Belastung. Zur Selbsteinschätzung der wahrgenommenen kognitiven Belastung werden häufig zwei unterschiedliche Items eingesetzt, das Mental-Effort Item zur wahrgenommenen Denkanstrengung von Paas (1992) und das Difficulty Item von Kalyuga, Chandler und Sweller (1999) zur wahrgenommenen Aufgabenschwierigkeit. Die wahrgenommene Denkanstrengung wird bei Paas (1992) auf einer neunstufigen Rating-Skala erfasst, die wahrgenommene Aufgabenschwierigkeit bei Kalyuga et al. (1999) auf einer siebenstufigen Rating-Skala.

Van Gog, Kirschner, Kester und Paas (2012, S. 833-839; nur Mental-Effort Item) empfehlen auf Basis ihrer empirischen Befunde, die Items unmittelbar nach jeder Aufgabe einer Aufgabenserie vorzulegen und über die jeweiligen Einschätzungen zu mitteln. Eine globale Einschätzung nach allen Aufgaben scheint im Vergleich eine weniger reliable Messung zu ergeben. Sweller et al. (2011, S. 85) heben hervor, dass Selbsteinschätzungsfragebögen wertvolle Indikatoren für die kognitive Belastung unmittelbar nach der Aufgabenbearbeitung sein können, allerdings nicht die tatsächliche kognitive Belastung im Bearbeitungsprozess erfassen können.

8.4.2 Begründung der Stichprobe

Der Grund für die Wahl Biologiestudierender als Stichprobe für Studie D besteht in dem hohen logistischen und materiellen Aufwand, der erforderlich ist, um die realen Experimentiermaterialien für die hands-on Aufgaben bereitzustellen. Befragungen von Schwarz, Effertz und Heinke (2013, S. 2) sowie Borawski, Theyßen und Heinke (2005, S. 2) zu den schulischen Physikvorkenntnissen der Biologiestudierenden an der RWTH-Aachen haben gezeigt, dass einerseits zwischen 60 % und 67 % der Studierenden nach der 10. Klasse keinen Physikunterricht mehr hatten. Andererseits haben knapp 20 % der Biologiestudierenden Physik bis zum Abitur belegt. Diese Ergebnisse lassen vermuten, dass die Studierenden breit gestreute, aber überwiegend mit der Zielgruppe des MeK-LSA Experimentiertests vergleichbare Vorerfahrungen mit physikalischen Experimenten besitzen.

76 % der 42 Biologiestudierenden waren weiblich. Mindestens 69 % der Studierenden haben das letzte Mal in der Sekundarstufe I am Physikunterricht teilgenommen. Die letzte Physiknote wird im Mittel als befriedigend angegeben (MW = 2,6; SD = 1,2), wobei nach Selbstauskunft im eigenen Physikunterricht eher selten Schülerexperimente durchgeführt worden sind (MW = 1,3; SD = 1,0; 0 ≙ gar nicht bis 5 ≙ sehr häufig). Diese personenbezogenen Daten deuten erwartungskonform auf heterogene schulische Vorkenntnisse in der untersuchten Stichprobe hin.

8.5 Studie E: Testformatvergleich (Anfertigen eines Messwertediagramms)

In Studie E wurde untersucht, ob Schülerinnen und Schüler beim Anfertigen eines Messwertediagramms im on-screen Format ähnliche Leistungen erzielen und eine ähnliche kognitive Belastung wahrnehmen wie beim Anfertigen eines inhaltlich identischen Messwertediagramms im hands-on Format. An Studie E haben elf Schülerinnen und acht Schüler einer 9. Gymnasialklasse aus Nordrhein-Westfalen teilgenommen. Die Teilnahme erfolgte im Klassenverband während der regulären Schulzeit. Die Schülerinnen und Schüler waren im Mittel 14,4 (SD = 0,6) Jahre alt. Studie E beschränkt sich auf eine Aufgabe zum Anfertigen eines Messwertediagramms, die sich dem Teilaufgabentyp *Datenauswertung durchführen* zuordnen lässt, und ist als (qualitative) Ergänzung zu Studie D zu sehen. Die Aufgabe zum Anfertigen eines Messwertediagramms (vgl. Abbildung 8.2 auf Seite 90) wurde sowohl im on-screen Format als auch in einem vergleichbar gestalteten hands-on Format umgesetzt (vgl. Abschnitt 7.3 auf Seite 81 zur Gestaltung des hands-on Formats). Vor der Bearbeitung der Aufgabe erhielten die Schülerinnen und Schüler, anhand einer Trainingsaufgabe zum Anfertigen eines Messwertediagramms, eine kurze Einführung in die Bedienung des Diagramm-Tools zur Erstellung von Messwertediagrammen.

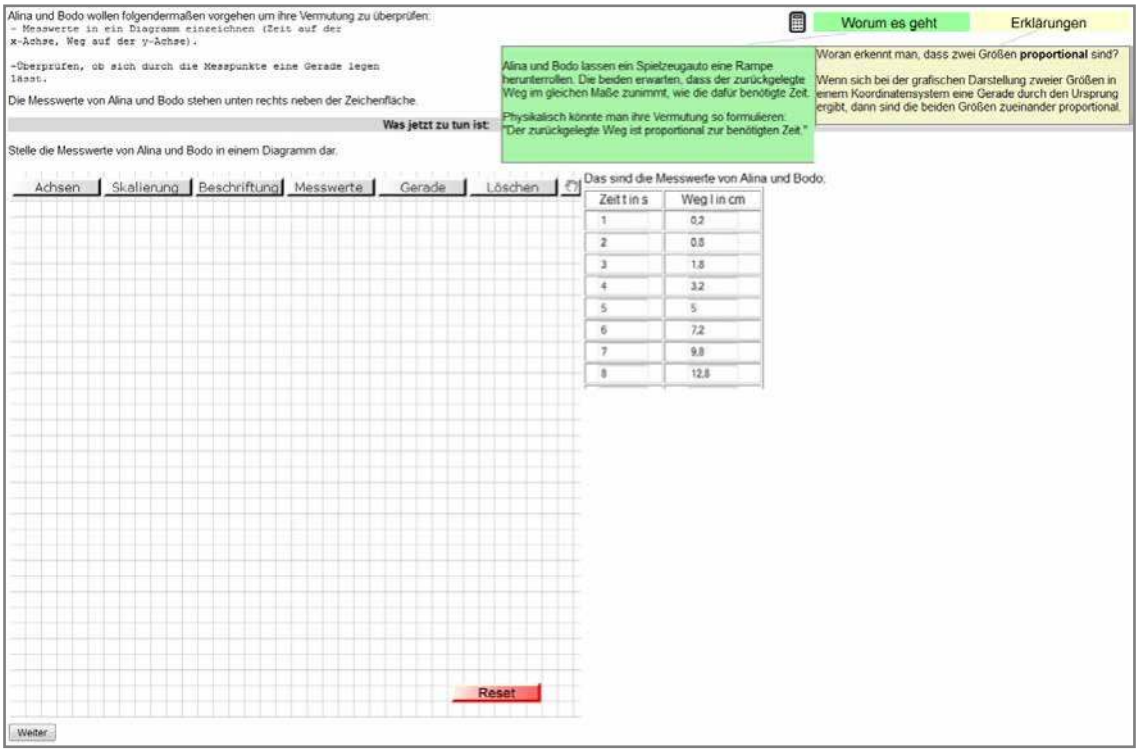


Abbildung 8.2: Aufgabenstellung zum Anfertigen eines Messwertediagramms in Studie E

Abbildung 8.3 zeigt den Ablauf der Studie. Die Schülerinnen und Schüler fertigten die Messwertediagramme im on-screen und hands-on Format an. Um einen möglichen Einfluss der Formatreihenfolge zu kontrollieren, wurde diese abgewechselt (vgl. Abbildung 8.3: Gruppe 1 & Gruppe 2). Die Bearbeitungszeit war für jedes Messwertediagramm auf sechs Minuten begrenzt. Zusätzlich zum Anfertigen der Messwertediagramme schätzten die Schülerinnen und Schüler in jedem Format die wahrgenommene kognitive Belastung ein (vgl. Abbildung 8.3). Die wahrgenommene kognitive Belastung wurde durch etablierte Selbsteinschätzungsskalen erfasst (vgl. Abschnitt 8.4.1 auf Seite 88: Zur Methode der Messung kognitiver Belastung). Die Bearbeitungen wurden durch Bildschirmscreenshots (on-screen) bzw. Protokollbögen (hands-on) dokumentiert.

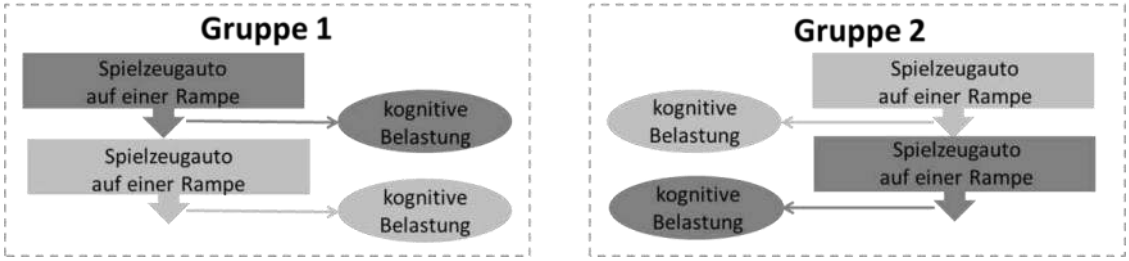


Abbildung 8.3: Ablauf von Studie E (dunkelgrau: hands-on Format; hellgrau: on-screen Format; Rechtecke: Erstellen eines Messwertediagramms zur Aufgabe Spielzeugauto auf einer Rampe; Ovale ≙ Einschätzung der wahrgenommenen kognitiven Belastung)

Prüfung der Annahmen aus Teil I des INA: Die Aufgaben umfassen relevante und repräsentative Inhalte und Anforderungen aus der Zieldomäne

Die zwei Annahmen zur ersten übergeordneten Aussage aus dem INA *Die Aufgaben umfassen relevante und repräsentative Inhalte und Anforderungen aus der Zieldomäne* werden in Kapitel 9 und Kapitel 10 evidenzbasiert geprüft und diskutiert.

9 Relevanz der Inhaltsbereiche (Annahme I.I)

In diesem Kapitel wird die folgende Annahme geprüft und diskutiert:

Die Aufgaben basieren auf Experimenten aus relevanten Inhaltsbereichen des Physikunterrichts der Sekundarstufe I (AI.I).

Zur Prüfung der Annahme wird auf die Ergebnisse der Lehrkräftebefragung zur Bekanntheit der Experimente aus den Aufgabenskizzen (vgl. Studie A in Abschnitt 8.1 auf Seite 83) und die Ergebnisse der Schülerbefragung zur Bekanntheit der Experimente aus den Testaufgaben (vgl. Studie B in Abschnitt 8.3 auf Seite 83) zurückgegriffen. Sind diese Experimente den Lehrkräften und den Schülerinnen und Schülern aus dem Physikunterricht der Sekundarstufe I bekannt, kann davon ausgegangen werden, dass die Experimente wahrscheinlich im Unterricht thematisiert worden sind und somit aus relevanten Inhaltsbereichen stammen. Aus diesen Ergebnissen kann jedoch nicht geschlossen werden, ob die normative Beschränkung auf drei Inhaltsbereiche (Elektrizitätslehre, Optik, Mechanik) zu einer Unterrepräsentation relevanter Inhaltsbereiche führt. In der Diskussion zu Annahme I.I wird daher auch Bezug auf die Ergebnisse einer von Karaböcek und Erb (2015) durchgeführten Lehrkräftebefragung zum *Einsatz von Experimenten im eigenen Physikunterricht* genommen, die Hinweise auf diesen Aspekt liefert.

9.1 Beitrag der Lehrkräftebefragung aus Studie A

In Studie A beantworteten Lehrkräfte zwei Fragen zur Bekanntheit der Experimente aus den Aufgabenskizzen (vgl. Abschnitt 8.1 auf Seite 83). Zu beiden Fragen wurde die mittlere Lehrkräfteeinschätzung für jedes Experiment bestimmt. Die Darstellung der Ergebnisse beschränkt sich auf die Experimente aus den Aufgabenskizzen, die in Teil I der Dissertation für den MeK-LSA Experimentiertest ausgewählt wurden[18]. Abbildung 9.1 auf Seite 92 zeigt die Ergebnisse zur Bekanntheit der Experimente. Die Experimente zu den Aufgaben *Totalreflexion* und *Fahrzeit auf der schiefen Ebene* sind nicht berücksichtigt, da diese Aufgaben erst nach Abschluss der Lehrkräftebefragung nachentwickelt wurden (Begründung: Abschnitt 4.7 auf Seite 51).

Die mittlere Lehrkräfteeinschätzung liegt für die Frage *wie wahrscheinlich Schülerinnen und Schüler dieses oder ein sehr ähnliches Experiment selbst durchgeführt haben* (Kurzform:

[18]Die in diesem Abschnitt gezeigten Ergebnisse zur Bekanntheit der Experimente aus den Aufgabenskizzen sind bei der Auswahl der Aufgaben in Teil I der Dissertation (vgl. Abschnitt 4.7) nicht berücksichtigt worden.

Durchgeführt?) bei vier von elf Experimenten bei einem Wert von mindestens 3,0 auf der vierstufigen Rating-Skala (1 ≙ sehr unwahrscheinlich bis 4 ≙ sehr wahrscheinlich) und bei fünf weiteren Experimenten bei mindestens 2,5. Für die Experimente *Ausdehnung eines Gummibandes* (MW = 2,3) und *Auftriebskraft im Wasser* (MW = 2,4) liegt die mittlere Lehrkräfteeinschätzung knapp unterhalb einer neutralen Einschätzung von 2,5. Die mittlere Lehrkräfteeinschätzung liegt für die Frage *wie wahrscheinlich Schülerinnen und Schüler dieses oder ein sehr ähnliches Experiment gesehen haben, ohne es selbst durchzuführen* (Kurzform: *Gesehen?*) bei sieben von elf Experimenten bei einem Wert von mindestens 3,0 auf der vierstufigen Rating-Skala und bei den vier weiteren Experimenten bei mindestens 2,5.

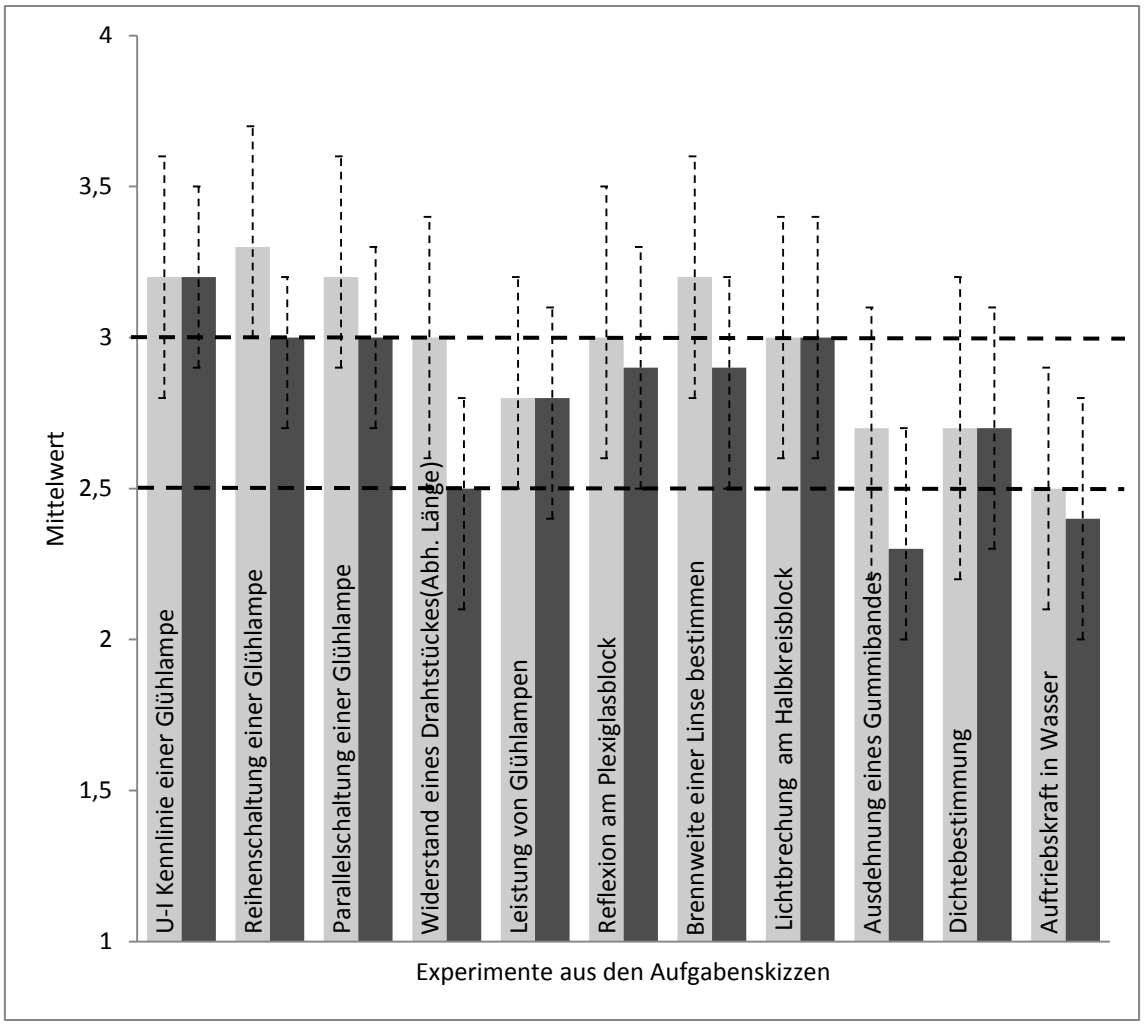


Abbildung 9.1: Lehrkräfteeinschätzung zur Bekanntheit der Experimente (hellgraue Balken: mittlere Einschätzung zur Frage Gesehen?*; dunkelgraue Balken: mittlere Einschätzung zur Frage* Durchgeführt?*; Aufgabennamen in hellgrauen Balken; senkrecht gestrichelte Linien ≙ 95 % Konfidenzintervall)*

Betrachtet man jeweils die untere und obere Grenze des 95 %-Konfidenzintervalls für die Lehrkräfteeinschätzungen zur Frage *Durchgeführt?* (vgl. Abbildung 9.1), so liegt die untere Grenze des Konfidenzintervalls für die Experimente *Widerstand eines Drahtstückes*, *Leistung von Glühlampen*, *Ausdehnung eines Gummibandes*, *Dichtebestimmung* und *Auftriebskraft in Wasser* unter einem Wert von 2,5. Die obere Grenze des Konfidenzintervalls liegt für die Experimente *Widerstand eines Drahtstückes*, *Ausdehnung eines Gummibandes* und *Auftriebskraft in Wasser* unter einem Wert von 3,0. Betrachtet man jeweils die untere und

obere Grenze des 95 %-Konfidenzintervalls für die Lehrkräfteeinschätzungen zur Frage *Gesehen?* (vgl. Abbildung 9.1 auf Seite 92), so liegt die untere Grenze für die Experimente aus dem Inhaltsbereich Mechanik (*Ausdehnung eines Gummibandes*, *Dichtebestimmung*, *Auftriebskraft in Wasser*) unter einem Wert von 2,5. Die obere Grenze des Konfidenzintervalls liegt nur für das Experiment *Auftriebskraft im Wasser* unter einem Wert von 3,0.

9.2 Beitrag der Schülerbefragung aus Studie B

In Studie B wurden Schülerinnen und Schüler nach der Bearbeitung der Testaufgaben zur Bekanntheit der Experimente aus den Testaufgaben befragt (vgl. Abschnitt 8.2 auf Seite 83). Nach jeder bearbeiteten Testaufgabe schätzten die Schülerinnen und Schüler ein, ob sie das durchgeführte Experiment aus dem Physikunterricht bereits kannten – und wenn ja, ob sie das Experiment bereits selbst durchgeführt hatten. Schülerinnen und Schüler, die bei der ersten Frage *Nein* und bei der zweiten Frage *Ja* angekreuzt haben, wurden aus der Analyse ausgeschlossen. Jede Testaufgabe wurde im Mittel von 374 (SD = 10) Schülerinnen und Schülern eingeschätzt. Die Schülerinnen und Schüler haben die Experimente aus den ausgearbeiteten Testaufgaben und nicht die Experimente aus den Aufgabenskizzen (vgl. Abschnitt 9.1) beurteilt. Abbildung 9.2 zeigt die Ergebnisse der Schülerbefragung zur Bekanntheit der Experimente.

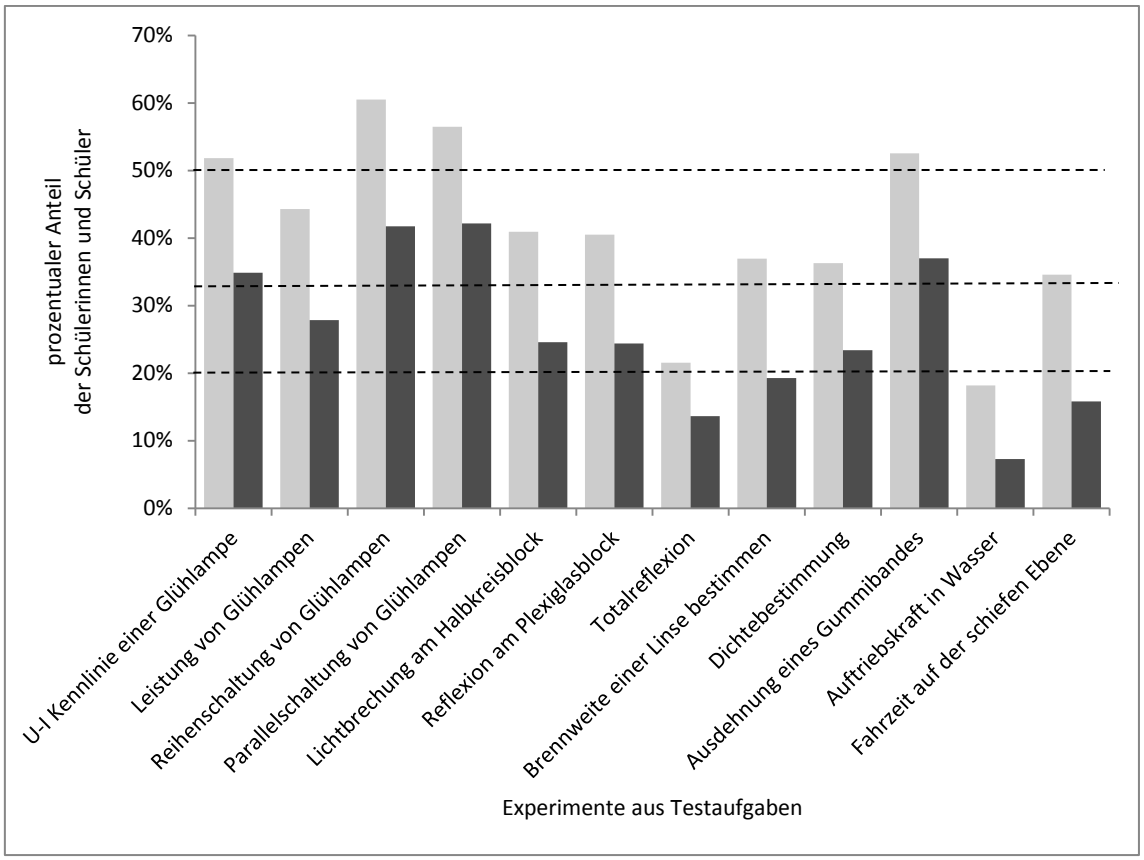


Abbildung 9.2: Ergebnisse der Schülerbefragung zur Bekanntheit der Experimente aus den Testaufgaben (hellgraue Balken: prozentualer Anteil der Schülerinnen und Schüler, die angegeben haben, das Experiment aus dem Physikunterricht zu kennen; dunkelgraue Balken: prozentualer Anteil der Schülerinnen und Schüler, die angegeben haben, das Experiment bereits im Physikunterricht selber durchgeführt zu haben)

Zehn von zwölf Experimenten sind mindestens einem Drittel der Schüler und Schülerinnen bekannt und davon wurden acht Experimente von mindestens 20 % der Schüler und Schülerinnen bereits selber durchgeführt. Deutlich nach unten weichen die Experimente *Totalreflexion* und *Auftriebskraft in Wasser* in der Bekanntheit ab. Diese sind ca. nur 20 % der Schülerinnen und Schüler bekannt. Deutlich nach oben weichen die Experimente aus dem Inhaltsbereich Elektrizitätslehre (*U-I Kennlinie einer Glühlampe, Leistung von Glühlampen, Reihenschaltung und Parallelschaltung von Glühlampen*) und das Experiment *Ausdehnung eines Gummibandes* in der Bekanntheit ab, die mindestens 50 % der Schülerinnen und Schüler bekannt sind.

9.3 Diskussion

Die Ergebnisse der Lehrkräftebefragung zur Bekanntheit der Experimente zeigen (vgl. Abschnitt 9.1), dass die für den MeK-LSA Experimentiertest ausgewählten Experimente mit hoher Wahrscheinlichkeit im Unterricht thematisiert worden sind. Dieser Befund ist erwartungskonform, da die Überführung der Testkonzeption in konkrete Testaufgaben u.a. auf umfangreichen Lehrplan- und Schulbuchanalysen basiert (vgl. Kapitel 4 auf den Seiten 37-53). Detailbetrachtungen zeigen, dass die Experimente aus dem Inhaltsbereich Mechanik wahrscheinlich am wenigsten bekannt sind, insbesondere wenn man die Einschätzungen zur Frage *Durchgeführt?* (vgl. Abbildung 9.1 auf Seite 92) berücksichtigt. Einschränkend bleibt festzuhalten, dass die Lehrkräfte ihre Einschätzungen bereits auf Ebene von Aufgabenskizzen und nicht auf der Ebene der Testaufgaben vorgenommen haben. Da sich die Inhaltsbereiche und Experimente zwischen Aufgabenskizzen und Testaufgaben nicht geändert haben, dürfte ein möglicher Einfluss auf die Einschätzung zur Bekanntheit der Experimente allerdings - wenn überhaupt vorhanden - sehr gering sein.

Die Einschätzungen der Schülerinnen und Schüler zur Bekanntheit der Experimente clustern in der Regel nach Inhaltsbereichen. Die Experimente aus dem Bereich der Elektrizitätslehre weisen den höchsten Bekanntheitsgrad auf. Danach folgen die Experimente aus dem Inhaltsbereich Optik (*Lichtbrechung am Halbkreisblock, Reflexion am Plexiglasblock, Brennweite einer Linse bestimmen),* wobei die nachentwickelte Aufgabe zur *Totalreflexion* im Bekanntheitsgrad deutlich nach unten abweicht. Die Einschätzungen zur Bekanntheit der Experimente aus dem Bereich der Mechanik zeigen ein uneinheitliches Bild. Während die Aufgabe *Ausdehnung eines Gummibandes* den Schülerinnen und Schülern ähnlich bekannt ist wie die Experimente aus dem Inhaltsbereich Elektrizitätslehre, ist die Bekanntheit der Experimente aus den Aufgaben *Dichtebestimmung* und *Fahrzeit auf der schiefen Eb*ene mit der Bekanntheit der Experimente aus dem Inhaltsbereich Optik vergleichbar. Das Experiment aus der Aufgabe *Auftriebskraft in Wasser* ist dagegen ähnlich unbekannt wie das Experiment zur Aufgabe *Totalreflexion*. Während sich die Experimente aus den Inhaltsbereichen Elektrizitätslehre und Optik auf *elektrische Stromkreise* bzw. *Optikexperimente auf dem Tisch* fokussieren, streuen die Experimente aus dem Inhaltsbereich Mechanik über unterschiedliche Teilbereiche der Mechanik (z. B. Mechanik der Flüssigkeiten). Aus diesem Grund ist das uneinheitliche Bild in der Mechanik plausibel erklärbar. Da die Schülerinnen und Schüler die Bekanntheit der Experimente einmalig am Ende der Aufgabenbearbeitung eingeschätzt haben

und ihnen keine weiteren Erläuterungen zur Verfügung standen, bleibt offen, worauf sich die Einschätzung der Schülerinnen und Schüler genau bezieht. Beispielsweise ist unklar, ob die Schülerinnen und Schüler sich bei ihrer Einschätzung auf das gesamte Experiment oder nur auf einzelne Teilschritte beziehen, oder wie ähnlich das im Unterricht gesehene bzw. durchgeführte Experiment gewesen sein muss, damit die Schülerinnen und Schüler sich an das Experiment erinnern. Aufgrund dieser Einschränkung kann nicht davon ausgegangen werden, dass Experimente, an die sich die Schülerinnen und Schüler nicht erinnern, im Unterricht nicht vorgekommen sind. Das erklärt möglicherweise die insgesamt relativ geringen prozentualen Anteile von Schülerinnen und Schülern, die angegeben haben, das Experiment im Unterricht gesehen oder durchgeführt zu haben.

Vergleicht man, trotz der oben beschriebenen Einschränkungen und unterschiedlichen Datenbasis der Studien A und B, die Einschätzungen der Schülerinnen und Schüler mit denen der Lehrkräfte, stellt man ähnliche Tendenzen bezogen auf die Bekanntheit der Experimente fest. Beispielsweise weisen sowohl aus Schüler- als auch aus Lehrkraftperspektive die Experimente aus dem Inhaltsbereich Elektrizitätslehre in der Regel den höchsten, und das Experiment zur *Auftriebskraft im Wasser* aus dem Inhaltsbereich Mechanik den geringsten Bekanntheitsgrad auf. Insgesamt kann auf Basis der Schüler- und Lehrkräftebefragung davon ausgegangen werden, dass die Experimente, die für den MeK-LSA Experimentiertest ausgewählt wurden, wahrscheinlich im Unterricht thematisiert worden sind. Die Inhaltsbereiche Elektrizitätslehre, Optik und Mechanik können folglich als relevant für die Zieldomäne (*Experimentieren im Physikunterricht der Sekundarstufe I*) bezeichnet werden. Es bleibt allerdings zu prüfen, ob die gewählte Beschränkung (vgl. Abschnitt 4.7 auf Seite 51) auf die Inhaltsbereiche Elektrizitätslehre, Optik und Mechanik zu einer Unterrepräsentation relevanter Inhaltsbereiche führt, da weitere Inhaltsbereiche (z. B. Wärmelehre) nicht berücksichtigt werden. Zur Prüfung wird Bezug auf die Ergebnisse einer von Karaböcek und Erb (2015) durchgeführten Lehrkräftebefragung zum *Einsatz von Experimenten im eigenen Physikunterricht* genommen. Die Ergebnisse dieser Befragung zeigen, dass Experimente zur Wärmelehre gegenüber Experimenten zur Elektrizitätslehre, Mechanik und Optik eine deutlich untergeordnete Rolle im Physikunterricht der Sekundarstufe I spielen (vgl. ebenda). Das Ausklammern des Inhaltsbereichs Wärmelehre aus dem MeK-LSA Experimentiertest schränkt demnach dessen Repräsentativität für die Inhaltsbereiche, in denen üblicherweise im Physikunterricht experimentiert wird, nicht bedeutsam ein.

10 Angemessenheit der Anforderungen (Annahme I.II)

In diesem Kapitel wird die folgende Annahme geprüft und diskutiert:

Die Aufgaben stellen Anforderungen, die Schülerinnen und Schüler aus ihrem Physikunterricht der Sekundarstufe I üblicherweise kennen (AI.II).

Zur Prüfung der Annahme wird zum einen auf die Ergebnisse der Lehrkräftebefragung zur Passung der Experimente aus den Aufgabenskizzen zu den experimentellen Anforderungen im Physikunterricht (vgl. Studie A in Abschnitt 8.1 auf Seite 83) zurückgegriffen. Zum anderen werden die Ergebnisse der Schülerbefragung zu experimentellen Anforderungen im Physikunterricht (vgl. Studie C in Abschnitt 8.3 auf Seite 83) berücksichtigt. Schätzen die Lehrkräfte die experimentellen Anforderungen als wahrscheinlich erfüllbar ein und geben die Schülerinnen und Schüler an, die gestellten experimentellen Anforderungen eher häufig in ihrem Physikunterricht erfüllen zu müssen, kann davon ausgegangen werden, dass es für diese experimentellen Anforderungen üblicherweise Lerngelegenheiten im Unterricht gibt.

10.1 Beitrag der Lehrkräftebefragung aus Studie A

Zur Passung der Experimente aus den Aufgabenskizzen zu den experimentellen Anforderungen im Physikunterricht wurden in Studie A Lehrkräfte befragt (vgl. Abschnitt 8.1 auf Seite 83). Die Lehrkräfte schätzten ein, wie wahrscheinlich Schülerinnen und Schüler das in der Aufgabenskizze beschriebene Experiment planen (Kurzform: *Planbarkeit?*), durchführen (Kurzform: *Durchführbarkeit?)* und auswerten (Kurzform: *Auswertbarkeit?*) können. Die Einschätzungen zu den drei Fragen sind Indikatoren für die Anforderungen, die üblicherweise im Physikunterricht gestellt werden, da bei einer wahrscheinlichen Erfüllbarkeit der Anforderungen davon ausgegangen werden kann, dass es für diese Anforderungen Lerngelegenheiten im Physikunterricht gibt. Zu den drei Fragen wurde die mittlere Lehrkräfteeinschätzung für jedes Experiment bestimmt. Die Einschätzungen zur Frage *Durchführbarkeit?* sind bereits in Teil I der Dissertation bei der Auswahl der Aufgaben für den MeK-LSA Experimentiertest berücksichtigt worden (vgl. Abschnitt 4.7 auf Seite 51). Im Folgenden geht es darum zu prüfen, ob die Einschätzungen zu dieser Frage – zumindest tendenziell – auch auf die Fragen *Planbarkeit?* und *Auswertbarkeit?* zutreffen.

Die Darstellung der Ergebnisse beschränkt sich auf die Experimente aus Aufgabenskizzen, die für den MeK-LSA Experimentiertest ausgewählt wurden. Abbildung 10.1 auf Seite 98 zeigt die Ergebnisse zur Erfüllbarkeit der experimentellen Anforderungen. Die Experimente zu den Aufgaben *Totalreflexion* und *Fahrzeit auf der schiefen Ebene* sind nicht berücksichtigt, da diese Aufgaben erst nach Abschluss der Lehrkräftebefragung nachentwickelt wurden (Begründung: Abschnitt 4.7 auf Seite 51).

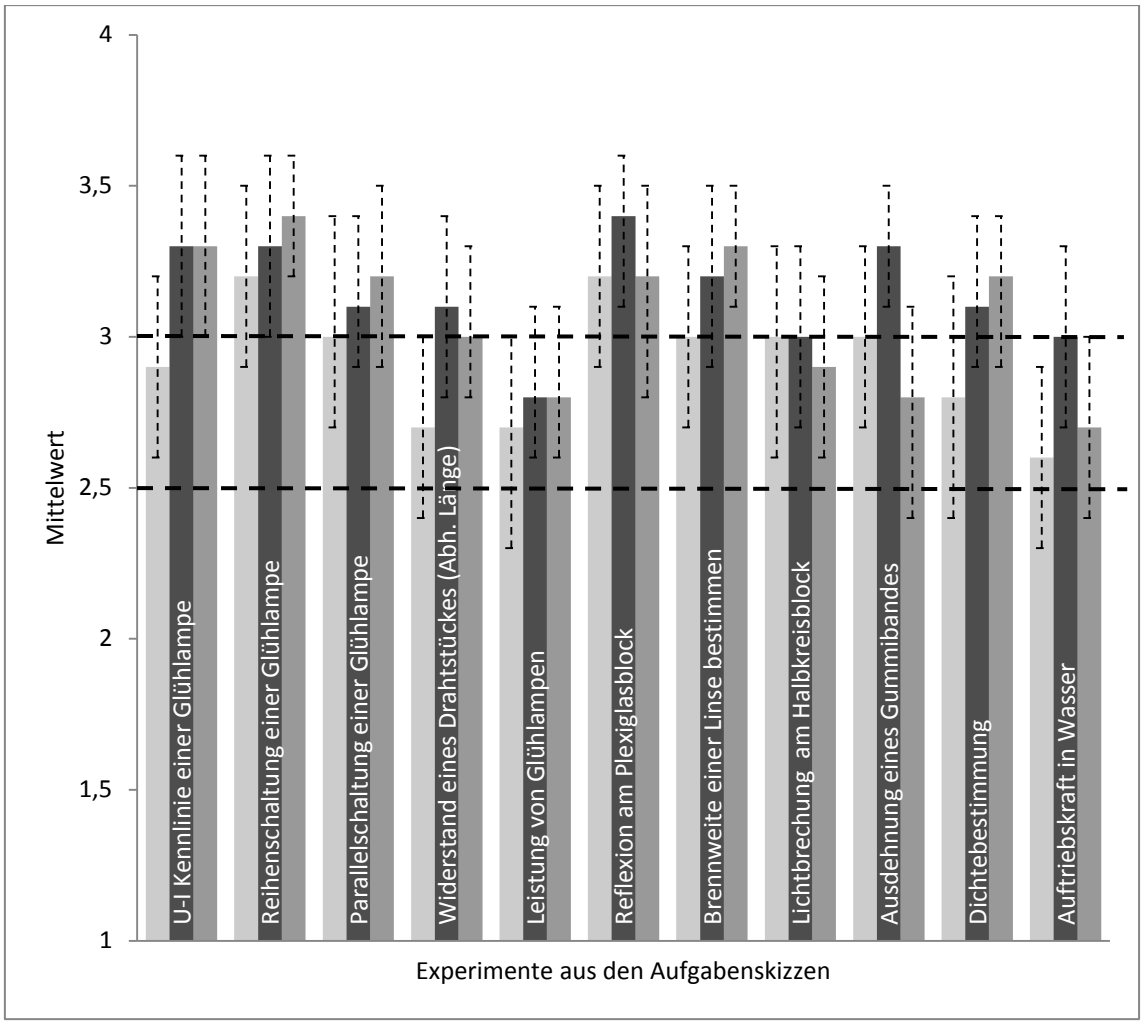


Abbildung 10.1: Lehrkräfteeinschätzung zur Erfüllbarkeit der experimentellen Anforderungen (hellgraue Balken: mittlere Einschätzung zur Frage Planbarkeit? *dunkelgraue Balken: mittlere Einschätzung zur Frage* Durchführbarkeit?*; mittellgraue Balken: mittlere Einschätzung zur Frage* Auswertbarkeit?*; senkrecht gestrichelte Linien ≙ 95 % Konfidenzintervalle)*

Die mittlere Lehrkräfteeinschätzung liegt für alle drei Fragen (*Planbarkeit?*, *Durchführbarkeit?*, *Auswertbarkeit?*) und alle elf Experimente über einem neutralen Wert von 2,5 auf der vierstufigen Rating-Skala (1 ≙ sehr unwahrscheinlich bis 4 ≙ sehr wahrscheinlich). Für die Frage *Planbarkeit?* liegen sechs von elf Experimenten bei einem Wert von mindestens 3,0. Für die Frage *Durchführbarkeit?* liegen erwartungskonform (aufgrund der Kriterien zur Aufgabenauswahl; Abschnitt 4.7 auf Seite 51) zehn von elf Experimenten bei einem Wert von mindestens 3,0. Für die Frage *Auswertbarkeit?* liegen sieben von elf Experimenten bei einem Wert von mindestens 3,0.

Betrachtet man jeweils die untere und obere Grenze des 95 %-Konfidenzintervalls für die Lehrkräfteeinschätzungen zur Frage *Planbarkeit?* (vgl. Abbildung 10.1), so liegt die untere Grenze des Konfidenzintervalls für die Experimente *Widerstand eines Drahtstückes*, *Leistung von Glühlampen*, *Dichtebestimmung* und *Auftriebskraft in Wasser* unter einem Wert von 2,5. Die obere Grenze des Konfidenzintervalls liegt nur für das Experiment *Auftriebskraft in Wasser* unter einem Wert von 3,0. Betrachtet man jeweils die untere und obere Grenze des 95 %-Konfidenzintervalls für die Lehrkräfteeinschätzungen zur Frage *Durchführbarkeit?*

(vgl. Abbildung 10.1 auf Seite 98), so liegt die untere Grenze des Konfidenzintervalls für alle elf Experimente über einem Wert von 2,5. Die obere Grenze des Konfidenzintervalls liegt für alle elf Experimente über einem Wert von 3,0. Betrachtet man jeweils die untere und obere Grenze des 95 %-Konfidenzintervalls für die Lehrkräfteeinschätzungen zur Frage *Auswertbarkeit?* (vgl. Abbildung 10.1 auf Seite 98), so liegt die untere Grenze des Konfidenzintervalls für die Experimente *Ausdehnung eines Gummibandes* und *Auftriebskraft in Wasser* unter einem Wert von 2,5. Die obere Grenze des Konfidenzintervalls liegt für alle Experimente mindestens bei einem Wert von 3,0.

10.2 Beitrag der Schülerbefragung aus Studie C

Im letzten Teil von Studie C (vgl. Abschnitt 8.3 auf Seite 83) wurden die Schülerinnen und Schüler zu experimentellen Anforderungen im Physikunterricht befragt. Dazu wurden die in den Teilaufgaben gestellten Anforderungen (z. B. ein Experiment selbstständig aufbauen) mit dem Schüler bzw. der Schülerin in der Testumgebung schrittweise durchgegangen. Zu jeder Anforderung (vgl. Tabelle 10.1) schätzten die Schülerinnen und Schüler auf einer vierstufigen Rating-Skala (1 ≙ selten bis 4 ≙ häufig) ein, wie häufig diese Anforderung im Physikunterricht gestellt wird.

Tabelle 10.1: Fragen an die Schülerinnen und Schüler zur Häufigkeit experimenteller Anforderungen im Physikunterricht und für die Auswertung zur Verfügung stehende Schülerantworten pro Frage

Fragen an die Schülerinnen und Schüler: Wie häufig musst du im Physikunterricht…	Antworten pro Frage
…selbstständig überlegen, welche Größen du messen musst?	36
….Geräte für ein Experiment auswählen?	36
…eine Skizze vor dem Experiment anfertigen?	36
…das Vorgehen für ein Experiment planen?	36
…ein Experiment selbstständig aufbauen?	36
…einen Protokollbogen vorbereiten?	36
…mehrere Messwerte aufnehmen?	36
…selbständig ein Auswertungsverfahren überlegen?	35
…Berechnungen vornehmen?	26
…ein Diagramm erstellen?	31
…eine Schlussfolgerung ziehen?	25
…eine Behauptung bestätigen oder widerlegen?	31

Für die Auswertung stehen 25 bis 36 Schülerantworten pro Frage zur Verfügung. Die im Vergleich zur Gesamtschülerzahl (106 Schülerinnen und Schüler in Studie C; vgl. Abschnitt 8.3 auf Seite 83) geringe Anzahl an Schülerantworten lässt sich im Wesentlichen auf zwei Gründe zurückführen. Zum einen wurde der Fragebogen zur Häufigkeit experimenteller Anforderungen im Physikunterricht erst nach der ersten Hälfte der Datenerhebungsphase eingesetzt[19]. Zum anderen wurde der Fragebogen im letzten Teil von Studie C vorgelegt, sodass der Fragebogen aus Zeitgründen bei einigen Schülerinnen und Schülern nur unvollständig oder gar nicht eingesetzt werden konnte.

[19] In der ersten Hälfte der Datenerhebungsphase wurden die Fragen zu den experimentellen Anforderungen nicht auf Rating-Skalen eingeschätzt.

Für jede Frage wurde die mittlere Einschätzung der Schülerinnen und Schüler bestimmt. Abbildung 10.2 zeigt die Ergebnisse der Schülerbefragung zur Häufigkeit der experimentellen Anforderungen im Unterricht.

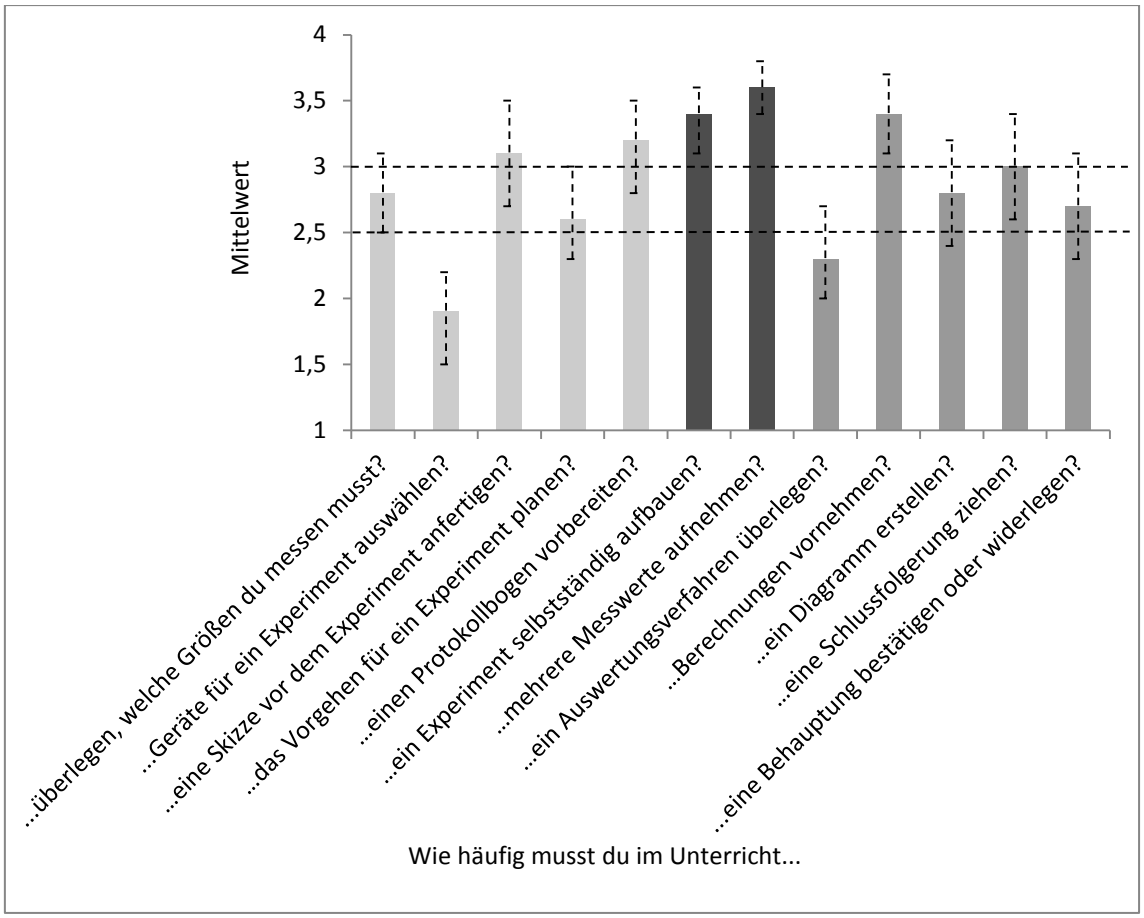


Abbildung 10.2: Einschätzungen der Schülerinnen und Schüler zur Häufigkeit experimenteller Anforderungen im Physikunterricht (Mittelwert und 95 % Konfidenzintervall; Grau-Abstufungen der Balken zeigen Bereiche des Experimentierens (von links nach rechts): Planung, Durchführung, Auswertung).

Für zehn der zwölf experimentellen Anforderungen liegt die mittlere Schülereinschätzung über einem neutralen Wert von 2,5 auf der vierstufigen Rating-Skala (1 ≙ selten bis 4 ≙ häufig). Die Anforderung *Geräte für ein Experiment auswählen (1,9)* liegt deutlich unter einem neutralen Wert von 2,5, während die Anforderung *Selbstständig ein Auswertungsverfahren überlegen* (2,4) nur knapp unter diesem Wert liegt. Mittelt man, im Sinne einer übergeordneten Anforderung *Versuchsplan entwerfen*, die Einschätzungen zu den Anforderungen *Geräte für ein Experiment auswählen*, *Skizze vor dem Experiment anfertigen* und *Vorgehen für ein Experiment planen*, ergibt sich eine mittlere Einschätzung von 2,6 auf der vierstufigen Rating-Skala. Sechs der zwölf experimentellen Anforderungen liegen sogar bei einem Wert von mindestens 3,0. Das Aufnehmen mehrerer Messwerte kommt nach Einschätzung der Schülerinnen und Schüler rein deskriptiv am häufigsten vor (3,6). Betrachtet man jeweils die untere und obere Grenze des 95 %-Konfidenzintervalls für die Schülereinschätzungen zur Häufigkeit der experimentellen Anforderungen (vgl. Abbildung 10.2), so liegt die untere Grenze des Konfidenzintervalls für die Anforderungen *Geräte für ein Experiment auswählen, das Vorgehen für ein Experiment planen,*

ein Auswertungsverfahren überlegen, ein Diagramm erstellen und *eine Behauptung bestätigen oder widerlegen* unter einem neutralen Wert von 2,5. Die obere Grenze des Konfidenzintervalls liegt nur für die Anforderungen *Geräte für ein Experiment auswählen* und *ein Auswertungsverfahren überlegen* unter einem Wert von 3,0.

10.3 Diskussion

Die Ergebnisse der Lehrkräftebefragung zur Erfüllbarkeit der experimentellen Anforderungen zeigen (vgl. Abschnitt 10.1), dass es für die im MeK-LSA Experimentiertest gestellten Anforderungen zur Planung, Durchführung und Auswertung von Experimenten mit hoher Wahrscheinlichkeit Lerngelegenheiten im Unterricht gibt. Einschränkend bleibt festzuhalten, dass die Lehrkräfte Ihre Einschätzungen bereits auf Ebene von Aufgabenskizzen und nicht auf der Ebene der Testaufgaben vorgenommen haben. Den Lehrkräften standen somit während der Einschätzung der Experimente aus den Aufgabenskizzen die genaue Formulierung der Aufgabenstellung auf Ebene der Teilaufgaben, die in den Teilaufgaben vorgegebenen Zwischenlösungen und der Bewertungsmaßstab zur Bewertung der Aufgabenbearbeitungen nicht zur Verfügung. Da insbesondere die von Alina und Bodo eingebrachten Zwischenlösungen die Schülerinnen und Schüler bei der Bearbeitung einer Aufgabe unterstützen, kann davon ausgegangen werden, dass die Einschätzungen der Lehrkräfte bezüglich der Lösungswahrscheinlichkeit eher noch positiver ausgefallen wären.

Die Einschätzungen der Schülerinnen und Schüler zur Häufigkeit experimenteller Anforderungen im Unterricht zeigen (vgl. Abschnitt 10.2), dass die im MeK-LSA Experimentiertest gestellten Anforderungen eher häufig im Unterricht vorkommen. Das gilt insbesondere für die Anforderungen im Bereich der Durchführung (*ein Experiment selbstständig aufbauen* und *mehrere Messwerte aufnehmen*). Dieser Befund unterstreicht, dass experimentelle Anforderungen im Bereich der Durchführung im Zentrum physikalischen Experimentierens der Sekundarstufe I stehen. Die relativ geringe Häufigkeit der Anforderungen *Geräte für ein Experiment auswählen* und *selbstständig ein Auswertungsverfahren überlegen* deutet darauf hin, dass im Unterricht in der Regel sehr stark vorgeplante und gelenkte Experimentieraufgaben eingesetzt werden. Diese Experimentieraufgaben berücksichtigen nur sehr selten eigene Planungs- und Auswertungsideen der Schülerinnen und Schüler, insbesondere solche, die sich nur bedingt vorab planen lassen. Ähnliche Befunde finden sich auch in der IPN-Videostudie (vgl. Tesch & Duit, 2004). Da zu stark vorgeplante und gelenkte Experimentieraufgaben weder einem fachdidaktischen Ideal von Experimentieren, noch den Kompetenzerwartungen zum Experimentieren in den Bildungsstandards in vollem Umfang entsprechen, werden im MeK-LSA Experimentiertest bewusst auch Anforderungen gestellt (z. B. *Geräte auswählen* oder *Datenauswertung planen*), die eigene Planungs- und Auswertungsideen der Schülerinnen und Schüler erfordern.

Prüfung der Annahmen aus Teil II des INA: Die beobachtete Performanz passt zur beabsichtigten Performanz

Die vier Annahmen zur zweiten übergeordneten Aussage aus dem INA *Die beobachtete Performanz passt zur beabsichtigten Performanz* werden in Kapitel 11 bis 14 evidenzbasiert geprüft und diskutiert.

11 Experimentbezogene Überlegungen (Annahme II.I)

In diesem Kapitel wird die folgende Annahme geprüft und diskutiert:

Die Schülerinnen und Schüler stellen bei der Bearbeitung der on-screen Aufgaben überwiegend experimentbezogene Überlegungen an (AII.I).

Zur Prüfung der Annahme wird auf Daten aus Studie C (vgl. Abschnitt 8.3 auf Seite 83) zurückgegriffen. Die Annahme wird explorativ geprüft, da für den MeK-LSA Experimentiertest bislang keine abgesicherten Theorien zu den ablaufenden Überlegungen (kognitiven Prozessen) während der Aufgabenbearbeitung vorliegen.

11.1 Beitrag des begleitenden Think-Aloud aus Studie C

Um zu untersuchen, welche Art von Überlegungen (z. B. experimentbezogene Überlegungen) die Schülerinnen und Schüler bei der Bearbeitung der Aufgaben anstellen, wurden sie dazu aufgefordert, während der Aufgabenbearbeitung alle Überlegungen laut auszusprechen (vgl. begleitendes Think-Aloud in Studie C). Im Folgenden werden das Kategoriensystem zur Kategorisierung der Daten (Abschnitt 11.1.1) und die zur Prüfung von Annahme II.I berücksichtigten Daten beschrieben (Abschnitt 11.1.2). Im Anschluss werden die Ergebnisse zur Prüfung von Annahme II.I dargestellt und diskutiert.

11.1.1 Kategorisierung der Daten

Die mittels Bildschirmaufzeichnungen dokumentierten Bearbeitungen von Teilaufgaben wurden zusammen mit den Audiomitschnitten zeitbasiert in 10 Sekunden-Intervallen kategorisiert. Die Intervalllänge von 10 Sekunden entspricht dabei *„auch der in anderen Arbeiten gefundenen ca. mittleren Dauer eines inhaltlich zusammenhängenden Gedankenganges“* (von Aufschnaiter & Rogge, 2010, S. 100). Bearbeitungen von Teilaufgaben werden im Folgenden auch als Datensätze bezeichnet. Zur zeitbasierten Kategorisierung der Datensätze hat Zirwes (2014) im Rahmen ihrer Abschlussarbeit (vgl. Abschnitt 7.1 auf Seite 77) ein Kategoriensystem mit fünf Oberkategorien ausgearbeitet und erprobt (Tabelle 11.1).

Tabelle 11.1: Kategorien zur Kategorisierung der Datensätze

Kategorien				
physikalisch-experimentell	reproduktiv	Computerbedienung	Sonstiges	keine Verbalisierung

Der Kategorie *physikalisch-experimentell* werden die eindeutig experimentbezogenen Überlegungen der Schülerinnen und Schüler zugeordnet (z. B. *Die Zeit muss ich hier nicht messen, daher brauche ich keine Stoppuhr*). Eine Bedrohung für die kognitiv-valide Erfassung von Experimentierfähigkeiten besteht in Anlehnung an Messick (1996, S. 5) darin, dass zu einem substantiellen Anteil konstrukt-irrelevante Überlegungen auf Seiten der Schülerinnen und Schüler initiiert werden. In diesem Zusammenhang identifiziert Gut (2012, S. 63) bei der Analyse des *HarmoS-Experimentiertests* zum Beispiel kompetenzirrelevante Aufgabenmerkmale, die sich auf das Erfassen der Aufgabe und das Geben der Antwort beziehen. Überlegungen dieser Art werden durch die Kategorie *reproduktiv* (Lesen der Aufgabenstellung und Mitsprechen der Antwort bei der Eingabe) erfasst. Die Aufgaben werden vollständig on-screen bearbeitet, daher können auch formatspezifische Überlegungen (z. B. zur Computerbedienung) eine Rolle spielen. Formatspezifische Überlegungen werden in der Kategorie *Computerbedienung* erfasst. Alle nicht zuordenbaren Überlegungen werden in die Kategorie *Sonstiges* übernommen. Für die Güte des Kategorisierungsverfahrens ergeben sich für die Datensätze zufriedenstellende Werte ($.59 < \kappa < .91$; mittleres $\kappa = .73$). Insgesamt sind 5 % aller 10 Sekunden-Intervalle (772 von 15896) aus den in der Analyse berücksichtigten Datensätzen doppelt kodiert worden (vgl. berücksichtigte Daten in Abschnitt 11.1.2).

11.1.2 Berücksichtigte Daten

Die Datenerhebung mit begleitendem Think-Aloud wurde für die vier on-screen Aufgaben *Ausdehnung eines Gummibandes*, *Brechung am Halbkreisblock*, *Leistung von Glühlampen* und *Reihenschaltung von Glühlampen* durchgeführt. Die Kategorisierung der erhobenen Daten mit dem in Abschnitt 11.1.1 beschriebenen Kategoriensystem beschränkt sich auf die Teilaufgabentypen *Versuchsplan entwerfen*, *Versuch aufbauen und testen*, *Messung durchführen und dokumentieren* sowie *Datenauswertung durchführen*. Die Auswahl der Aufgaben und die Auswahl der in der Kategorisierung berücksichtigten Teilaufgabentypen wurden in Abschnitt 7.2 auf Seite 80 begründet.

Insgesamt konnten 644 Datensätze mit dem Kategoriensystem aus Abschnitt 11.1.1 kategorisiert werden. In die Analyse zur Prüfung von Annahme II.I gehen nur Datensätze ein, bei denen die Schülerinnen und Schüler ihre Überlegungen in mindestens 50 % der 10 Sekunden-Intervalle verbalisiert haben. Bei einem niedrigeren Anteil wird davon ausgegangen, dass die Schülerinnen und Schüler ihre Überlegungen zu einem zu geringen Anteil verbalisiert haben, um daraus valide Rückschlüsse auf die kognitiven Prozesse ziehen zu können. Aufgrund dieses Kriteriums mussten jedoch nur 12 % aller erhobenen Datensätze (77 von 644) aus den weiteren Analysen ausgeschlossen werden. Das zeigt, dass der Großteil der Schülerinnen und Schüler in der untersuchten Stichprobe in der Lage war, eigene Überlegungen in ausreichendem Maße zu verbalisieren. Für die Analyse zur Prüfung von Annahme II.I stehen somit 567 Datensätze zur Verfügung. Tabelle 11.2 auf Seite 105 zeigt die Anzahl der Datensätze getrennt nach Aufgaben und Teilaufgabentypen.

Tabelle 11.2: Anzahl in der Analyse zur Prüfung von Annahme II.I berücksichtigter Datensätze getrennt nach Aufgaben und Teilaufgabentypen

Teilaufgabentyp	**Aufgaben**				**Summe**
	Ausdehnung eines Gummibands	Brechung am Halbkreisblock	Leistung von Glühlampen	Reihenschaltung von Glühlampen	
Versuchsplan entwerfen	38	38	37	34	147
Versuch aufbauen und testen	36	36	38	34	144
Messung durchführen und dokumentieren	36	38	34	34	142
Datenauswertung durchführen	35	37	31	31	134
Summe	145	149	140	133	567

11.1.3 Datenanalyse

Die in der Analyse zur Prüfung von Annahme II.I berücksichtigten Datensätze (vgl. Tabelle 11.2) wurden mit dem in Abschnitt 11.1.1 beschriebenen Kategoriensystem zeitbasiert (10 Sekunden-Intervalle) kategorisiert. Jedem Intervall wird genau eine Kategorie (z. B. *physikalisch-experimentell*) zugewiesen (vgl. Kodiermanual zur Kategorisierung der Daten aus dem begleitenden Think-Aloud in Anhang A.5 auf Seite 187). Wenn in einem Intervall Überlegungen auftreten, die man verschiedenen Kategorien zuordnen kann, dann wird die Kategorie zugewiesen, die den größeren zeitlichen Anteil in diesem Intervall einnimmt. Eine Ausnahme bildet die Kategorie *keine Äußerung*. Diese Kategorie wird nur zugewiesen, wenn in einem kompletten Intervall keine Schüleräußerung vorliegt. Das heißt, äußert der Schüler bzw. die Schülerin in zwei Sekunden des Intervalls physikalische-experimentelle Überlegungen und in den verbleibenden acht Sekunden keine weiteren Überlegungen, wird dem 10 Sekunden-Intervall die Kategorie *physikalisch-experimentell* zugewiesen, obwohl die Kategorie *keine Äußerung* den größten zeitlichen Anteil in diesem Intervall einnimmt. Die Häufigkeitsverteilung der Kategorien kann folglich nicht unmittelbar mit der zeitlichen Häufigkeitsverteilung der Kategorien gleichgesetzt werden, stellt aber zumindest einen Indikator für die zeitliche Häufigkeitsverteilung dar (vgl. Rogge, 2010, S. 178). Die Gesamtzahl der pro Datensatz kategorisierten Zeitintervalle hängt darüber hinaus von der individuellen Bearbeitungsdauer der Teilaufgabe ab, da für die Bearbeitung von Teilaufgaben keine feste Bearbeitungszeit vorgegeben war. Zur Ergebnisdarstellung werden in Anlehnung an Rogge (2010, S. 177-178) mittlere prozentuale Anteile der Kategorien betrachtet, relativ zur Gesamtzahl der pro Datensatz kodierten Zeitintervalle:

$$\left(\frac{1}{n}\sum_{i=1}^{n}\frac{AZ_{i,K}}{AZ_i}\right)\cdot 100\ \%$$

mit $AZ_{i,K}$ ≙ Anzahl der in Datensatz i mit Kategorie K kodierten Zeitintervalle

AZ_i ≙ Anzahl der in Datensatz i kodierten Zeitintervalle und n ≙ Anzahl der Datensätze

11.1.4 Ergebnisse

Tabelle 11.3 auf Seite 106 zeigt die mittleren prozentualen Anteile der Kategorien getrennt nach Teilaufgabentypen und über alle Teilaufgabentypen hinweg (*Gesamt*). Insgesamt zeigt sich, dass Schülerinnen und Schüler über alle Teilaufgabentypen hinweg im Mittel in 60 % der

Zeitintervalle physikalisch-experimentelle Überlegungen anstellen. Je nach Teilaufgabentyp liegt der mittlere prozentuale Anteil zwischen 50 % und 70 %. Überlegungen zur Computerbedienung spielen beim Teilaufgabentyp *Versuch aufbauen und testen* mit im Mittel 9 % im Vergleich zu den anderen Teilaufgabentypen (2 % bis 3 %) die größte Rolle. Jedoch ist der mittlere prozentuale Anteil über alle Teilaufgabentypen hinweg als sehr gering zu bezeichnen (4 %).

Tabelle 11.3: Mittlere prozentuale Anteile der Kategorien (angegeben in der Form „mittlerer Anteil (Standardabweichung) in %"; Gesamt ≙ über alle Teilaufgabentypen hinweg)

Teilaufgabentyp	**Kategorien**				
	physikalisch-experimentell	reproduktiv	Computerbedienung	Sonstiges	keine Verbalisierung
Versuchsplan entwerfen	50 (16)	24 (13)	2 (3)	11 (10)	13 (14)
Versuch aufbauen und testen	55 (21)	8 (9)	9 (11)	15 (14)	13 (13)
Messung durchführen und dokumentieren	70 (18)	6 (6)	3 (4)	11 (10)	10 (13)
Datenauswertung durchführen	65 (20)	11 (13)	3 (5)	10 (12)	11 (14)
Gesamt	60 (20)	12 (13)	4 (7)	12 (12)	12 (13)

Der mittlere prozentuale Anteil reproduktiver Überlegungen liegt beim Teilaufgabentyp *Versuchsplan entwerfen* mit 24 % deutlich über den mittleren prozentualen Anteilen reproduktiver Überlegungen bei den anderen Teilaufgabentypen (6 % bis 11 %).

11.2 Diskussion

Die oben dargestellten Ergebnisse zeigen, dass experimentbezogene Überlegungen während der Bearbeitung aller Teilaufgabentypen die größte Rolle spielen. Gleichzeitig sind Überlegungen zur Computerbedienung den experimentbezogenen Überlegungen in der Regel am deutlichsten untergeordnet. Das spricht dafür, dass das on-screen Format keine zusätzlichen Denkkapazitäten bindet, solange keine Bedienprobleme vorliegen (vgl. Abschnitt 13.2 auf Seite 123). Überlegungen der Kategorie *reproduktiv* nehmen nur beim Teilaufgabentyp *Versuchsplan entwerfen* einen vergleichsweise hohen Stellenwert ein. Die erfolgreiche Bearbeitung dieses Teilaufgabentyps erfordert in hohem Maße eigenständige Planungsideen der Schülerinnen und Schüler, welche ein grundlegendes Verständnis der übergeordneten Aufgabenstellung voraussetzen. Es ist daher plausibel, dass die Schülerinnen und Schüler bei diesem Teilaufgabentyp vergleichsweise viel Zeit darauf verwenden, einen Bezug zwischen der übergeordneten Aufgabenstellung und der vorgegebenen Grundidee herzustellen, um auf diese Weise die eigenen Planungsideen zu konkretisieren. Detailbetrachtungen zeigen allerdings, dass dieser Erklärungsansatz nur teilweise trägt, da beim Teilaufgabentyp *Versuchsplan entwerfen* auch eine Konfundierung zwischen der

Methode begleitendes Think-Aloud*,* dem Kategoriensystem zur Auswertung der Daten und den Anforderungen des Teilaufgabentyps beobachtet werden kann, insbesondere bei den Teilschritten *Skizze anfertigen* und *Vorgehensweise beschreiben*. Werden beispielsweise Schülerüberlegungen verbal formuliert und dann beim Eintippen der Lösung oder beim Zeichnen der Skizze nochmals laut mitgesprochen, ohne dass eine neue Überlegung geäußert wird, so wurden diese wiederholten Überlegungen der Kategorie *reproduktiv* zugeordnet. Der vergleichsweise hohe Anteil reproduktiver Überlegungen beim Teilaufgabentyp *Versuchsplan entwerfen* ist daher zusätzlich zum oben beschriebenen Erklärungsansatz auch auf ein methodisches Problem zurückzuführen, jedoch nicht auf ein grundsätzliches Problem des Aufgabenformats. Ganz im Gegenteil: Bedenkt man, dass das Erfassen der Aufgabenstellung zwar kein unmittelbar konstrukt-relevanter, aber ein notwendiger Bestandteil jeder Aufgabenbearbeitung ist, so kann der mittlere prozentuale Anteil reproduktiver Überlegungen über alle Teilaufgabentypen hinweg (12 %) sogar als gering bezeichnet werden. Das spricht dafür, dass die für den MeK-LSA Experimentiertest getroffenen Strukturierungsmaßnahmen dazu beitragen, die Aufgabenstellung schnell erfassen zu können, sodass ausreichend Zeit für konstrukt-relevante Überlegungen während der Bearbeitung zur Verfügung steht.

Als Einschränkungen müssen festgehalten werden, dass die zeitbasierte Kodierung nur eine Annäherung an die zeitliche Verteilung der Überlegungen darstellt (vgl. Abschnitt 11.1.3), und dass mit dem Verzicht auf eine Transkription der Daten ein Informationsverlust einhergeht, der qualitative Detailanalysen erschwert. Für die gewählte Vorgehensweise spricht, dass zum einen die Intervalllänge (10 Sekunden) zumindest in der Größenordnung eines mittleren Gedankengangs liegt (vgl. Abschnitt 11.1.1). Zum anderen erfolgt die Kategorisierung der Intervalle auf Basis eines inhaltlich sinnvoll zu interpretierenden Kategoriensystems, sodass insgesamt von einer zufriedenstellenden Beurteilung der Gedankengänge auszugehen ist. Gleichzeitig können mit der gewählten Vorgehensweise (keine zeitaufwendige Transkription der Daten erforderlich) vergleichsweise große Fallzahlen untersucht werden, sodass Aussagen auf einer relativ breiten empirischen Basis getroffen werden können. Bedenkt man darüber hinaus, dass routinierte Handlungen automatisiert ablaufen und daher nicht verbalisiert werden können (vgl. Sandmann, 2014, S. 188), stellt der gefundene hohe Anteil experimentbezogener Überlegungen (gesamt: 60 %) vermutlich eher eine Unterschätzung des wahren Anteils konstrukt-relevanter Überlegungen dar.

12 Demonstration experimenteller Performanz (Annahme II.II)

In diesem Kapitel wird die folgende Annahme geprüft und diskutiert:

Die Bearbeitung des Tests ermöglicht das Zeigen experimenteller Performanz (AII.II).

Der MeK-LSA Experimentiertest basiert auf vorstrukturierten, unabhängig voneinander zu bearbeitenden on-screen (Teil-)Aufgaben mit interaktiven Simulationen (vgl. Abschnitt 3.1 auf Seite 29). Es kann zwar angenommen werden, dass die Schülerinnen und Schüler die im Test gestellten experimentellen Anforderungen wahrscheinlich aus dem Physikunterricht kennen (vgl. Kapitel 10 auf den Seiten 97-101), das Aufgabenformat wird den Schülerinnen und Schülern jedoch nicht bekannt sein. Um das Aufgabenformat kennenzulernen, bearbeiten die Schülerinnen und Schüler daher zu Beginn des Tests eine Trainingsaufgabe. Die Konzeption des Tests (v.a. konsekutive Struktur mit Zwischenlösungen) und die Trainingsaufgabe (Kennenlernen des Aufgabenformats) sollen sicherstellen, dass Schülerinnen und Schüler die Möglichkeit haben, beim Bearbeiten des Tests experimentelle Performanz zu zeigen (vgl. Annahme II.II). Bedrohungen für das Beibehalten der Annahme bestehen darin, dass die getroffenen Maßnahmen von den Schülerinnen und Schülern nicht in der intendierten Weise wahrgenommen werden oder die Schülerinnen und Schüler sogar irritieren und in der Folge das Zeigen experimenteller Performanz nicht möglich ist. Zur Prüfung der Annahme wird daher zum einen auf die Ergebnisse der Schülerbefragung zur Wahrnehmung der Testsituation (vgl. Studie C in Abschnitt 8.3 auf Seite 83) zurückgegriffen. Zum anderen werden die Studien F und G (vgl. Tabelle 7.1 auf Seite 79) berücksichtigt, in denen die oben genannten Aspekte empirisch untersucht wurden.

12.1 Beitrag der Schülerbefragung aus Studie C

Die Schülerinnen und Schüler wurden in Studie C nach der Bearbeitung der on-screen Aufgaben *Leistung von Glühlampen* und *Brechung am Halbkreisblock* zur Wahrnehmung der Testsituation befragt. Hierzu schätzten sie die fünf in Tabelle 12.1 auf Seite 110 dargestellten Fragen auf einer vierstufigen Rating-Skala ein (++ ≙ positive Einschätzung; + ≙ eher positive Einschätzung; - ≙ eher negative Einschätzung; -- ≙ negative Einschätzung). Die erste Frage zielt auf die Wahrnehmung des konsekutiven Aufgabenformats ab, Fragen zwei und drei zielen auf die Wahrnehmung der Testbedienung ab und die letzten beiden Fragen zielen auf die Wahrnehmung der Rahmenhandlung von Alina und Bodo ab, die dazu beitragen soll, die Zwischenlösungen authentisch einzubinden. Die Antworten auf diese Fragen liefern Hinweise, ob das Bearbeiten des Tests das Zeigen experimenteller Performanz ermöglicht.

Tabelle 12.1: Fragen an Schülerinnen und Schüler zur Wahrnehmung der Testsituation

Fragen an die Schülerinnen und Schüler				**Leistung von Glühlampen**	**Brechung am Halbkreisblock**
Frage1: Wie hast du diese „Stationenarbeit“ am Computer empfunden?				34	36
angenehm	eher angenehm	eher unangenehm	unangenehm		
☐	☐	☐	☐		
++	+	-	--		
Frage 2:Wie bist du mit der Bedienung der Simulationen zurecht gekommen?				34	35
gut	eher gut	eher schlecht	schlecht		
☐	☐	☐	☐		
++	+	-	--		
Frage 3: Du musstest dich durch den Test hindurchklicken: Wie gut bist du insgesamt mit der Bedienung des Tests zurechtgekommen?				34	36
gut	eher gut	eher schlecht	schlecht		
☐	☐	☐	☐		
++	+	-	--		
Frage 4: Zwischendurch wurden immer wieder Lösungen von Alina und Bodo gezeigt. Wie hilfreich fandest du die Lösungen insgesamt?				33	36
hilfreich	eher hilfreich	eher nicht hilfreich	nicht hilfreich		
☐	☐	☐	☐		
++	+	-	--		
Frage 5: Es gab sicherlich auch Stellen, wo du selbst eine Lösung gefunden hast. Wie fandest du es dann, dass du im nächsten Schritt mit der Lösung von Alina und Bodo weiterarbeiten solltest?				34	33
hilfreich	eher hilfreich	eher störend	störend		
☐	☐	☐	☐		
++	+	-	--		

Für die Auswertung stehen 33 bis 34 Schülerantworten pro Frage für die Aufgabe *Leistung von Glühlampen* und 33 bis 36 Schülerantworten pro Frage für die Aufgabe *Brechung am Halbkreisblock* zur Verfügung (vgl. Tabelle 12.1). Die im Vergleich zur Gesamtschülerzahl (106 Schülerinnen und Schüler in Studie C; vgl. Abschnitt 8.3 auf Seite 83) geringe Anzahl an Schülerantworten lässt sich im Wesentlichen darauf zurückführen, dass der Fragebogen zur Wahrnehmung der Testsituation erst nach der ersten Hälfte der Datenerhebungsphase eingesetzt wurde (vgl. Abschnitt 10.2 auf Seite 99). Abbildung 12.1 auf Seite 111 zeigt die Ergebnisse der Schülerbefragung zur Wahrnehmung der Testsituation für die Aufgaben *Leistung von Glühlampen* und *Brechung am Halbkreisblock*. Für jede Aufgabe und jede Frage ist die prozentuale Verteilung der Schülerantworten dargestellt.

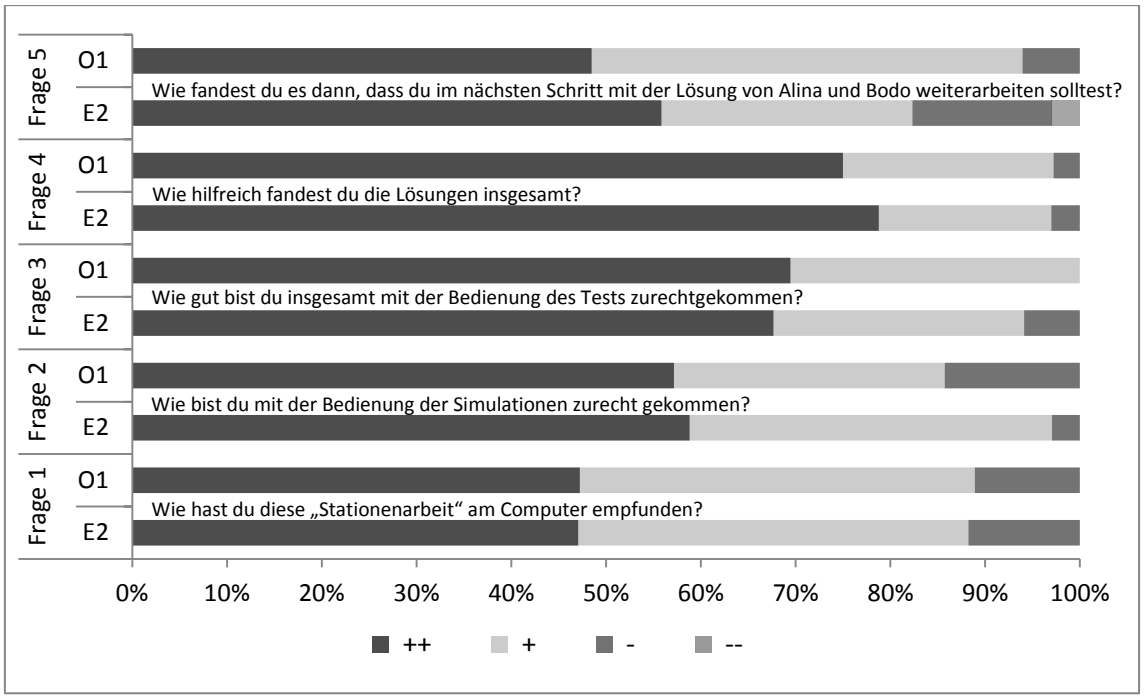


Abbildung 12.1: Ergebnisse der Schülerbefragung zur Wahrnehmung der Testsituation (O1 ≙ Brechung am Halbkreisblock; E2 ≙ Leistung von Glühlampen; ++ ≙ positive Einschätzung, + ≙ eher positive Einschätzung, -≙ eher negative Einschätzung und --≙ negative Einschätzung)

Die Schülerinnen und Schüler schätzten die Fragen zur Wahrnehmung der Testsituation über beide Aufgaben hinweg mehrheitlich positiv bzw. eher positiv ein (Minimum: 82 % bei Frage 5 zur Aufgabe *Leistung von Glühlampen*; Maximum: 100 % bei Frage 3 zur Aufgabe *Brechung am Halbkreisblock*). Das konsekutive Aufgabenformat (Frage 1), die Bedienung des Tests (Fragen 2 und 3) und die Zwischenlösungen von Alina und Bodo (Fragen 4 und 5) wurden mehrheitlich positiv wahrgenommen. Nur wenige Schülerinnen und Schüler (*Leistung von Glühlampen*: 6; *Brechung am Halbkreisblock*: 2) fanden es störend, dass sie nach eigener Lösung einer Teilaufgabe mit der Zwischenlösung von Alina und Bodo weiterarbeiten sollten (Frage 5).

12.2 Evaluation der Trainingsaufgabe (Beitrag aus Studie F)

Um das Aufgabenformat kennenzulernen und den Umgang mit dem Aufgabenformat zu üben, bearbeiten die Schülerinnen und Schüler zu Beginn des Tests eine Trainingsaufgabe, die alle acht im Aufgabenentwicklungsmodell (vgl. Abbildung 3.1 auf Seite 30) enthaltenen Teilaufgabentypen umfasst. Die Aufgabe *Widerstand eines Drahtstückes* aus dem Inhaltsbereich Elektrizitätslehre (vgl. Abschnitt 4.7 auf Seite 51) wurde in Studie F als automatisch ablaufende Trainingsaufgabe ausgearbeitet. Alle Erklärungen zum Aufgabenformat wurden dort audio-visuell umgesetzt. Beispielsweise hören die Schülerinnen und Schüler beim Teilaufgabentyp *Versuch aufbauen und testen* eine Erklärung, wie man Geräte auf der Simulationsfläche verschiebt, während gleichzeitig eine Videosequenz abläuft, die das Verschieben von Geräten zeigt. Unmittelbar im Anschluss können die Schülerinnen und Schüler alle Bedienmöglichkeiten bei diesem Teilaufgabentyp selbstständig ausprobieren. Beim Ausprobieren geht es nicht um die Bearbeitung der eigentlichen Aufgabenstellung, sondern ausschließlich um das Üben der Testbedienung. Zusätzlich zur Erklärung der Testbedienung wird auch die Zwischenlösung von Alina und Bodo vorgestellt, um den Schülerinnen und

Schülern aufzuzeigen, welche Art von Lösungen (z. B. Fließtext oder Stichpunkte) und welcher Lösungsumfang erwartet werden (Erwartungshorizont). Ein vollständiges Skript der Trainingsaufgabe ist in Anhang A.6 auf Seite 188 zu finden.

Exkurs: Vorgehensweisen zur Vorbereitung auf einen Test

Brunner, Artelt, Krauss & Baumert (2007, S. 112) unterscheiden in Anlehnung an Allalouf und Ben-Shakhar (1998, S. 32) drei miteinander verknüpfte Vorgehensweisen, um Testteilnehmende auf die Bearbeitung eines Tests vorzubereiten:

- *familiarity approach*: Testeilnehmende werden mit den Kernanforderungen der Testsituation vertraut gemacht. Zu diesem Zweck wird eine authentische Testsituation simuliert, in der die Teilnehmenden unter testähnlichen Bedingungen eine Beispielaufgabe bzw. einen Beispieltest bearbeiten.
- *content approach*: Testteilnehmende werden gezielt auf die Testinhalte vorbereitet.
- *test-wiseness approach*: Testteilnehmende werden mit allgemeinen Strategien zur Bewältigung des Tests oder mit Strategien zur Bearbeitung bestimmter Aufgabentypen vertraut gemacht.

Ziel dieser Vorgehensweisen ist es, die Testperformanz der Testteilnehmenden zu steigern.

Die Trainingsaufgabe folgt dem *familiarity approach* (vgl. Exkurs: Vorgehensweisen zur Vorbereitung auf einen Test). Daraus könnte sich die Gefahr ergeben, dass die Schülerinnen und Schüler zusätzlich gezielt auf die Testinhalte vorbereitet werden (*content approach*). Da der MeK-LSA Experimentiertest den Schülerinnen und Schülern unbekannt ist, kann ein einmaliges Durchlaufen einer Trainingsaufgabe nicht als *test-wiseness approach* bezeichnet werden. Inwieweit die Trainingsaufgabe dazu beiträgt, die Schülerinnen und Schüler im Sinne des *familiarity approach* mit dem Aufgabenformat vertraut zu machen, ohne sie gleichzeitig gezielt auf die Testinhalte vorzubereiten, wurde in Studie F empirisch untersucht.

12.2.1 Datenbasis und Vorgehensweise

An Studie F haben 13 Schülerinnen und Schüler der 9. Klasse aus zwei Schulen (eine Realschule, ein Gymnasium) in Nordrhein-Westfalen teilgenommen. Die Erhebung wurde im zweiten Schulhalbjahr 2012/13 durchgeführt. Die Erhebungszeit war auf 45 Minuten pro Schülerin bzw. Schüler begrenzt. Nach einer Trainingsübung zum begleitenden Think-Aloud (vgl. Abschnitt 8.3.1 auf Seite 84: Zur Methode der Analyse kognitiver Prozesse) bearbeiteten die Schülerinnen und Schüler in Einzelarbeit die vollständig automatisch ablaufende Trainingsaufgabe *Widerstand eines Drahtstückes* und die Testaufgabe *U-I-Kennlinie einer Glühlampe.* Während der Bearbeitung beider Aufgaben wurden die Schülerinnen und Schüler dazu aufgefordert, alle Überlegungen, die ihnen durch den Kopf gehen, laut auszusprechen (begleitendes Think-Aloud). Zusätzlich zum begleitenden Think-Aloud fand für die Testaufgabe *U-I-Kennlinie einer Glühlampe* unmittelbar nach der Bearbeitung jeder Teilaufgabe eine

Nachbefragung zur Vorgehensweise bei der Aufgabenbearbeitung und zur Bedienung des Tests statt (z. B. Wusstest du wie die Geräteauswahl zu bedienen ist? Woher?). Zum Abschluss wurden die Schülerinnen und Schüler zum wahrgenommenen Nutzen der Trainingsaufgabe befragt. Die Bearbeitungen und alle Äußerungen wurden über Bildschirmaufzeichnungen und Audiomitschnitte dokumentiert.

12.2.2 Datenauswertung

Um beurteilen zu können, inwieweit die Trainingsaufgabe die Schülerinnen und Schüler im Sinne des *familiarity approach* mit dem Aufgabenformat vertraut macht, wurden die Schülerhandlungen und Schüleräußerungen nach Indikatoren durchsucht. Es wurde für jeden Teilaufgabentyp (z. B. *Versuch aufbauen und testen*) erfasst, welche Aspekte nach der Bearbeitung der Trainingsaufgabe,

a) …Schwierigkeiten bei der Testbedienung bereiten,
b) …keine Schwierigkeiten bei der Testbedienung bereiten,
c) …dem Erwartungshorizont nicht entsprechen.

Indikatoren für Kriterium a) sind Handlungen oder Äußerungen, die darauf hindeuten, dass nicht bekannt ist, wie einzelne Elemente zu bedienen sind (z. B. „Wie konnte man nochmal die Geräte drehen?“). Indikatoren für Kriterium b) sind Handlungen oder Äußerungen, die zeigen, dass zu bedienende Elemente unmittelbar gefunden und zweckmäßig eingesetzt werden (z. B. „Um Kabel zu legen muss ich den Button *Kabel legen* aktivieren“). Indikatoren für Kriterium c) sind deutliche Abweichungen der Schülerlösungen bei der Bearbeitung der Testaufgabe von Alinas und Bodos Lösung, bezogen auf die Art und den Umfang der Lösung (z. B. realitätsgetreue Darstellung der Geräte in der Versuchsskizze anstelle einer symbolhaften Darstellung). Bei der Anwendung von Kriterium c) wird die Korrektheit der Schülerlösung nicht berücksichtigt, d. h. eine realitätsgetreue Darstellung der Geräte in einer Skizze wird zwar als Indikator gewertet, der nicht dem Erwartungshorizont entspricht, eine solche Darstellung kann aber trotzdem zu einer geeigneten Lösung führen.

Zusätzlich zu den Kriterien a) bis c) wurden Schülerhandlungen und Schüleräußerungen erfasst, die

d) …für eine gezielte Vorbereitung auf die Inhalte der Testaufgabe durch die Trainingsaufgabe sprechen.

Indikatoren für Kriterium d) sind Schülerhandlungen und Schüleräußerungen bei der Bearbeitung der Testaufgabe *U-I-Kennlinie einer Glühlampe*, die aufgrund einer Erinnerung an die Trainingsaufgabe durchgeführt werden (z. B. „Ich soll jetzt wieder eine Tabelle machen, glaube ich.“) oder die einen inhaltlichen Bezug zur Trainingsaufgabe herstellen, der für die Bearbeitung der Testaufgabe keine inhaltliche Bedeutung hat (z. B. Auswählen von Isolatorfüßen, weil diese in der Trainingsaufgabe verwendet wurden). Für die Darstellung der Ergebnisse (vgl. Abschnitt 12.2.3) werden die Indikatoren (Schülerhandlungen, Schüleräußerungen, Schülerlösungen) zu den Kriterien a) bis d) nach Bedienaspekten bzw.

Antwortaspekten gruppiert und paraphrasiert, um die Häufigkeit der Aspekte bestimmen zu können. Der paraphrasierte Indikator *Bedienung der Geräteauswahl* (vgl. Tabelle 12.2) gibt beispielsweise an, wie viele Schülerinnen und Schüler Handlungen bzw. Äußerungen vorgenommen haben, aus denen man schließen kann, dass die Bedienung der Geräteauswahl keine Schwierigkeiten bereitet hat.

12.2.3 Ergebnisse

Tabelle 12.2 stellt die Ergebnisse für die Kriterien a) (Schwierigkeiten bei der Testbedienung) und b) (keine Schwierigkeiten bei der Testbedienung) getrennt nach Teilaufgabentypen gegenüber.

Tabelle 12.2: Ergebnisse für die Kriterien a) (Schwierigkeiten bei der Testbedienung) und b) (keine Schwierigkeiten bei der Testbedienung) getrennt nach Teilaufgabentypen (N= Anzahl von Schülerinnen und Schülern)

Teilaufgabentyp	**Schwierigkeiten bei der Testbedienung**	**keine Schwierigkeiten bei der Testbedienung**[20]
Versuchsplan entwerfen	Betrachten des gesamten Bildschirminhalts durch Scrollen (N = 5)	Bedienung der Geräteauswahl (N = 13) Bedienung des Skizzen-Tools (N = 13)
Messprotokoll vorbereiten	Bearbeiten/ Löschen einer bereits gezeichneten Tabelle (N = 2)	Erstellen von Tabellen (N = 10)
Versuch aufbauen und testen	Bedienung der Funktion Kabel legen (N = 2)	Verschieben der Geräte (N = 13) Spannungsquelle an-/ ausschalten (N = 4) Drehschalter an Multimetern bedienen (N = 8)
Messung durchführen und dokumentieren	-	Spannungsquelle an-/ ausschalten (N = 12)
Datenauswertung durchführen (Diagramm)	Bedienung von ein bis zwei Funktionen des Diagramm-Tools unklar (N = 4)	Bedienung des Diagramm-Tools (N = 9)
Schlüsse ziehen	-	Ausfüllen des Textfeldes (N = 10)

Die Ergebnisse (vgl. Tabelle 12.2) zeigen, dass den Schülerinnen und Schülern – bis auf wenige Bedienelemente – die Bedienung des Tests während der Bearbeitung der Testaufgabe *U-I-Kennlinie einer Glühlampe* keine Schwierigkeiten bereitet hat. Schwierigkeiten bei mehr als zwei Schülerinnen und Schülern ergaben sich lediglich beim Teilaufgabentyp *Versuchsplan entwerfen* und beim Teilaufgabentyp *Datenauswertung durchführen*. Beim Teilaufgabentyp *Versuchsplan entwerfen* wurde nicht erkannt, dass gescrollt werden muss, um den gesamten Bildschirminhalt betrachten zu können. Beim Teilaufgabentyp *Datenauswertung durchführen* bereitete die Bedienung einzelner Diagramm-Tool-Funktionen (z. B. Funktion des Buttons

[20] Angaben mit N < 13 bedeuten in der Regel nicht, dass die anderen Schülerinnen und Schüler mit diesem Aspekt Schwierigkeiten hatten, sondern lediglich, dass kein paraphrasierter Indikator zu diesem Aspekt vorliegt.

Skalierung) Schwierigkeiten, wobei die Mehrzahl der Schülerinnen und Schüler keine Schwierigkeiten mit der Bedienung des Diagramm-Tools hatten.

Tabelle 12.3 stellt die Ergebnisse für Kriterium c) (Aspekte, die dem Erwartungshorizont nicht entsprechen) getrennt nach Teilaufgabentypen gegenüber.

Tabelle 12.3: Ergebnisse für Kriterium c) (Aspekte, die dem Erwartungshorizont nicht entsprechen) getrennt nach Teilaufgabentypen (N= Anzahl von Schülerinnen und Schülern)

Teilaufgabentyp	**Aspekte, die dem Erwartungshorizont nicht entsprechen**
Versuchsplan entwerfen	Skizze mit realitätsgetreuer Darstellung der Geräte (N = 1)
Messprotokoll vorbereiten	Aufgabenstellung unklar (N = 3)
Versuch aufbauen und testen	-
Messung durchführen und dokumentieren	Aufgabenstellung unklar (N = 2)
Datenauswertung durchführen (Diagramm)	vorgegebene Messwerte nicht in Koordinatensystem eingetragen (N = 1)
Schlüsse ziehen	-

Die Ergebnisse (vgl. Tabelle 12.3) zeigen, dass die Schülerinnen und Schülern in der Regel wussten, was von ihnen bei der Bearbeitung der Testaufgabe *U-I-Kennlinie einer Glühlampe* erwartet wird. Die Schülerlösungen entsprachen lediglich beim Teilaufgabentyp *Messprotokoll vorbereiten* bei drei Schülerinnen und Schülern nicht dem Erwartungshorizont. Den Schülerinnen und Schülern war bei diesem Teilaufgabentyp nicht klar, in welcher Form und in welchem Umfang die Lösung erfolgen soll. Die Ergebnisse (vgl. Tabelle 12.2 auf Seite 114 und Tabelle 12.3) sprechen insgesamt dafür, dass die Schülerinnen und Schüler durch die Trainingsaufgabe im Sinne des *familiarity approach* auf die Testaufgabe vorbereitet werden.

Bei den Teilaufgabentypen *Versuchsplan entwerfen* und *Messprotokoll vorbereiten* wurden Indikatoren für Kriterium d) (Schüleräußerungen und Schülerhandlungen, die für eine gezielte Vorbereitung auf die Testinhalte sprechen) gefunden (vgl. Tabelle 12.4 auf Seite 116). Eine gezielte Vorbereitung auf die Testinhalte findet relativ häufig beim Teilaufgabentyp *Versuchsplan entwerfen* statt. In der Testaufgabe *U-I-Kennlinie einer Glühlampe* wurden bei diesem Teilaufgabentyp zum einen häufig Geräte ausgewählt, die aus der Trainingsaufgabe bekannt sind (N = 3-4) und zum anderen Versuchsskizzen angefertigt, die der Skizze aus der Trainingsaufgabe ähneln (N = 8).

Tabelle 12.4: Ergebnisse für Kriterium d) (Schüleräußerungen und Schülerhandlungen, die für eine gezielte Vorbereitung auf die Testinhalte sprechen) getrennt nach Teilaufgabentypen (N = Anzahl von Schülerinnen und Schülern)

Teilaufgabentyp	**Schüleräußerungen und Schülerhandlungen, die für eine gezielte Vorbereitung auf die Testinhalte sprechen**
Versuchsplan entwerfen	Auswahl von Geräten, die aus der Trainingsaufgabe bereits bekannt sind (Isolatorfüße: N = 3; Spannungsquelle: N = 4) Versuchsskizze, die der Skizze aus der Trainingsaufgabe ähnelt (N = 8)
Messprotokoll vorbereiten	Tabelle als Darstellungsform für das Messprotokoll, da auch in der Trainingsaufgabe eine Tabelle gezeichnet wurde (N = 2)

Über alle Teilaufgabentypen hinweg stellen die Schülerinnen und Schüler bei der Bearbeitung der Testaufgabe *U-I Kennlinie einer Glühlampe* allerdings kaum inhaltliche Bezüge zur Trainingsaufgabe her. Die Gefahr einer Vorbereitung auf die Testaufgabe im Sinne des *content approach* scheint daher durch die Trainingsaufgabe nur bei den beiden oben beschriebenen Teilaufgabentypen zu bestehen.

12.3 Vergleich von konsekutivem und nicht-konsekutivem Aufgabenformat (Beitrag aus Studie G)

Das konsekutive Aufgabenformat des MeK-LSA Experimentiertests soll Schülerinnen und Schüler, die Schwierigkeiten mit der eigenständigen Strukturierung des Bearbeitungsprozesses haben, bei der Bearbeitung der Testaufgaben unterstützen. Durch die Vorstrukturierung in Teilaufgaben soll das Zeigen aller im Aufgabenentwicklungsmodell beschriebenen Fähigkeiten (vgl. Abbildung 3.1 auf Seite 30) ermöglicht werden. Durch das Bereitstellen von Zwischenlösungen sollen darüber hinaus Folgefehler einer falsch oder fehlerhaft bearbeiteten Teilaufgabe vermieden werden. Auf der anderen Seite müssen die Schülerinnen und Schüler durch die Vorstrukturierung möglicherweise nach einer (auch erfolgreich) bearbeiteten Teilaufgabe ihre Bearbeitungsstrategie anpassen. Folglich könnte durch die Vorgabe der Bearbeitungsreihenfolge das Zeigen experimenteller Performanz beeinflusst werden. Um diese Aspekte empirisch untersuchen zu können, wurde in Studie G als Vergleichsmaßstab ein nicht-konsekutives Aufgabenformat eingesetzt. Das nicht-konsekutive Aufgabenformat ist gekennzeichnet durch eine Simulationsfläche, auf der mit realitätsnahen Geräten virtuell experimentiert werden kann, sowie ein Onlineprotokoll, in dem weitere Teilaufgaben (z. B. Versuchsskizze anfertigen) bearbeitet werden können (vgl. eXkomp-Aufgabe in Abbildung 2.5 auf Seite 25). Die Bearbeitungsreihenfolge der Teilaufgaben ist nicht vorgegeben und Folgefehler werden nicht abgefangen.

In Studie G wurde untersucht, ob …

- Unterschiede im Schülerhandeln zwischen dem konsekutiven und dem nicht-konsekutiven Aufgabenformat feststellbar sind.
- die Schülerinnen und Schüler im konsekutiven Aufgabenformat alle Teilfähigkeiten zeigen können.
- im konsekutiven Format Folgefehler tatsächlich vermieden werden.

12.3.1 Datenbasis und Vorgehensweise

An Studie G haben zehn Schülerinnen und Schüler der Klassen 9 und 10 teilgenommen (Gymnasium und Oberschule). Die Schülerinnen und Schüler bearbeiteten die folgende übergeordnete experimentelle Aufgabenstellung im nicht-konsekutiven und im konsekutiven Aufgabenformat:

„Du sollst für 3 unterschiedliche Metalle (A, B, C) herausfinden, welches am besten elektrischen Strom leitet."

Die Aufgabenstellung wurde jeweils in das konsekutive und das nicht-konsekutive Aufgabenformat überführt (vgl. Eckloff, 2014, S. 23). Tabelle 12.5 zeigt den Ablauf der Datenerhebung in Studie G.

Tabelle 12.5: Ablauf der Datenerhebung in Studie G

Trainingsübung	begleitendes Think-Aloud
Trainingsaufgabe	Kennenlernen des nicht-konsekutiven Aufgabenformats
Aufgabe	Aufgabenbearbeitung im nicht-konsekutiven Format
Trainingsaufgabe	Kennenlernen des konsekutiven Aufgabenformats
Aufgabe	Aufgabenbearbeitung im konsekutiven Format

Nach einer Trainingsübung zum begleitenden Think-Aloud (vgl. Abschnitt 8.3.1 auf Seite 84: Zur Methode der Analyse kognitiver Prozesse) bearbeiteten die Schülerinnen und Schüler in Einzelarbeit die experimentelle Aufgabenstellung (siehe oben) zunächst im nicht-konsekutiven Aufgabenformat. Im Anschluss bearbeiteten die Schülerinnen und Schüler die gleiche Aufgabenstellung im konsekutiven Aufgabenformat. Vor der jeweiligen Aufgabenbearbeitung wurde eine Trainingsaufgabe zum Kennenlernen des jeweiligen Aufgabenformats durchgeführt (vgl. Tabelle 12.5). Während der Bearbeitung der Aufgaben wurden die Schülerinnen und Schüler dazu aufgefordert, alles laut auszusprechen, was ihnen bei der Bearbeitung der Aufgaben durch den Kopf geht (begleitendes Think-Aloud). Die Bearbeitungen und alle Äußerungen wurden über Bildschirmaufzeichnungen und Audiomitschnitte dokumentiert.

12.3.2 Datenauswertung

Um mögliche Unterschiede im Schülerhandeln zwischen dem konsekutiven und dem nicht-konsekutiven Aufgabenformat untersuchen zu können, wurden die Bildschirmaufzeichnungen und Audiomitschnitte zeitbasiert (10 Sekunden-Intervalle) kategorisiert. Zur Kategorisierung der Datensätze hat Eckloff (2014) im Rahmen seiner Abschlussarbeit (vgl. Abschnitt 7.1 auf Seite 77) in Anlehnung an Rogge (2010, S. 99) ein Kategoriensystem ausgearbeitet (vgl. Abbildung 12.2 auf Seite 118).

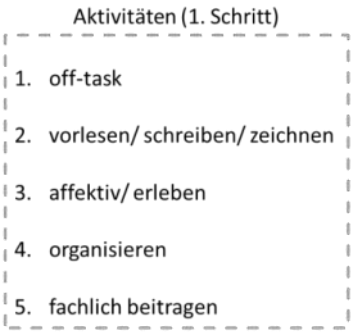


Abbildung 12.2: Von Eckloff (2014) in Anlehnung an Rogge (2010, S. 99) ausgearbeitetes Kategoriensystem zur Kategorisierung der Handlungen und Äußerungen im konsekutiven und nicht-konsekutiven Aufgabenformat (Kurzform; 1. Kategorisierungsschritt)

In einem ersten Schritt wird jedem 10 Sekunden-Intervall eine von fünf Aktivitäten zugeschrieben (vgl. Abbildung 12.2). Aktivitäten der Kategorie *fachlich beitragen* beziehen sich auf experimentbezogene Handlungen und Überlegungen und werden in Anlehnung an von Aufschnaiter und Rogge (2010, S. 103) noch hinsichtlich ihrer Qualität unterschieden. Für eine detaillierte Darstellung des Kategoriensystems sei an dieser Stelle auf Eckloff (2014) verwiesen. Für die Prüfung von Annahme II.II (*Die Bearbeitung des Tests ermöglicht das Zeigen experimenteller Performanz)* ist relevant, ob sich der Anteil fachlicher Beiträge zwischen dem konsekutiven und dem nicht-konsekutiven Aufgabenformat unterscheidet. Darüber ist zu prüfen, ob die Schülerinnen und Schüler die Aufgabenstellungen im konsekutiven Aufgabenformat vollständiger bearbeiten als im nicht-konsekutiven Format und ob durch das konsekutive Format tatsächlich Folgefehler vermieden werden.

12.3.3 Ergebnisse

Tabelle 12.6 zeigt die gemittelten prozentualen Anteile der 10 Sekunden-Intervalle an den Aktivitäten, getrennt nach konsekutivem und nicht-konsekutivem Aufgabenformat.

Tabelle 12.6: Gemittelte prozentuale Anteile der 10 Sekunden-Intervalle an den Aktivitäten (Kategorien), getrennt nach konsekutivem und nicht-konsekutivem Aufgabenformat (Angabe: MW (SD))

Aufgabenformat	**Kategorien**				
	off-task	vorlesen/ schreiben/ zeichnen	affektiv/ erleben	organisieren	fachlich beitragen
nicht-konsekutiv	6 (3)	9 (5)	9 (4)	18 (4)	58 (5)
konsekutiv	6 (3)	19 (4)	9 (4)	13 (4)	53 (7)

In beiden Aufgabenformaten nimmt die Kategorie *fachlich beitragen* mit einem gemittelten prozentualen Anteil von jeweils über 50 % der Zeitintervalle den größten Anteil ein. Ein bedeutsamer Unterschied zwischen den Aufgabenformaten findet sich nur für die Kategorie

vorlesen/schreiben/zeichnen (Wilcoxon-Test[21]: z = -2,70; p = .007; r = -.60), wobei diese Kategorie im konsekutiven Aufgabenformat einen höheren Anteil einnimmt.

Im konsekutiven Aufgabenformat wurden alle Teilaufgaben bearbeitet, während im nicht-konsekutiven Aufgabenformat insgesamt 35 % (28 von 80) der Teilaufgaben nicht bearbeitet wurden. Am häufigsten unbearbeitet blieben die Teilaufgaben *Datenauswertung durchführen* (N = 8) und *Schlüsse ziehen* (N = 8). In Einzelfallanalysen konnte Eckloff (2014) Hinweise finden, dass im konsekutiven Format Folgefehler vermieden werden. Beispielsweise gelingt es einem Probanden im konsekutiven Format, mit dem vorgegebenen Versuchsaufbau eine Messreihe aufzunehmen, obwohl mit dem eigenen Versuchsaufbau keine vollständigen Messungen möglich wären, da kein Spannungsmessgerät im Versuchsaufbau integriert wurde.

12.4 Diskussion

Die Ergebnisse der Schülerbefragung aus Studie C (vgl. Abschnitt 12.1). zeigen, dass die Testsituation von den Schülerinnen und Schülern mehrheitlich positiv wahrgenommen wird. Das trifft gleichermaßen auf die Wahrnehmung des konsekutiven Aufgabenformats, die Wahrnehmung der Testbedienung und die Wahrnehmung der Zwischenlösungen von Alina und Bodo zu. Lediglich in wenigen Fällen wurde es als störend empfunden, dass nach erfolgreicher eigener Lösung mit der Zwischenlösung von Alina und Bodo weitergearbeitet werden musste. Es finden sich jedoch keine Hinweise, die auf demotivierende Effekte durch die Gestaltung der Testsituation schließen lassen. Insgesamt sprechen die Ergebnisse der Schülerbefragung dafür, dass die getroffenen Strukturierungsmaßnahmen die Schülerinnen und Schüler bei der Bearbeitung der Aufgaben unterstützen. Offen bleibt, inwieweit die Einschätzungen aufgrund sozialer Erwünschtheit zu positiv ausfallen. Die Testleiter waren allerdings dazu angehalten, eine Gesprächssituation zu schaffen, die eine ehrliche Feedbackkultur fördert.

Die Ergebnisse aus Studie F (Evaluation der Trainingsaufgabe) sprechen dafür, dass die Schülerinnen und Schüler durch die Trainingsaufgabe im Sinne des *familiarity approach* auf die Testaufgabe vorbereitet werden. Die Gefahr einer Vorbereitung auf die Testaufgabe im Sinne des *content approach* scheint durch die Trainingsaufgabe dagegen nur bei den Teilaufgabentypen *Versuchsplan entwerfen* und *Messprotokoll vorbereiten* zu bestehen und ist mit hoher Wahrscheinlichkeit auf den identischen Inhaltsbereich (Elektrizitätslehre) von Trainingsaufgabe und Testaufgabe zurückzuführen. Auf Basis der Ergebnisse aus Studie F wurde die Trainingsaufgabe für den weiteren Projektverlauf überarbeitet (z. B. angepasste Erklärung zur Bedienung des Diagramm-Tools; Austausch der Geräte und zu zeichnender Skizze beim Teilaufgabentyp *Versuchsplan entwerfen*). Da in der automatisch ablaufenden Trainingsaufgabe audio-visuelle Erklärungen eingesetzt wurden, hätte jeder Schüler bzw. jede Schülerin bei der Testbearbeitung einen Kopfhörer benötigt. Aus pragmatischen Gründen wurde daher für den weiteren Projektverlauf auf eine vollständig automatisierte

[21] Aufgrund einer möglichen Alpha-Fehler Kumulierung wird das Signifikanz-Niveau mit der Bonferroni-Korrektur auf $\alpha_k = \frac{.05}{5} = .01$ festgelegt.

Trainingsaufgabe verzichtet. Die nicht automatisierte Trainingsaufgabe basiert auf einem mit der automatisierten Trainingsunit vergleichbaren Skript. Das Skript umfasst Erklärungen und Handlungsanweisungen die vom Testleiter unter Zuhilfenahme von animierten Bildern und Videosequenzen vorgetragen werden. Es kann also plausibel angenommen werden, dass sich die Ergebnisse aus Studie F auf die nicht automatisierte Trainingsaufgabe übertragen lassen. Zusätzlich werden die in Studie F aufgetretenen Schwierigkeiten (vgl. Abschnitt 12.2) durch die Überarbeitung der Trainingsaufgabe minimiert. Einschränkend muss erwähnt werden, dass aufgrund des Studiendesigns keine Aussage darüber getroffen werden kann, inwieweit die Schülerinnen und Schüler die Handlungsmöglichkeiten und den Erwartungshorizont auch ohne Durchlaufen einer Trainingsaufgabe verstanden hätten. Insgesamt liegen jedoch keine belastbaren Hinweise vor, die gegen den Einsatz der Trainingsaufgabe zum Kennenlernen des Aufgabenformats sprechen.

Die Ergebnisse aus Studie G zeigen, dass sich der Anteil fachlicher Beiträge (experimentbezogene Überlegungen und Handlungen) zwischen dem konsekutiven Aufgabenformat des MeK-LSA Experimentiertests und dem nicht-konsekutiven Aufgabenformat nicht unterscheidet und experimentbezogene Überlegungen und Handlungen in beiden Aufgabenformaten den größten Anteil einnehmen. Bedeutsame Unterschiede zwischen den Aufgabenformaten ergeben sich lediglich in der Kategorie *vorlesen/schreiben/zeichnen*, wobei der gemittelte prozentuale Anteil im konsekutiven Format höher als im nicht-konsekutiven Format ist. Dieser Befund deutet darauf hin, dass Schülerinnen und Schüler durch das konsekutive Format eher die Notwendigkeit sehen, die Aufgabenstellungen zu lesen und durchgängig zu bearbeiten. Das spricht dafür, dass das konsekutive Aufgabenformat die Schülerinnen und Schüler in beabsichtigter Weise bei der Strukturierung des Bearbeitungsprozesses unterstützt, ohne die grundsätzliche experimentelle Vorgehensweise zu verändern. Darüber hinaus zeigt sich, dass die Schülerinnen und Schüler im konsekutiven Aufgabenformat des MeK-LSA Experimentiertests durch die Vorgabe von Zwischenlösungen (Vermeidung von Folgefehlern) Teilaufgaben bearbeiten, die sie im nicht-konsekutiven Aufgabenformat nicht bearbeitet haben bzw. bearbeiten konnten. Einschränkend muss erwähnt werden, dass aufgrund des Studiendesigns mögliche Lerneffekte zugunsten des konsekutiven Aufgabenformats nicht ausgeschlossen werden können. Andererseits könnte es auch sein, dass einige, bereits im nicht-konsekutiven Format geäußerten, fachlichen Überlegungen im nachfolgenden konsekutiven Format nicht mehr so vertieft wiederholt und verbalisiert wurden, sodass man den Anteil fachlicher Beiträge im konsekutiven Format vielleicht sogar unterschätzt.

Die zur Prüfung von Annahme II.II (*Die Bearbeitung des Tests ermöglicht das Zeigen experimenteller Performanz)* durchgeführten Studien basieren auf vergleichsweise kleinen Gelegenheitsstichproben und unterliegen studiendesignbedingten Einschränkungen (siehe oben). Die Ergebnisse sind daher auf der einen Seite entsprechend vorsichtig zu interpretieren. Auf der anderen Seite zeigen die Studienergebnisse, dass die vom Testentwicklungsteam getroffenen Maßnahmen (z. B. konsekutives Aufgabenformat; Trainingsaufgabe zum Kennenlernen des Aufgabenformats) von den Schülerinnen und Schülern in der intendierten Weise wahrgenommen und genutzt werden.

13 Anteil experimentbezogener Überlegungen (Annahme II.III)

In diesem Kapitel wird die folgende Annahme geprüft und diskutiert:

Der Anteil experimentbezogener Überlegungen ist bei der Bearbeitung von on-screen Aufgaben vergleichbar hoch wie bei der Bearbeitung inhaltlich identischer hands-on Aufgaben (AII.III).

In Kapitel 11 (Seiten 103-107) konnte gezeigt werden, dass Annahme II.I (*Die Schülerinnen und Schüler stellen bei der Bearbeitung der on-screen Aufgaben überwiegend experimentbezogene Überlegungen an)* beibehalten werden kann. Aus diesem Befund kann allerdings nicht unmittelbar geschlossen werden, dass die experimentbezogenen Überlegungen in beiden Testformaten vergleichbar sind und somit auch Annahme II.III beibehalten werden kann. Zur Prüfung von Annahme II.III werden daher die on-screen und hands-on Bearbeitungen der Aufgabe *Ausdehnung eines Gummiban*des aus dem begleitenden Think-Aloud herangezogen (vgl. Studie C in Abschnitt 8.3 auf Seite 83).

13.1 Beitrag des begleitenden Think-Aloud aus Studie C

Um zu untersuchen, ob sich die Art der Überlegungen (z. B. experimentbezogene Überlegungen) von Schülerinnen und Schülern zwischen der Bearbeitung einer Aufgabe im hands-on und on-screen Format unterscheidet, haben Schülerinnen und Schüler die Aufgabe *Ausdehnung eines Gummibandes* entweder im on-screen oder im hands-on Format bearbeitet[22] und alle Überlegungen während der Aufgabenbearbeitung laut ausgesprochen (begleitendes Think-Aloud). Die Analyse beschränkt sich auf die Teilaufgabentypen *Versuchsplan entwerfen*, *Versuch aufbauen und testen*, *Messung durchführen und dokumentieren* sowie *Datenauswertung durchführen*. Die Auswahl der Aufgabe *Ausdehnung eines Gummibandes* und die in der Analyse berücksichtigten Teilaufgabentypen wurden in Abschnitt 7.2 auf Seite 80 begründet.

13.1.1 Kategorisierung der Daten

Die mittels Bildschirmaufzeichnungen dokumentierten Bearbeitungen von Teilaufgaben wurden zusammen mit den Audiomitschnitten zeitbasiert kategorisiert (10 Sekunden-Intervalle). Die on-screen Bearbeitungen von Teilaufgaben (Datensätze) der Aufgabe *Ausdehnung eines Gummibandes* wurden mit dem in Abschnitt 11.1.1 auf Seite 103 beschriebenen Kategoriensystem ausgewertet. Türck (2014) hat dieses Kategoriensystem im Rahmen ihrer Abschlussarbeit (vgl. Abschnitt 7.1 auf Seite 77) für das hands-on Format adaptiert. Da bei der Bearbeitung von hands-on Aufgaben keine Äußerungen zur *Computerbedienung* zu erwarten sind, wurde diese Kategorie durch eine vergleichbare Kategorie *manueller Umgang* (z. B. *Das wackelt hier aber ganz schön*) ersetzt (vgl. Türck, 2014). Die dieser Kategorie zugeordneten Überlegungen beziehen sich auf das erforderliche manuelle Geschick und nicht auf physikalisch-experimentelle Überlegungen zur Bearbeitung der

[22] Ein Schüler hat die Aufgabe in beiden Formaten bearbeitet.

Aufgabe. Alle weiteren Kategorien konnten aus dem Kategoriensystem für die on-screen Datensätze übernommen werden.

Für die Güte des Kategorisierungsverfahrens ergeben sich sowohl für die on-screen, als auch für die hands-on Datensätze der Aufgabe *Ausdehnung eines Gummibandes* zufriedenstellende Werte (on-screen: .64< κ < .77; mittleres κ = .70; hands-on: .64 < κ < .90; mittleres κ = .78). Insgesamt sind für die Aufgabe *Ausdehnung eines Gummibandes* 9 % aller Zeitintervalle (485 von 5151) aus den on-screen Datensätzen und 20 % (626 von 3172) aller Zeitintervalle aus den hands-on Datensätzen doppelt kodiert worden.

In die Analyse gingen nur Datensätze ein, bei denen die Schülerinnen und Schüler ihre Überlegungen in mindestens 50 % der 10 Sekunden-Intervalle verbalisiert haben (vgl. Begründung in Abschnitt 11.1.2 auf Seite 104). Aufgrund dieses Kriteriums mussten lediglich 10 % aller erhobenen Datensätze für die Aufgabe *Ausdehnung eines Gummibandes* (26 von 259) von der Auswertung ausgeschlossen werden. Insgesamt sind 233 Datensätze aus dem *begleitenden* Think-Aloud in die Analyse eingeflossen. Tabelle 13.1 zeigt die Anzahl der in der Analyse berücksichtigten Datensätze getrennt nach Format und Teilaufgabentyp.

Tabelle 13.1: Anzahl in der Analyse berücksichtigter Datensätze getrennt nach Format und Teilaufgabentyp

Teilaufgabentyp	Ausdehnung eines Gummibands		Summe
	on-screen	hands-on	
Versuchsplan entwerfen	38	23	61
Versuch aufbauen und testen	36	22	58
Messung durchführen und dokumentieren	36	23	59
Datenauswertung durchführen	35	20	55
Summe	145	88	233

13.1.2 Ergebnisse

Tabelle 13.2 auf Seite 123 zeigt den mittleren prozentualen Anteil der 10 Sekunden-Intervalle getrennt nach Kategorien, Teilaufgabentypen und Format. Die Anteile sind analog zu Kapitel 11 (vgl. Abschnitt 11.1.3 auf Seite 105) zu interpretieren. Den größten Anteil nehmen bei allen vier Teilaufgabentypen, unabhängig vom Testformat, physikalisch-experimentelle Überlegungen ein. Eine zweifaktorielle Varianzanalyse (abhängige Variable: Anteil physikalisch-experimenteller Überlegungen[23]; Faktoren: Teilaufgabentyp und Format) zeigt, dass der Anteil physikalisch-experimenteller Überlegungen durch den Teilaufgabentyp ($F(3,225)=55{,}22$, $p<.001$, $\eta_p^2=.424$), nicht aber durch das Format ($F(1,225)=2{,}58$, $p=.110$, $\eta_p^2=.011$) und nicht

[23] Varianzhomogenität gegeben: $F(7, 225) = 1{,}51$; $p=.164$

durch eine Interaktion von Teilaufgabentyp und Format (F(3,225)=1,87, p=.135, η_p^2=.024) beeinflusst wird.

Tabelle 13.2: mittlerer prozentualer Anteil der 10 Sekunden Intervalle getrennt nach Bewertungskategorien, Teilaufgabentypen und Format (hellgrau: on-screen; dunkelgrau: hands-on) für die Aufgabe Ausdehnung eines Gummibandes (Angabe: MW (SD) in %)

Teilaufgabentyp	**Kategorien**									
	physikalisch-experimentell		reproduktiv		Computerbedienung/ manueller Umgang		Sonstiges		keine Verbalisierung	
Versuchsplan entwerfen	50 (16)	43 (11)	24 (12)	32 (14)	2 (3)	0 (1)	9 (7)	9 (8)	16 (15)	16 (14)
Versuch aufbauen und testen	53 (15)	57 (17)	4 (4)	6 (6)	20 (12)	10 (10)	10 (8)	8 (7)	13 (13)	19 (11)
Messung durchführen und dokumentieren	78 (12)	71 (15)	4 (3)	5 (5)	3 (3)	3 (6)	5 (5)	6 (6)	10 (12)	15 (14)
Datenauswertung durchführen	78 (15)	75 (10)	2 (2)	5 (5)	4 (5)	1 (2)	4 (5)	6 (5)	11 (13)	13 (10)
Gesamt	64 (20)	61 (19)	9 (11)	12 (14)	7 (10)	4 (7)	7 (7)	7 (7)	13 (14)	16 (12)

13.2 Diskussion

Die Ergebnisse (siehe oben) zeigen, dass experimentbezogene Überlegungen während der Bearbeitung aller Teilaufgabentypen unabhängig vom Format die größte Rolle spielen. Gleichzeitig sind formatspezifische Überlegungen (Computerbedienung; manueller Umgang) den experimentbezogenen Überlegungen deutlich untergeordnet. Detailanalysen zeigen, dass der Anteil formatspezifischer Überlegungen beim Teilaufgabentyp *Versuch aufbauen und testen* in beiden Formaten am größten ist. Dieser Befund ist erwartungskonform, da der Teilaufgabentyp *Versuch aufbauen und testen* die meisten Freiheitsgrade beim Umgang mit dem Experimentiermaterial zulässt. Im on-screen Format kann der hohe Anteil von Überlegungen zur Computerbedienung beim Teilaufgabentyp *Versuch aufbauen und testen* (20 %) auf Bedienprobleme im Umgang mit der interaktiven Simulation (z. B. Befestigen der Messzeiger am Maßstab) zurückgeführt werden. Für den weiteren Projektverlauf wurden daher weitere Hilfestellungen zur Bedienung der interaktiven Simulation zur Verfügung gestellt, sodass sich dieser Anteil reduziert, im Idealfall sogar dem Anteil im hands-on Format (10 %) angenähert haben sollte.

Als Einschränkungen ergeben sich zum einen analog zu den Ergebnissen in Kapitel 11, dass die zeitbasierte Kodierung nur eine Annäherung an die zeitliche Verteilung der Überlegungen darstellt, und dass mit dem Verzicht auf eine Transkription der Daten ein Informationsverlust einhergeht, der qualitative Detailanalysen erschwert (zur Bedeutsamkeit der Einschränkungen: vgl. Abschnitt 11.2 auf Seite 106). Zum anderen beschränkt sich der Formatvergleich lediglich auf eine Aufgabe (*Ausdehnung eines Gummibandes*). Folglich ist die Belastbarkeit der Daten eingeschränkt. Auf der anderen Seite stellen die Ergebnisse zur Aufgabe *Ausdehnung eines*

Gummibandes eine Worst-Case Abschätzung dar, da bei dieser Aufgabe in der Oberflächenstruktur maximale Unterschiede bezüglich der Handlungsmöglichkeiten im on-screen und hands-on Format bestehen (vgl. Abschnitt 7.2 auf Seite 80). Schon bei dieser Worst-Case Abschätzung zeigen sich keine bedeutsamen Formatunterschiede bezüglich des Anteils experimentbezogener Überlegungen. Es ist folglich eher unwahrscheinlich, dass sich bei Aufgaben, die geringere Unterschiede bezüglich der Handlungsmöglichkeiten aufweisen, ein bedeutsamer Einfluss des Testformats auf den Anteil experimentbezogener Überlegungen zeigt.

14 Kognitive Belastung (Annahme II.IV)

In diesem Kapitel wird die folgende Annahme geprüft und diskutiert:

Die wahrgenommene kognitive Belastung ist bei on-screen Aufgaben vergleichbar hoch wie bei inhaltlich identischen hands-on Aufgaben (AII.IV).

Eine Bedrohung für das Beibehalten der Annahme besteht darin, dass formatspezifische Eigenschaften, wie z. B. die virtuellen Handlungen in den Simulationen oder das manuelle Handling im Realexperiment, zu einer nicht experimentbezogenen kognitiven Belastung des Arbeitsgedächtnisses führen könnten. Für die eigentliche Bearbeitung einer inhaltlich identischen Aufgabe würde in der Folge ein unterschiedlicher Anteil an Denkkapazität in beiden Formaten zur Verfügung stehen. Zur Prüfung der Annahme werden die Daten zur wahrgenommenen kognitiven Belastung aus den Studie D und E (vgl. Abschnitte 8.4 und 8.5 auf Seite 86 bzw. 89) berücksichtigt.

14.1 Messung der kognitiven Belastung

Zur Messung der wahrgenommenen kognitiven Belastung wurde in den Studien D und E das Mental-Effort Item von Paas (1992) und das Difficulty Item von Kalyuga et al. (1999) eingesetzt. (vgl. Abschnitt 8.4.1 auf Seite 88: Zur Methode der Messung kognitiver Belastung). Die beiden Items werden im Folgenden als CL-Itemtypen bezeichnet. Die Formulierung der CL-Itemtypen wurde für die vorliegende Arbeit für jeden Teilaufgabentyp angepasst. Abbildung 14.1 zeigt die beiden angepassten Items beispielhaft für den Teilaufgabentyp *Versuch aufbauen und testen.*

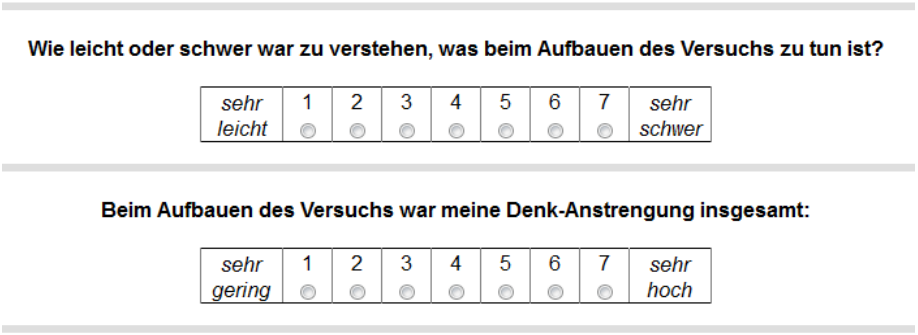


Abbildung 14.1: Items zur Messung der wahrgenommenen kognitiven Belastung (angepasste Version für Teilaufgabentyp Versuch aufbauen und testen*; oben: Difficulty Item nach Kalyuga et al. (1999); unten: Mental-Effort Item nach Paas (1992))*

Im Gegensatz zum Originalitem wird auch für das Mental-Effort Item von Paas (1992) anstelle einer neunstufigen Rating-Skala eine siebenstufige Rating-Skala verwendet (vgl. Schwamborn, Thillmann, Opfermann & Leutner, 2011). Ein möglicher Einfluss des Formats wird varianzanalytisch überprüft (Details in den Abschnitten 14.2 und 14.3). Die Bearbeitungsreihenfolge der Formate wurde in den Studien D und E abgewechselt (vgl. Abbildungen 8.1 und 8.3 auf Seite 87 bzw. 90). Auf diese Weise kann auch ein möglicher Einfluss der Formatreihenfolge kontrolliert werden.

14.2 Beitrag der Einschätzung zur wahrgenommenen kognitiven Belastung aus Studie D

42 Biologiestudierende schätzten für drei Aufgaben (*Ausdehnung eines Gummi*bandes, *U-I Kennlinie einer Glühlamp*e, *Brechung am Halbkreisblock*) jeweils unmittelbar nach der Bearbeitung ihre Denkanstrengung und die empfundene Aufgabenschwierigkeit auf einer siebenstufigen Rating-Skala für jeden untersuchten Teilaufgabentyp (*Versuchsplan entwerfen*, *Versuch aufbauen und testen*, *Messung durchführen und dokume*ntieren) in beiden Formaten (on-screen, hands-on) ein (vgl. Abbildung 8.1 auf Seite 87). In Anlehnung an van Gog et al. (2012) wurde die über die drei Aufgaben gemittelte wahrgenomme kognitive Belastung getrennt nach Teilaufgabentyp (z. B. *Versuch aufbauen und testen*), Format (on-screen, hands-on) und CL-Itemtyp bestimmt. Die Ergebnisse sind in Abbildung 14.2 dargestellt.

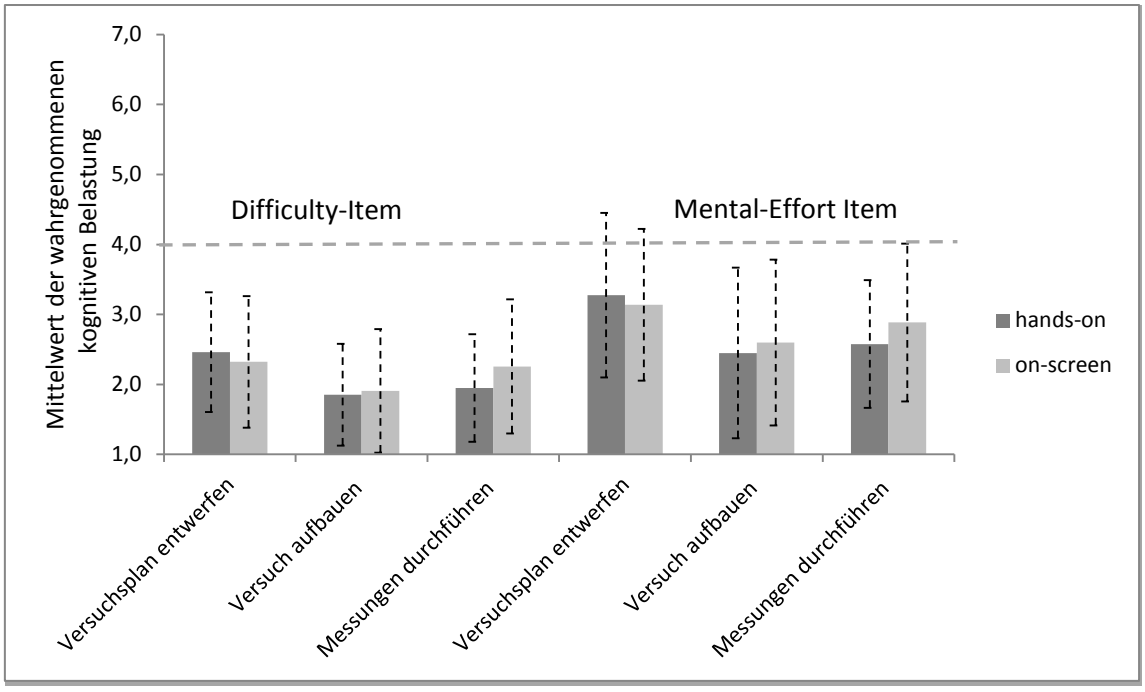


Abbildung 14.2: Über die Aufgaben gemittelte wahrgenommene kognitive Belastung getrennt nach Format, CL-Itemtyp und Teilaufgabentyp (1 ≙ sehr niedrige kognitive Belastung; 7 ≙ sehr hohe kognitive Belastung; senkrecht gestrichelte Linien ≙ Standardabweichung; waagerecht gestrichelte Linie ≙ mittlerer kognitiver Belastung)

Rein deskriptiv zeigt sich, dass die über die drei Aufgaben gemittelte wahrgenommene kognitive Belastung in beiden Formaten, beiden CL-Itemtypen und bei allen drei Teilaufgabentypen deutlich unter einer mittleren Einschätzung von 4,0 (maximal: 3,3) liegt. Eine Varianzanalyse mit den beiden messwiederholten Faktoren Format und CL-Itemtyp unter Berücksichtigung der Bearbeitungsreihenfolge des Formats und des Teilaufgabentyps als Zwischensubjektfaktoren belegt, dass die wahrgenommene kognitive Belastung durch den CL-Itemtyp ($F(1,78)= 77.70$, $p<.001$, $\eta_p^2=.499$) und den Teilaufgabentyp ($F(2,78)= 3.51$, $p=.035$, $\eta_p^2=.083$) beeinflusst wird, jedoch nicht durch die Bearbeitungsreihenfolge ($F(1,78)= 1.77$, $p=.188$, $\eta_p^2=.022$) und das Format ($F(1,78) = 1.64$, $p = .204$, $\eta_p^2 = .021$).

14.3 Beitrag der Einschätzung zur wahrgenommenen kognitiven Belastung aus Studie E

In Studie E (vgl. Abschnitt 8.5 auf Seite 89) schätzten 19 Schülerinnen und Schüler für die Aufgabe *Spielzeugauto auf einer Rampe* jeweils unmittelbar nach dem Anfertigen eines Messwertediagramms ihre Denkanstrengung und die empfundene Aufgabenschwierigkeit auf einer siebenstufigen Rating-Skala ein. Abbildung 14.3 zeigt die mittlere wahrgenommene kognitive Belastung getrennt nach Format (on-screen, hands-on) und CL-Itemtyp.

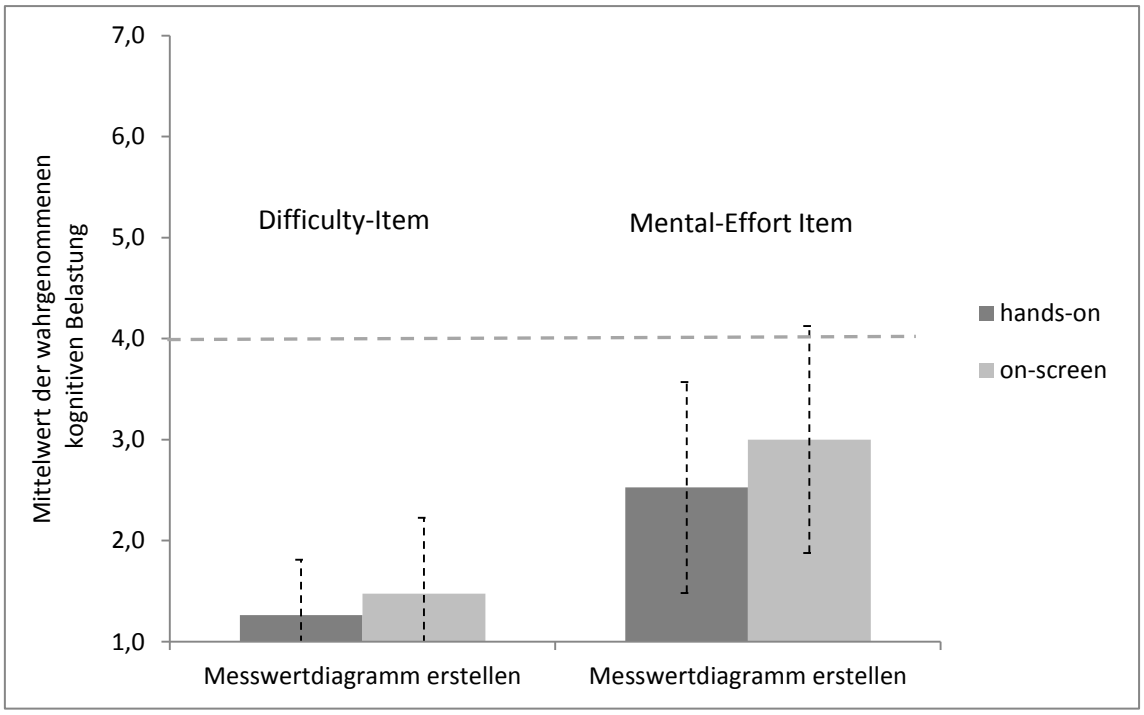


Abbildung 14.3: mittlere wahrgenommene kognitive Belastung getrennt nach Format und CL-Itemtyp (1 ≙ sehr niedrige kognitive Belastung; 7 ≙ sehr hohe kognitive Belastung; senkrecht gestrichelte Linien ≙ Standardabweichung; waagerecht gestrichelte Linie ≙ mittlere kognitive Belastung)

Rein deskriptiv zeigt sich, dass die mittlere wahrgenommene kognitive Belastung in beiden Formaten und CL-Itemtypen deutlich unter einer mittleren Einschätzung von 4,0 (maximal: 3,0) liegt. Eine Varianzanalyse mit den beiden messwiederholten Faktoren Format und CL-Itemtyp unter Berücksichtigung der Bearbeitungsreihenfolge des Formats als Zwischensubjektfaktor belegt, dass die wahrgenommene kognitive Belastung durch den CL-Itemtyp ($F(1,17) = 44{,}78$, $p < .001$, $\eta_p^2 = .725$) beeinflusst wird, jedoch nicht durch die Bearbeitungsreihenfolge ($F(1,17) = 1{,}48$, $p = .240$, $\eta_p^2 = .080$) und nicht durch das Testformat ($F(1,17) = 3{,}38$, $p = .084$, $\eta_p^2 = .166$).

14.4 Diskussion

Die Ergebnisse zur Einschätzung der wahrgenommenen kognitiven Belastung zeigen, dass sowohl im on-screen als auch im hands-on Format ein vergleichbarer Anteil an Denkkapazität genutzt wurde (vgl. Abschnitte 14.2 & 14.3). Dieser Anteil liegt unabhängig vom Teilaufgabentyp deutlich unterhalb einer neutralen Einschätzung von 4,0 auf der siebenstufigen Rating-Skala (Skala von 1 bis 7). Das spricht dafür, dass weder bei der Bearbeitung der on-screen Aufgaben, noch bei der Bearbeitung der hands-on Aufgaben

formatspezifische Eigenschaften (z. B. virtuelle Handlungen in den Simulationen oder manuelles Handling der Geräte im Realexperiment) zu einer Überschreitung der Verarbeitungskapazität des Arbeitsgedächtnisses führen. In beiden Formaten steht folglich genügend Kapazität im Arbeitsgedächtnis zur Verfügung, um die experimentellen Aufgaben (potentiell) erfolgreich zu bewältigen. Die wahrgenommene kognitive Belastung hängt allerdings zum einen erwartungskonform von der Art des CL-Itemtyps ab (Difficulty geringer als Mental-Effort), und zum anderen vom Teilaufgabentyp. Detailbetrachtungen bezüglich des Teilaufgabentyps in Studie D zeigen beispielsweise, dass die wahrgenommene kognitive Belastung beim Teilaufgabentyp *Versuchsplan entwerfen* am höchsten ist und beim Teilaufgabentyp *Versuch aufbauen und testen* am geringsten. Das ist plausibel, da die Studierenden aus der Schule und aus dem physikalischen Praktikum mit dem eigenständigen Aufbauen eines Versuchs (nach Anleitung) vertrauter sind als mit der eigenständigen Planung eines Experiments.

Die Einschätzungen zur wahrgenommenen kognitiven Belastung aus Studie E fügen sich stimmig in das Bild aus Studie D ein, trotz der unterschiedlichen Stichproben (Studie D: Biologiestudierende; Studie E: Schülerinnen und Schüler einer 9. Klasse) und der unterschiedlichen Teilaufgabentypen (Studie D: *Versuchsplan entwerfen*, *Versuch aufbauen und testen*, *Messung durchführen und dokumentieren*; Studie E: *Datenauswertung durchführen*).

Die größte Einschränkung ergibt sich aufgrund der gewählten Stichprobe in Studie D (Biologiestudierende). Auch wenn die Studierenden breit gestreute, aber überwiegend mit der Zielgruppe des MeK-LSA Experimentiertests vergleichbare Vorerfahrungen mit physikalischen Experimenten besitzen (vgl. Abschnitt 8.4.2 auf Seite 89), kann aus den Ergebnissen nicht unmittelbar auf die wahrgenommene kognitive Belastung in der Zielgruppe (Schülerinnen und Schüler am Ende der Sekundarstufe I) geschlossen werden. Die Ergebnisse für die Biologiestudierenden liefern jedoch keine Hinweise, die *gegen* das Beibehalten von Annahme II.IV sprechen. Die Ergebnisse aus Abschnitt 10.2 auf Seite 99 zur Häufigkeit experimenteller Anforderungen im Physikunterricht zeigen sogar, dass die in Studie D an die Studierenden gestellten Anforderungen im Unterricht mindestens ähnlich häufig vorkommen wie die in Studie E gestellten Anforderungen. Die in Studie D gestellten Anforderungen *ein Experiment selbstständig aufbauen* und *mehrere Messwerte aufnehmen* kommen tendenziell sogar deutlich häufiger vor. Daher ist nicht zu erwarten, dass die Verarbeitungskapazität im Arbeitsgedächtnis von Schülerinnen und Schülern bei diesen Anforderungen überschritten wird, auch wenn bisher keine empirischen Daten vorliegen, die diese Annahme weiter absichern.

Prüfung der Annahmen aus Teil III des INA: Die beobachtete Performanz wird in geeigneter Art und Weise in Testwerte überführt

Die zwei Annahmen zur dritten übergeordneten Aussage aus dem INA *Die beobachtete Performanz wird in geeigneter Art und Weise in Testwerte überführt* werden in Kapitel 15 und Kapitel 16 geprüft und diskutiert.

15 Bewertungsmaßstab und experimentelle Performanz (Annahme III.I)

In diesem Kapitel wird die folgende Annahme geprüft und diskutiert:

Der Bewertungsmaßstab spiegelt die beobachtete experimentelle Performanz wider (AIII.I).

Zur Beurteilung der Qualität der Bearbeitungen von Teilaufgaben wurde ein detaillierter Bewertungsmaßstab für die Aufgaben des MeK-LSA Experimentiertests durch das Testentwicklungsteam ausgearbeitet (vgl. Abschnitt 3.3 auf Seite 34). Laut zugrundeliegender Definition *experimenteller Kompetenz* (vgl. Abschnitt 1.3 auf Seite 13) gehört die fachliche Korrektheit zur experimentellen Performanz. Die fachliche Korrektheit wird im Bewertungsmaßstab explizit berücksichtigt. Darüber hinaus wird in der Definition *experimentelle Kompetenz* als latente Fähigkeit verstanden, die auf mindestens intuitiv regelbasiertem Vorgehen basiert. Die Regelbasiertheit der Vorgehensweise ist im Bewertungsmaßstab nicht explizit berücksichtigt, da in der Regel keine verbalisierten Überlegungen der Schülerinnen und Schüler vorliegen, anhand derer die Regelbasiertheit der Vorgehensweise beurteilt werden könnte. Zur Prüfung von Annahme III.I muss daher untersucht werden, ob eine höhere Bewertungsstufe auch mit einem höheren Anteil regelbasierten Vorgehens einhergeht.

15.1 Beitrag des begleitenden Think-Aloud aus Studie C

In Abschnitt 11.1 (Seiten 103-106) wurde für 567 on-screen Bearbeitungen von Teilaufgaben (Datensätze) aus Studie C (vgl. Abschnitt 8.3 auf Seite 83) der mittlere prozentuale Anteil physikalisch-experimenteller Überlegungen bestimmt. Die Datensätze beziehen sich auf die vier on-screen Aufgaben *Ausdehnung eines Gummibandes*, *Brechung am Halbkreisblock*, *Leistung von Glühlampen* und *Reihenschaltung von Glühlampen* und die Teilaufgaben des Typs *Versuchsplan entwerfen*, *Versuch aufbauen und testen*, *Messung durchführen und dokumentieren* sowie *Datenauswertung durchführen*. Die Auswahl der Datensätze wurde in Abschnitt 11.1.2 auf Seite 104 begründet.

15.1.1 Regelbasiertheit physikalisch-experimenteller Überlegungen

Zur Prüfung von Annahme III.I (*Der Bewertungsmaßstab spiegelt die beobachtete experimentelle Performanz wider*) werden die physikalisch-experimentellen Überlegungen der 567 on-screen Datensätze in Anlehnung an von Aufschnaiter und Rogge (2010; S. 101-106), zunächst nach ihrer *Regelbasiertheit* zeitbasiert (10 Sekunden-Intervalle) kategorisiert.

In einem ersten Schritt erfolgt die Beschreibung und Begründung der am Kategoriensystem von von Aufschnaiter und Rogge (ebenda) vorgenommenen Modifikationen. Während von Aufschnaiter und Rogge (2010, S. 99-101) natürliche Kommunikationsprozesse in Kleingruppen analysieren, wurden in Studie C unnatürliche Kommunikationsprozesse (begleitendes Think-Aloud) analysiert. Um diese beiden Arten von Kommunikationsprozessen gegeneinander abzugrenzen, wird bei den Beschreibungen der Kategorien in dieser Arbeit anstelle des Begriffs *Vorgehen* der Begriff *Überlegungen* verwendet (vgl. Abbildung 15.1). Die Überlegungen werden aber immer vor dem Hintergrund konkreter Handlungen bewertet. Beim begleitenden Think-Aloud werden die Schülerinnen und Schüler dazu aufgefordert, ihre Überlegungen laut auszusprechen (vgl. Abschnitt 8.3.1 auf Seite 84: Zur Methode der Analyse kognitiver Prozesse), sodass alle Verbalisierungen als explizit aufgefasst werden können. Aus diesem Grund wird der Begriff *explizit* durch den Begriff *generalisierend* ersetzt. Da aufgrund der Testkonzeption (vgl. Abschnitt 3.1 auf Seite 29) nur wenig generalisierende Überlegungen erwartet wurden, ist auf eine weitere Differenzierung der generalisierend regelbasierten Überlegungen verzichtet worden.

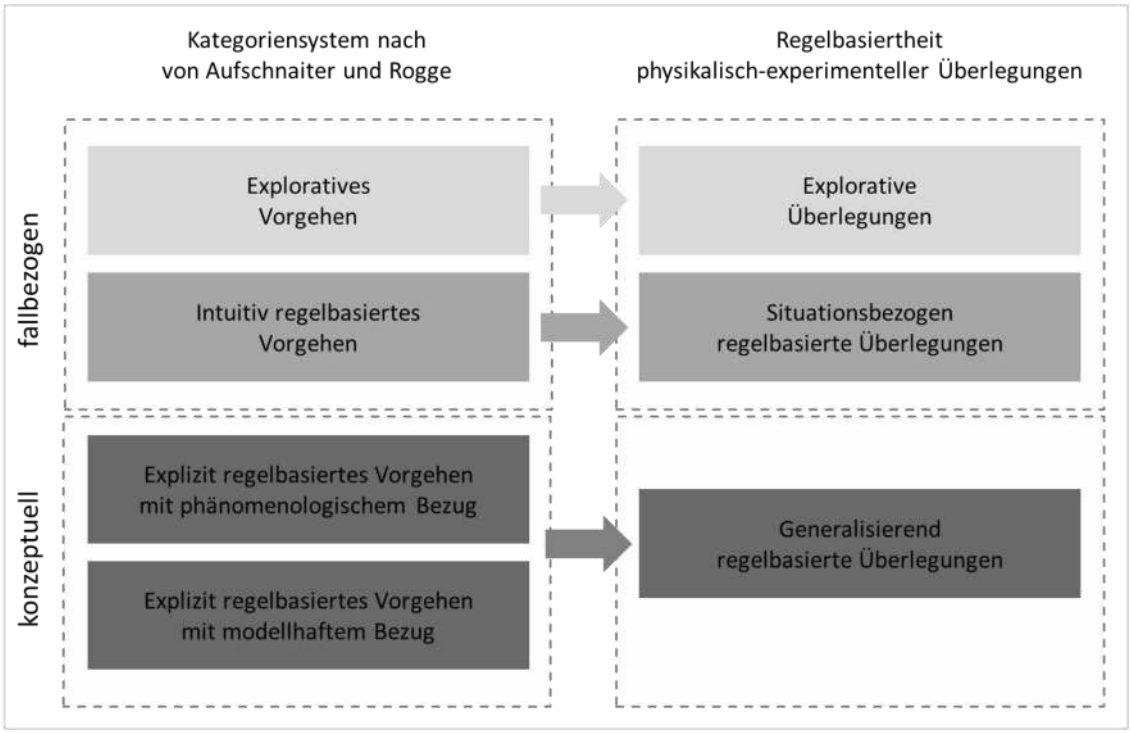


Abbildung 15.1: Vergleich des Kategoriensystems nach von Aufschnaiter und Rogge (2010, S. 103) und den adaptierten Kategorien zur Regelbasiertheit physikalisch-experimenteller Überlegungen

Tabelle 15.1 auf Seite 131 zeigt die Beschreibungen der Kategorien *explorative Überlegungen*, *situationsbezogen regelbasierte Überlegungen* und *generalisierend regelbasierte Überlegungen* nach Zirwes (2014). Diese Kategorien sind im Kodiermanual zur Analyse der Regelbasiertheit physikalisch-experimenteller Überlegungen für jeden Teilaufgabentyp ausdifferenziert und mit Beispielen veranschaulicht.

Tabelle 15.1: Kategorien zur Unterscheidung der Regelbasiertheit *physikalisch-experimenteller Überlegungen nach Zirwes (2014)*

Kategorien	Beschreibung: Überlegungen…	Beispiele für Überlegungen aus Studie C
explorative Überlegungen	sind gekennzeichnet durch eine fehlende Sicherheit bzw. eine fehlende Systematik.	*„Passt das Teil hier dran? Ne aber vielleicht das hier?“*
situationsbezogen regelbasierte Überlegungen	beziehen sich konkret auf die vorliegende (Experimentier-) Situation. Vorhandenes Wissen wird bezogen auf diese Situation angewendet.	*„Die Zeit muss ich hier nicht messen, daher brauche ich keine Stoppuhr.“*
generalisierend regelbasierte Überlegungen	beziehen sich nicht auf die konkret vorliegende (Experimentier-) Situation. Vorhandenes Wissen wird oder Erkenntnisse aus dem konkreten Experiment werden generalisiert.	*„Um eine Spannung zu messen, muss das Messgerät auf Volt eingestellt sein.“*

Für die Güte des Kategorisierungsverfahrens ergeben sich für die Datensätze mindestens akzeptable Werte ($.54 < \kappa < .84$; mittleres $\kappa = .70$). Insgesamt sind 5 % aller 10 Sekunden-Intervalle (772 von 15896) aus den in der Analyse berücksichtigten Datensätzen doppelt kodiert worden. Da der Anteil generalisierend regelbasierter Überlegungen sehr gering ist (weniger als 1 %), werden die regelbasierten Überlegungen (situationsbezogen regelbasiert, generalisierend regelbasiert) für die folgende Analyse zusammengefasst.

15.1.2 Qualität der Bearbeitungen von Teilaufgaben

Die Bewertung der Qualität der 567 Bearbeitungen von Teilaufgaben aus Studie C erfolgte auf Basis von Bildschirmaufzeichnungen mit dem MeK-LSA Bewertungsmaßstab (vgl. Abschnitt 3.3 auf Seite 34). Der Bewertungsmaßstab umfasst je nach Teilaufgabentyp bis zu drei Bewertungsstufen: 0 (ungeeignet), 1 (teilweise geeignet) und 2 (geeignet). Die Bearbeitungen von Teilaufgaben wurden durch zwei geschulte Kodierer bewertet. Die Übereinstimmung der Kodierer kann als gut bezeichnet werden, da es nur bei knapp 8 % der Bearbeitungen zu abweichenden Bewertungen kam.[24] Um Urteilsfehler bei der Bewertung zu minimieren, wurde bei den abweichenden Bewertungen eine Konsenskodierung durchgeführt. Stichprobenartig wurden im Rahmen der Konsenskodierung auch übereinstimmende Bewertungen auf Urteilsfehler überprüft. Durch diese Vorgehensweise wird sichergestellt, dass Urteilsfehler keinen bedeutsamen Einfluss auf die Zuweisung der Bewertungsstufen haben.

[24]Die Doppelkodierung der Daten erfolgte getrennt nach (Teil-)Aufgaben. Insgesamt werden im Bewertungsmaßstab nur drei Stufen unterschieden. Darüber hinaus ist die Grundhäufigkeit einzelner Bewertungsstufen, die durch einen Kodierer vergeben werden, bei einigen (Teil-)Aufgaben sehr gering. Bei dieser Datenlage ist Cohens Kappa - entgegen üblicher Praxis - kein inhaltlich sinnvoll zu interpretierendes Maß für die Urteilerübereinstimmung, da die tatsächliche Übereinstimmung zu stark unterschätzt wird.

15.1.3 Vorgehensweise

Um Annahme III.I (*Der Bewertungsmaßstab spiegelt die beobachtete experimentelle Performanz wider*) zu prüfen, wird die *Regelbasiertheit* der Überlegungen (vgl. Abschnitt 15.1.1) folgendermaßen mit der Qualität der Bearbeitungen von Teilaufgaben (vgl. Abschnitt 15.1.2) in Beziehung gesetzt:

1. Ermittlung der Besetzung der drei Bewertungsstufen (2: geeignet, 1: teilweise geeignet, 0: ungeeignet) mit Bearbeitungen von Teilaufgaben. So erhält man pro Bewertungsstufe eine Teilmenge der Bearbeitungen.
2. Prüfung, ob - jeweils im Vergleich zu den anderen beiden Bewertungsstufen - Bearbeitungen mit Bewertungsstufe 0 im Mittel der geringsten Anteil regelbasierter Überlegungen und Bearbeitungen mit Bewertungsstufe 2 im Mittel den höchsten Anteil regelbasierter Überlegungen aufweisen.
3. Berechnung einer einfaktoriellen Varianzanalyse (abhängige Variable: Anteil regelbasierter Überlegungen; Faktor: Bewertungsstufe) um zu prüfen, ob sich die drei ermittelten Teilmengen im mittleren prozentualen Anteil regelbasierter Überlegungen unterscheiden.

15.1.4 Ergebnisse

Tabelle 15.2 zeigt die Besetzung der Bewertungsstufen mit Bearbeitungen von Teilaufgaben sowie die mittleren prozentualen Anteile und den Median regelbasierter Überlegungen getrennt nach Bewertungsstufen. Es zeigt sich, dass Bewertungsstufe 0 mit 77 Datensätzen am geringsten und Bewertungsstufe 1 mit 262 Datensätzen am höchsten besetzt ist. Die 77 Datensätze mit Bewertungsstufe 0 haben rein deskriptiv den geringsten mittleren Anteil regelbasierter Überlegungen (42 %), die 228 Datensätze mit Bewertungsstufe 2 den höchsten mittleren Anteil regelbasierter Überlegungen (54 %).

Tabelle 15.2: *Besetzung der Bewertungsstufen mit Bearbeitungen von Teilaufgaben sowie mittlerer prozentualer Anteil regelbasierter Überlegungen und Median regelbasierter Überlegungen getrennt nach Bewertungsstufen*

	Bewertungsstufen		
	0	1	2
Besetzung der Bewertungsstufen mit Bearbeitungen von Teilaufgaben	77	262	228
mittlerer prozentualer Anteil regelbasierter Überlegungen: MW (SD) in %	42 (24)	53 (23)	54 (21)
MEDIAN in %	43	52	56

Eine einfaktorielle Varianzanalyse (abhängige Variable: Anteil regelbasierter Überlegungen; Faktor: Bewertungsstufe) zeigt, dass sich der mittlere Anteil regelbasierter Überlegungen zwischen den Bewertungsstufen signifikant ($F(2,564)=8{,}753$; $p<.001$; $\eta_p^2 = .030$) unterscheidet. Post-Hoc Tests nach Bonferroni zeigen, dass sich die Bewertungsstufen 0 und 1 ($p = .001$) und die Bewertungsstufen 0 und 2 signifikant ($p < .001$) unterscheiden. Die Bewertungsstufen 1 und 2 unterscheiden sich jedoch nicht signifikant voneinander ($p=1.000$).

15.2 Diskussion

Die Ergebnisse zum Zusammenhang zwischen der Regelbasiertheit experimenteller Überlegungen und der Qualität der Bearbeitungen von Teilaufgaben zeigen (vgl. Abschnitt 15.1.4), dass bei ungeeigneten Lösungen im Mittel ein signifikant niedrigerer Anteil regelbasierter Überlegungen vorliegt als bei teilweise geeigneten und geeigneten Lösungen. Das spricht dafür, dass der Bewertungsmaßstab zwischen ungeeigneten und mindestens teilweise geeigneten Lösungen im Hinblick auf die Regelbasiertheit der Überlegungen angemessen differenziert. Betrachtet man den Median regelbasierter Überlegungen, so differenziert der Bewertungsmaßstab rein deskriptiv sogar auch zwischen teilweise geeigneten und geeigneten Lösungen in intendierter Art und Weise.

Allerdings ergibt sich kein bedeutsamer Unterschied im mittleren Anteil regelbasierter Überlegungen zwischen teilweise geeigneten und geeigneten Lösungen. Ein plausibler Grund ist, dass die Regelbasiertheit eine Schwelle darstellt, die überschritten werden muss, um Bewertungsstufe 1 zu erreichen. Danach ist die fachliche Korrektheit entscheidend, die explizit auch im Bewertungsmaßstab berücksichtigt wird. Darüber hinaus ist zu berücksichtigen, dass sich die Ergebnisse auf mittlere Anteile regelbasierter Überlegungen beziehen und folglich keine Aussage über die Abfolge der Überlegungen möglich ist. Bei der Bearbeitung von experimentellen Aufgabenstellungen gibt es allerdings häufig Schlüsselstellen, die entscheidend für den Bearbeitungserfolg sind (z. B. Festlegung des Messbereichs). Möglicherweise spielt daher gerade bei der Unterscheidung von teilweise geeigneten und geeigneten Lösungen die Abfolge regelbasierter Überlegungen eine entscheidendere Rolle als die Quantität regelbasierter Überlegungen. Der nicht vorhandene Unterschied im mittleren Anteil regelbasierter Überlegungen zwischen teilweise geeigneten und geeigneten Lösungen stellt folglich keine Bedrohung für das Beibehalten von Annahme III.I (*Der Bewertungsmaßstab spiegelt die beobachtete experimentelle Performanz wider)* dar.

16 Vergleich von Testwerten (Annahme III.II)

In diesem Kapitel wird die folgende Annahme geprüft und diskutiert:

Bearbeitungen von on-screen Aufgaben werden ähnliche Testwerte zugewiesen wie inhaltlich identischen hands-on Aufgaben (AIII.II).

In den Studien D und E wurden on-screen Aufgaben und inhaltlich identische hands-on Aufgaben eingesetzt und bearbeitet (vgl. Abschnitte 8.4 und 8.5 auf Seite 86 bzw. 89). Zur Prüfung von Annahme III.II werden diesen Aufgabenbearbeitungen Testwerte durch den MeK-LSA Bewertungsmaßstab zugewiesen. Auf diese Weise kann der Frage nachgegangen werden, ob Personen bei der Bearbeitung von on-screen gestellten Aufgaben ähnliche Leistungen erzielen wie bei der Bearbeitung inhaltlich gleicher hands-on Aufgaben (gemessen an den jeweiligen Testwerten).

16.1 Vergleich von on-screen und hands-on Aufgabenbearbeitungen (Beitrag aus Studie D)

Der genaue Ablauf von Studie D ist in Abschnitt 8.4 auf Seite 86 beschrieben. An der Studie haben insgesamt 42 Biologiestudierende der RWTH-Aachen teilgenommen. Zur Prüfung von Annahme III.II (*Bearbeitungen von on-screen Aufgaben werden ähnliche Testwerte zugewiesen wie inhaltlich identischen hands-on Aufgaben)* stehen jeweils zwischen 25 und 29 hands-on und on-screen Bearbeitungen für insgesamt 15 Teilaufgaben zur Verfügung. Es handelt sich um die Teilaufgaben vom Typ *Versuchsplan entwerfen*, *Versuch aufbauen und testen*, sowie *Messung durchführen und dokumentieren* bei den Aufgaben *Ausdehnung eines Gummibandes*, *Reihenschaltung von Glühlampen*, *U-I-Kennlinie einer Glühlampe*, *Brechung am Halbkreisblock*, *Brennweitenbestimmung einer Linse*. Insgesamt konnten 420 Bearbeitungen von Teilaufgaben pro Format (on-screen und hands-on) bewertet werden. Die Bewertung der Bearbeitungen erfolgte auf Basis von Bildschirmaufzeichnungen (on-screen) bzw. Videoaufzeichnungen (hands-on) mit dem MeK-LSA Bewertungsmaßstab (vgl. Abschnitt 3.3 auf Seite 34). Alle Bearbeitungen wurden durch zwei geschulte Kodierer bewertet. Die Übereinstimmung der Kodierer kann als sehr gut bezeichnet werden, da es nur in knapp 3 % der Bearbeitungen von Teilaufgaben (23 von insgesamt 840) zu abweichenden Bewertungen kam.[25] Um Urteilsfehler bei der Bewertung zu minimieren, wurde bei den 23 abweichenden Bewertungen eine Konsenskodierung durchgeführt. Stichprobenartig wurden im Rahmen der Konsenskodierung auch übereinstimmende Bewertungen auf Urteilsfehler überprüft. Durch diese Vorgehensweise wird sichergestellt, dass Urteilsfehler keinen bedeutsamen Einfluss auf die

[25] Die Doppelkodierung der Daten erfolgte getrennt nach (Teil-)Aufgaben. Insgesamt werden im Bewertungsmaßstab nur drei Stufen unterschieden. Darüber hinaus ist die Grundhäufigkeit einzelner Bewertungsstufen, die durch einen Kodierer vergeben werden, bei einigen (Teil-)Aufgaben sehr gering. Bei dieser Datenlage ist Cohens Kappa - entgegen üblicher Praxis - kein inhaltlich sinnvoll zu interpretierendes Maß für die Urteilerübereinstimmung, da die tatsächliche Übereinstimmung zu stark unterschätzt wird.

Zuweisung der Bewertungsstufen haben.[26] Zur Untersuchung von Zusammenhängen zwischen den Testwerten in beiden Formaten werden Rangkorrelationen (Kendall's tau-b) berechnet. Um inhaltlich interpretierbare Rangkorrelationen zwischen den Testwerten in beiden Formaten über alle Aufgaben eines Teilaufgabentyps zu berechnen, müssen die Aufgaben eines Teilaufgabentyps für jedes Format – als notwendige Voraussetzung – eine reliable Skala bilden. Das ist bei den vorliegenden Daten in beiden Formaten nicht der Fall. Aufgrund des Studiendesigns (nur zwei bearbeitete Teilaufgabentypen pro Person) ist eine Skalenbildung über alle Teilaufgabentypen eines Formats ebenfalls nicht möglich. Im Folgenden werden daher für jeden Teilaufgabentyp jeder Aufgabe Rangkorrelationen berechnet, da diese sich inhaltlich sinnvoll interpretieren lassen. Bei dieser Vorgehensweise besteht allerdings die Gefahr einer Alphafehlerkumulation. Aus diesem Grund werden die Signifikanzniveaus (1 % und 5 %) für die Teilaufgabentypen mittels der konservativen Bonferroni-Korrektur ($\alpha_{k,0,01} = \frac{0,01}{5} = 0,002$ bzw. $\alpha_{k,0,05} = \frac{0,05}{5} = 0,01$) angepasst. Tabelle 16.1 zeigt die Ergebnisse der Korrelationsanalysen.

Tabelle 16.1: Korrelationen zwischen den Testwerten in den on-screen und hands-on Aufgaben (** signifikant auf dem korrigierten 1%-Niveau; * signifikant auf dem korrigierten 5%-Niveau; n.s.= keine signifikante Korrelation)

Aufgabe	Teilaufgabentyp		
	Versuchsplan entwerfen	Versuch aufbauen und testen	Messung durchführen und dokumentieren
Ausdehnung eines Gummibandes	.63**	n.s.	.62*
Reihenschaltung von Glühlampen	.69**	.86**	n.s.
U-I-Kennlinie einer Glühlampe	.66**	n.s.	n.s.
Brechung am Halbkreisblock	.85**	n.s.	.66**
Brennweitenbestimmung einer Linse	.56*	n.s.	n.s.

Für den Teilaufgabentyp *Versuchsplan entwerfen* zeigen sich über alle Aufgaben hinweg bedeutsame Zusammenhänge zwischen den Testwerten in beiden Formaten. Für den Teilaufgabentyp *Versuch aufbauen und testen* zeigt sich dagegen nur für die Aufgabe Reihenschaltung von Glühlampen ein bedeutsamer Zusammenhang. Für den Teilaufgabentyp *Messung durchführen und dokumentieren* zeigen sich für die Aufgaben *Ausdehnung eines Gummibandes* und *Brechung am Halbkreisblock* bedeutsame Zusammenhänge. Für sieben Teilaufgaben zeigen sich nicht signifikante Korrelationen (n.s.; vgl. Tabelle 16.1).

Zur Untersuchung eines möglichen Reihenfolgeneffekts ist ein Vergleich der Korrelationen getrennt nach der Bearbeitungsreihenfolge mit den vorliegenden Daten nicht sinnvoll möglich, da die Teilgruppen (12 bis 15 Testteilnehmende) für aussagefähige Korrelationsanalysen zu klein sind. Qualitative Analysen (Vergleich der absoluten und prozentualen Übereinstimmungen der Testwerte in beiden Formaten, getrennt nach der

[26] Diese Vorgehensweise war besonders wichtig, weil der Bewertungsmaßstab nur eine relativ geringe Anzahl von drei Stufen hat und Kodierfehler somit schnell eine vorhandene Korrelation verringern oder eine nicht vorhandene vortäuschen könnten.

Bearbeitungsreihenfolge) liefern aber zumindest keine Hinweise, die auf einen Reihenfolgeneffekt hindeuten.

16.2 Testformatvergleich zum Anfertigen eines Messwertediagramms (Beitrag aus Studie E)

Der genaue Ablauf der Studie ist in Abschnitt 8.5 auf Seite 89 beschrieben. An der Studie haben 19 Schülerinnen und Schüler einer 9. Gymnasialklasse aus Nordrhein-Westfalen teilgenommen. Zur Prüfung von Annahme III.II (*Bearbeitungen von on-screen Aufgaben werden ähnliche Testwerte zugewiesen wie inhaltlich identischen hands-on Aufgaben*) stehen jeweils 19 hands-on und on-screen Bearbeitungen zum *Anfertigen eines Messwertediagramms* der Aufgabe *Spielzeugauto auf einer Rampe* zur Verfügung. Die Bewertung der Bearbeitungen erfolgte auf Basis von Screenshots (on-screen) bzw. Protokollbögen (hands-on) mit dem MeK-LSA Bewertungsmaßstab (vgl. Abschnitt 3.3 auf Seite 34). Zur Untersuchung von Zusammenhängen zwischen den Testwerten in beiden Formaten werden Rangkorrelationen (Kendall's tau-b) berechnet. Für die Teilaufgabe zum Anfertigen eines Messwertediagramms zeigt sich eine nicht signifikante Korrelation, wobei nur neun von 19 Schülerinnen und Schülern in beiden Formaten den gleichen Testwert erreicht haben.

16.3 Diskussion

Die Ergebnisse aus Studie D (vgl. Abschnitt 16.1) zeigen für acht von 15 Teilaufgaben signifikante Zusammenhänge zwischen den Testwerten in beiden Formaten. Die signifikanten Zusammenhänge können dabei sogar als hoch bezeichnet werden: Zum einen liegen sie in der Regel noch über den Werten, die in bisherigen Studien zur Austauschbarkeit von on-screen und hands-on Experimenten gefunden wurden (Schreiber et al., 2014; Shavelson et al., 1999). Zum anderen handelt es sich um manifeste Korrelationen auf Ebene einzelner Teilaufgaben, sodass die tatsächliche Höhe der Zusammenhänge eher unterschätzt wird. Für sieben von 15 Teilaufgaben zeigen sich allerdings keine signifikanten Korrelationen zwischen beiden Formaten. Für die nicht signifikanten Korrelationen kann es im Wesentlichen zwei mögliche Gründe geben:

1. Die Testteilnehmenden erhalten bei identischen Aufgaben in beiden Formaten unterschiedliche Testwertzuweisungen.

2. Die Testteilnehmenden erreichen bei inhaltlich identischen Aufgaben in beiden Formaten vergleichbare Testwerte, aber es liegt insgesamt wenig Varianz im Leistungsspektrum vor.

Beide Gründe sind nicht wünschenswert, allerdings spricht nur der erste Grund gegen die Vergleichbarkeit der Formate. Schaut man sich die Verteilung der Testwerte in Form von Blasendiagrammen genauer an, so liegt die Ursache für die nicht signifikanten Korrelationen fast immer in der zu geringen Varianz im Leistungsspektrum (vgl. beispielhaft Abbildung 16.1 auf Seite 138), wodurch sich hohe prozentuale Übereinstimmungen ergeben (vgl. Tabelle 16.2 auf Seite 138).

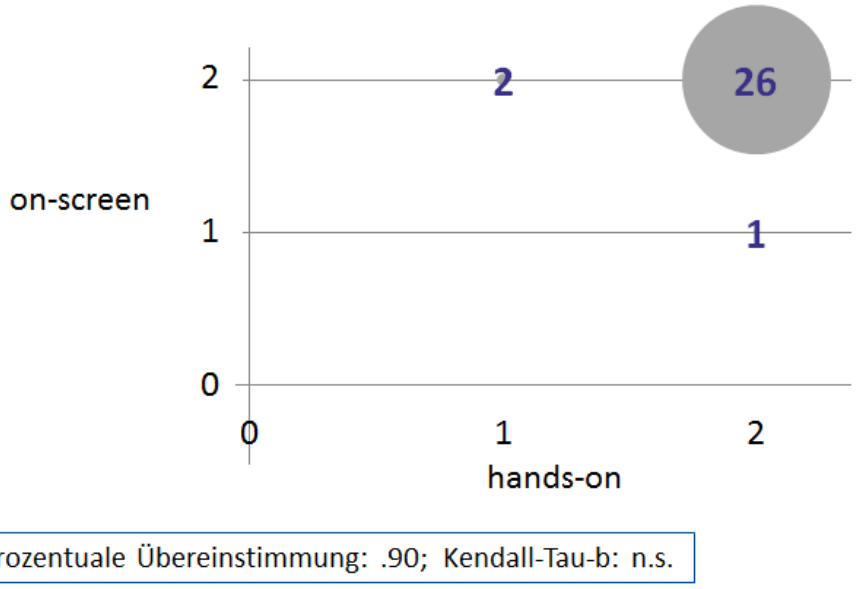


Abbildung 16.1: Blasendiagramm zur Verteilung der Testwerte in beiden Formaten (Aufgabe: Brennweitenbestimmung einer Linse*; Teilaufgabe:* Versuch aufbauen und testen*)*

Tabelle 16.2: Prozentuale Übereinstimmung zwischen den Testwerten in den on-screen und hands-on Aufgaben für die nicht signifikanten Korrelationen(s. K.: signifikante Korrelation)

Aufgabe	Teilaufgabentyp	
	Versuch aufbauen und testen	Messung durchführen und dokumentieren
Ausdehnung eines Gummibandes	.76	s. K.
Reihenschaltung von Glühlampen	s. K.	.88
U-I-Kennlinie einer Glühlampe	.79	.85
Brechung am Halbkreisblock	.79	s. K.
Brennweitenbestimmung einer Linse	.90	.54

Eine Ausnahme bildet die Teilaufgabe *Messung durchführen und dokumentieren* der Aufgabe *Brennweitenbestimmung einer Linse*. Bei dieser Teilaufgabe erhalten zwölf von 26 Testteilnehmenden unterschiedliche Testwertzuweisungen (vgl. Abbildung 16.2 auf Seite 139). Das deutet für diese Teilaufgabe auf unterschiedliche Bearbeitungsprozesse in beiden Formaten hin. Für die zu prüfende Annahme ist jedoch die Zuweisung der Testwerte und nicht die Gleich- oder Ungleichheit von Bearbeitungsprozessen entscheidend.

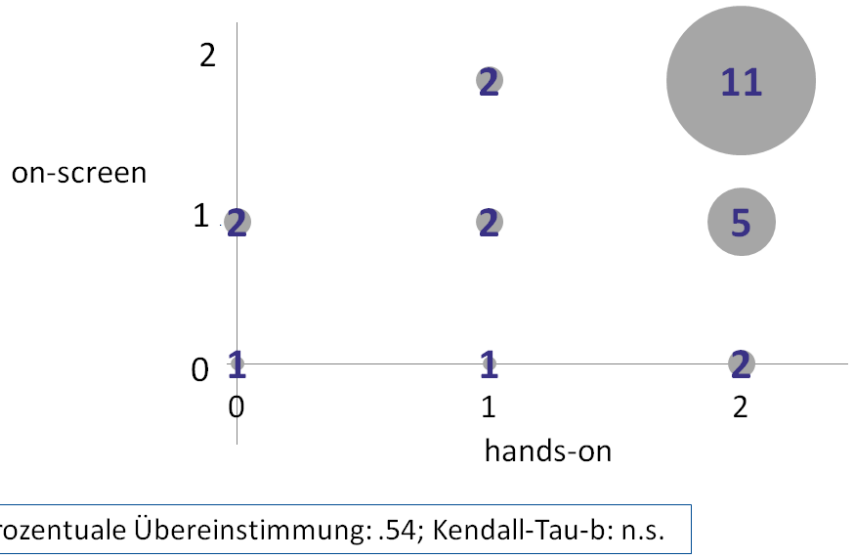


Abbildung 16.2: Blasendiagramm zur Verteilung der Testwerte in beiden Formaten (Aufgabe: Brennweitenbestimmung einer Linse*; Teilaufgabe:* Messungen durchführen und dokumentieren*)*

Insbesondere die Teilaufgaben zum *Versuch aufbauen und testen* waren für die Studierenden offenbar deutlich zu leicht, sodass fast durchgängig Stufe 2 (geeignet) erreicht wurde. Folglich lassen sich für diese Teilfähigkeiten keine statistisch bedeutsamen Zusammenhänge in Form von Rangkorrelationen nachweisen. Andererseits finden sich auch keine schlüssigen Hinweise die *gegen* eine Äquivalenz der Leistungen sprechen. Ganz im Gegenteil: Die hohen prozentualen Übereinstimmungen (vgl. Tabelle 16.2 auf Seite 138) bei geringer Varianz im Leistungsspektrum sprechen zumindest für eine Äquivalenz am oberen Ende des Leistungsspektrums.

Die größte Einschränkung ergibt sich aufgrund der gewählten Stichprobe in Studie D (Biologiestudierende). Auch wenn die Studierenden breit gestreute, aber überwiegend mit der Zielgruppe des MeK-LSA Experimentiertests vergleichbare Vorerfahrungen mit physikalischen Experimenten besitzen (vgl. Abschnitt 8.4.2 auf Seite 89), kann aus den Ergebnissen nicht unmittelbar auf die Leistungen in der Zielgruppe (Schülerinnen und Schüler am Ende der Sekundarstufe I) geschlossen werden. Die Ergebnisse für die Biologiestudierenden liefern jedoch keine Hinweise, die *gegen* das Beibehalten von Annahme III.II (*Bearbeitungen von on-screen Aufgaben werden ähnliche Testwerte zugewiesen wie inhaltlich identischen hands-on Aufgaben)* sprechen.

Beim Testformatvergleich zum *Anfertigen eines Messwertediagramms* zeigt sich keine bedeutsame Korrelation zwischen den Formaten (vgl. Abschnitt 16.2 und Abbildung 16.3 auf Seite 140). Der Grund dafür ist, dass zehn von 19 Schülerinnen und Schüler in beiden Formaten unterschiedliche Testwertzuweisungen erhalten. In einer Detailanalyse zu den aufgetretenen Schwierigkeiten beim Anfertigen des Messwertediagramms findet Jansen (2014) einerseits, dass die Schülerinnen und Schüler unabhängig vom Format in der Lage sind, eine angemessene Achsenbelegung und eine angemessene Achsenskalierung vorzunehmen. Andererseits scheint den Schülerinnen und Schülern das genaue Eintragen der Messwerte und eine angemessene Achsenbeschriftung im on-screen Format schwerer zu fallen als im hands-on Format.

Da sowohl das genaue Eintragen der Messwerte als auch die Achsenbeschriftung im MeK-LSA Bewertungsmaßstab berücksichtigt werden, und die untersuchte Stichprobe mit 19 Schülerinnen und Schüler eher klein ist, ist es plausibel erklärbar, dass sich keine bedeutsame Korrelation zwischen den Formaten ergibt.

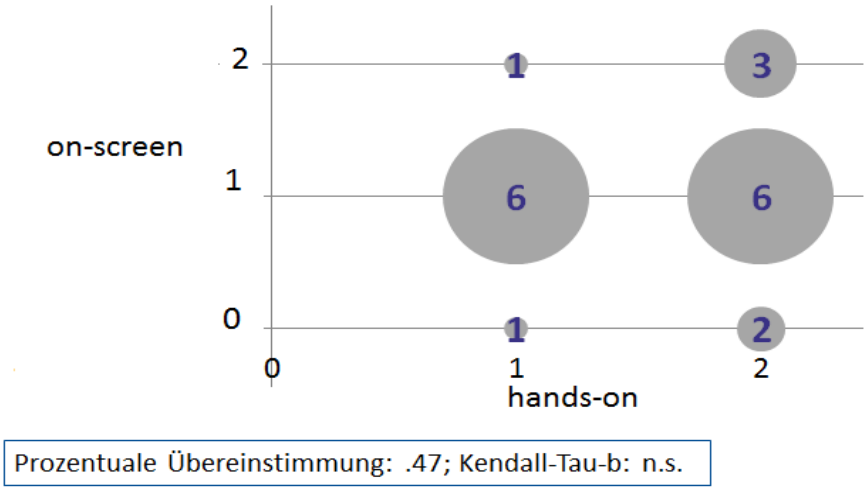


*Abbildung 16.3: Blasendiagramm zur Verteilung der Testwerte in beiden Formaten (*Anfertigen eines Messwertediagramms*)*

Somit sprechen die Ergebnisse auf der einen Seite gegen die Vergleichbarkeit der Formate beim *Anfertigen eines Messwertediagramms*. Auf der anderen Seite ist zu bedenken, dass die aufgetretenen Schwierigkeiten mit hoher Wahrscheinlichkeit auf Probleme bei der Bedienung einzelner Funktionen des Diagramm-Tools im on-screen Format (z. B. Bedienung des Buttons *Beschriftung*) zurückzuführen sind. Die unterschiedlichen Testwerte, trotz inhaltlich identischer Aufgabenstellungen zum Anfertigen eines Messwertediagramms, sind daher mit hoher Wahrscheinlichkeit keine Folge einer komplett unterschiedlichen Vorgehensweise oder eines ungeeigneten Bewertungsmaßstabs. Die Ergebnisse zeigen vielmehr die Notwendigkeit, Hilfen für das Eintragen von Messwerten und die Achsenbeschriftung zur Verfügung zu stellen, sprechen aber nicht generell gegen den Einsatz dieses Aufgabentyps im on-screen Format.

17 Bewertung des Interpretations-Nutzungs-Arguments (Validitätsargumentation)

In diesem Kapitel wird für den MeK-LSA Experimentiertest bewertet, bis zu welchem Grad die folgende Testwertinterpretation beibehalten werden kann:

Die Testwerte können valide als Ausdruck von Experimentierfähigkeiten aufgefasst werden.

Den Bezugsrahmen für die Validitätsbewertung des MeK-LSA Experimentiertests bildet das in Kapitel 6 (Seiten 67-75) beschriebene Interpretations-Nutzungs-Argument (INA).

Um die Zieldomäne Experimentieren im Physikunterricht der Sekundarstufe I im MeK-LSA Experimentiertest abzubilden, sind drei aufeinander aufbauende Schritte durchgeführt worden (vgl. Abbildung 17.1).

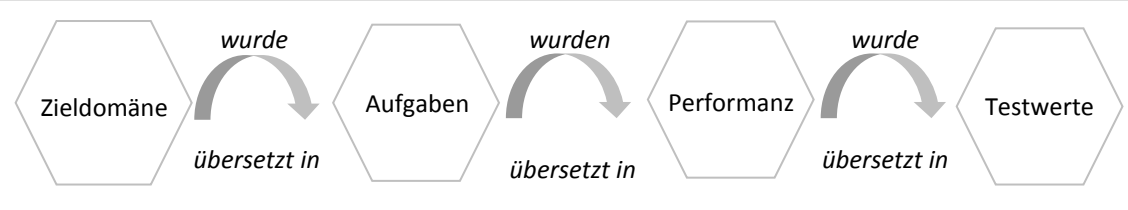


Abbildung 17.1: Durchgeführte Schritte von der Zieldomäne Experimentieren im Physikunterricht der Sekundarstufe I *bis zur Zuweisung von Testwerten*

Die von den drei Schritten zu erfüllenden Qualitätsanforderungen sind im INA durch die folgenden übergeordneten Aussagen beschrieben:

1. Die Aufgaben umfassen relevante und repräsentative Inhalte und Anforderungen aus der Zieldomäne (Teil I des INA).
2. Die beobachtete Performanz passt zur beabsichtigten Performanz (Teil II des INA).
3. Die beobachtete Performanz wird in geeigneter Art und Weise in Testwerte überführt (Teil III des INA).

Diese drei übergeordneten Aussagen sind im INA in Form weiterer Annahmen konkretisiert worden (vgl. Kapitel 6 auf den Seiten 67-75). Die evidenzbasierte Prüfung der Annahmen erfolgte in den Kapiteln 9 bis 16. Auf Grundlage dieser Prüfungen wird in den Abschnitten 17.1 bis 17.3 bewertet, bis zu welchem Grad die Aussagen bestätigt werden können. Eine Aussage gilt als bestätigt, wenn alle zu dieser Aussage gehörenden Annahmen beibehalten werden können. Bei der Bewertung wird explizit auch berücksichtigt, inwieweit sich bei der Prüfung Einschränkungen für das Beibehalten der Annahmen ergeben haben. Dabei lassen sich im Wesentlichen zwei Arten von Einschränkungen unterscheiden:

1. *Bedrohende Einschränkungen* ergeben sich aufgrund von Evidenz, die gegen das Beibehalten einer Annahme sprechen und in der Folge die Validitätsargumentation schwächen (z. B.: generierte Evidenz spricht gegen die Vergleichbarkeit von Testwerten zwischen on-screen und hands-on Format beim Erstellen eines Messwertediagramms).

2. *Nicht-bedrohende Einschränkungen* ergeben sich aufgrund einer eingeschränkten Datenbasis. Die Beschränkung der Analysen auf ausgewählte (Teil-)Aufgaben, die Untersuchung von (kleinen) Gelegenheitsstichproben sowie der Detailgrad der Datenerhebung und der Analyseverfahren sind Beispiele für eine eingeschränkte Datenbasis. Solche Beschränkungen waren aufgrund des bereits sehr hohen Aufwands für die Prüfung der Annahmen nicht zu vermeiden. Die auf dieser Datenbasis generierte Evidenz spricht aber im Allgemeinen nicht gegen das Beibehalten einer Annahme, weil wesentliche und besonders kritische Aspekte des Testverfahrens bei den Prüfungen durchgängig berücksichtigt worden sind (vgl. Abschnitt 7.2 auf Seite 80). Folglich schwächen die nicht-bedrohenden Einschränkungen die Validitätsargumentation nicht bedeutsam.

17.1 Bewertung der ersten Aussage: Die Aufgaben umfassen relevante und repräsentative Inhalte und Anforderungen aus der Zieldomäne (Teil I des INA)

In diesem Abschnitt wird bewertet, inwieweit die übergeordnete Aussage zum ersten Schritt (von der Zieldomäne zu den Aufgaben) bestätigt wird (vgl. Tabelle 17.1). Die Aussage gilt als bestätigt, wenn die zur Aussage gehörigen Annahmen, auf Basis der Prüfung in den Kapiteln 9 bis 10, ohne bedrohende Einschränkungen beibehalten werden können.

Tabelle 17.1: Bewertung der ersten übergeordneten Aussage im INA

Die **Aufgaben** umfassen relevante und repräsentative Inhalte und Anforderungen aus der Zieldomäne (Teil 1 des INA)			
Annahme	**Annahme beibehalten?**	**Einschränkungen**	
		bedrohend	**nicht-bedrohend**
I.I: Die Aufgaben basieren auf Experimenten aus relevanten Inhaltsbereichen des Physikunterrichts der Sekundarstufe I.	✓	-	ausgewählte Aufgaben; Gelegenheitsstichprobe(n); Detailgrad der Datenerhebung (z. B. Aufgabenskizzen)
I.II: Die Aufgaben stellen experimentelle Anforderungen, die Schülerinnen und Schüler aus ihrem Physikunterricht der Sekundarstufe I üblicherweise kennen.	✓	-	

Die Ergebnisse der Lehrkräftebefragung zur Bekanntheit der Experimente aus den Aufgabenskizzen, die Ergebnisse der Schülerbefragung zur Bekanntheit der Experimente aus den Testaufgaben sowie die Ergebnisse einer externen Lehrkräftebefragung zum Einsatz von Experimenten im eigenen Physikunterricht (Karaböcek & Erb, 2015) zeigen, dass Annahme I.I trotz Berücksichtigung von normativen Entscheidungen des Testentwicklungsteams bei der Auswahl der Testaufgaben beibehalten werden kann (vgl. Abschnitt 9.3 auf Seite 94). Dieser Befund ist erwartungskonform, da bereits die Überführung der Testkonzeption in konkrete Testaufgaben auf umfangreichen Lehrplan- und Schulbuchanalysen basiert (vgl. Kapitel 4 auf den Seiten 37-53). Einzelne Experimente weisen aus Lehrer- und Schülersicht einen etwas geringeren Bekanntheitsgrad auf (z. B. *Totalreflexion* oder *Auftriebskraft in Wasser*). Aus diesem Befund ergibt sich allerdings keine bedrohende Einschränkung. Wesentlich ist, dass die Schülerinnen und Schüler im Unterricht experimentelle Lerngelegenheiten in den relevanten

Inhaltsbereichen (Elektrizitätslehre, Optik, Mechanik) hatten, die es ihnen ermöglichen, die Aufgaben erfolgreich zu bearbeiten. Es ist daher nicht problematisch, wenn einzelne Experimente im Unterricht nicht bzw. nur vereinzelt durchgeführt worden sind.

Die in den Aufgaben gestellten experimentellen Anforderungen werden von den befragten Lehrkräften als erfüllbar eingeschätzt (vgl. Abschnitt 10.1 auf Seite 97). Die experimentellen Anforderungen sind darüber hinaus aus Sicht der befragten Schülerinnen und Schüler häufig Bestandteil des eigenen Physikunterrichts (vgl. Abschnitt 10.2 auf Seite 99). Das gilt insbesondere für die Anforderungen im Bereich der Durchführung von Experimenten.

Da einige Daten zur Prüfung der Annahmen bereits während der Entwicklung des MeK-LSA Experimentiertests erhoben (z. B. Lehrkräftebefragung) bzw. im Rahmen breiter angelegter Studien ergänzend miterfasst wurden (Large-Scale Studie: Bekanntheit der Experimente; Studie zu Aufgabenbearbeitungsprozessen: Häufigkeit experimenteller Anforderungen im Physikunterricht), ist der Detailgrad dieser Daten zum Teil eingeschränkt. So haben die Lehrkräfte die Bekanntheit der Experimente und die gestellten experimentellen Anforderungen beispielsweise auf Basis von Aufgabenskizzen und nicht auf Basis der endgültigen Testaufgaben eingeschätzt. Eine ausführliche Diskussion dieser Einschränkungen findet sich in den Abschnitten 9.3 und 10.3 auf Seite 94 bzw. 101, wobei die Ergebnisse ohne diese Einschränkungen wahrscheinlich noch positiver ausgefallen wären. Insgesamt stützen die Befunde, die sich aus der Prüfung der Annahmen I.I und I.II (vgl. Tabelle 17.1 auf Seite 142) ergeben, die erste übergeordnete Aussage (*Die* Aufgaben *umfassen relevante und repräsentative Inhalte und Anforderungen aus der Zieldomäne).* Zusammenfassend lässt sich daher festhalten:

Die Zieldomäne wurde adäquat in Testaufgaben übersetzt.

Die Argumentation zur ersten übergeordneten Aussage zeigt schlüssig auf, dass die Aufgaben des MeK-LSA Experimentiertests relevante und repräsentative Inhalte und Anforderungen der Zieldomäne abbilden. Im Gesamtzusammenhang des Assessment-Diskurses experimenteller Kompetenz (vgl. Gut, 2012) bleibt allerdings kritisch zu diskutieren, inwieweit die Aufgaben des MeK-LSA Experimentiertests die relevanten und repräsentativen Inhalte der Zieldomäne abbilden können (vgl. Abschnitt 18.1 auf Seite 151).

17.2 Bewertung der zweiten Aussage: Die beobachtete Performanz passt zur beabsichtigten Performanz (Teil II des INA)

In diesem Abschnitt wird bewertet, inwieweit die übergeordnete Aussage zum zweiten Schritt (von den Aufgaben zur Performanz) bestätigt wird (vgl. Tabelle 17.2). Die Aussage gilt als bestätigt, wenn die zur Aussage gehörigen Annahmen, auf Basis der Prüfung in den Kapiteln 11 bis 14, ohne bedrohende Einschränkungen beibehalten werden können.

Tabelle 17.2: Bewertung der zweiten übergeordneten Aussage im INA

Die beobachtete **Performanz** passt zur beabsichtigten Performanz (Teil II des INA)			
Annahme	**Annahme beibehalten?**	**Einschränkungen**	
		bedrohend	**nicht-bedrohend**
II.I: Die Schülerinnen und Schüler stellen bei der Bearbeitung der on-screen Aufgaben überwiegend experimentbezogene Überlegungen an.	✓	-	ausgewählte Aufgaben und Teilaufgabentypen; Gelegenheitsstichprobe; Detailgrad der Analyse
II.II: Die Bearbeitung des Tests ermöglicht das Zeigen experimenteller Performanz.	✓	-	kleine Gelegenheitsstichprobe(n); Detailgrad der Datenerhebung
II.III: Der Anteil experimentbezogener Überlegungen ist bei der Bearbeitung von on-screen Aufgaben vergleichbar hoch wie bei der Bearbeitung inhaltlich identischer hands-on Aufgaben.	✓	-	ausgewählte Aufgabe und Teilaufgabentypen; Gelegenheitsstichprobe; Detailgrad der Analyse
II.IV: Die wahrgenommene kognitive Belastung ist bei on-screen Aufgaben vergleichbar hoch wie bei inhaltlich identischen hands-on Aufgaben.	✓	-	ausgewählte Aufgaben und Teilaufgabentypen; Gelegenheitsstichprobe

Im Folgenden werden zunächst die Beiträge aus der Prüfung der Annahmen II.I bis II.IV (vgl. Tabelle 17.2) zur Bewertung der zweiten Aussage beschrieben. Anschließend erfolgt eine zusammenfassende Bewertung der zweiten Aussage.

Überwiegend experimentbezogene Überlegungen? (Prüfung von Annahme II.I)

Die Untersuchung der Aufgabenbearbeitungsprozesse zeigt, dass die Schülerinnen und Schüler bei der Bearbeitung der on-screen Aufgaben überwiegend (im Mittel: 60 %) experimentbezogene Überlegungen anstellen (vgl. Abschnitt 11.1.4 auf Seite 105). Es ist daher plausibel anzunehmen, dass die beobachteten Schülerhandlungen in der Mehrzahl auf experimentbezogenen Überlegungen basieren. Die verwendete Kategorie *experimentbezogene Überlegungen* stellt keine Neuentwicklung dar, sondern lehnt sich eng an ein etabliertes Kategoriensystem zur (experimentellen) Kompetenzmodellierung von von Aufschnaiter und Rogge (2010) an. Die Ankopplung an dieses Kategoriensystem erhöht die Aussagefähigkeit der Untersuchung und wirkt bewusst einem *Confirmation Bias* entgegen, der sich durch eine vollständige Neuentwicklung dieser Kategorie hätte ergeben können. Aufgrund der zugrundeliegenden Auswahlkriterien (vgl. Abschnitt 7.2 auf Seite 80) stellt die Fokussierung auf ausgewählte Aufgaben und Teilaufgabentypen keine bedrohende Einschränkung dar. Gleiches gilt für die zeitbasierte Analyse der Daten, die zwar qualitative Detailanalysen erschwert, aber

dennoch eine zufriedenstellende Beurteilung der Gedankengänge ermöglicht (vgl. Abschnitt 11.2 auf Seite 106).

Zeigen experimenteller Performanz möglich? (Prüfung von Annahme II.II)

Zur Prüfung von Annahme II.II (*Die Bearbeitung des Tests ermöglicht das Zeigen experimenteller Performanz*) sind die folgenden Indikatoren herangezogen worden: die Wahrnehmung der Testsituation durch Schülerinnen und Schüler und der Vergleich der experimentellen Vorgehensweise im konsekutiven und nicht-konsekutiven Aufgabenformat. Die Ergebnisse der Schülerbefragung zur Wahrnehmung der Testsituation (vgl. Abschnitt 12.1 auf Seite 109) zeigen, dass die vom Testentwicklungsteam getroffenen Maßnahmen (z. B. konsekutives Aufgabenformat, Trainingsaufgabe zum Kennenlernen des Aufgabenformats) von den Schülerinnen und Schülern positiv wahrgenommen werden. Die Trainingsaufgabe trägt dazu bei, die Schülerinnen und Schüler im Sinne des *familiarity approach* mit dem Aufgabenformat vertraut zu machen, ohne die Schülerinnen und Schüler gleichzeitig auf die Testinhalte vorzubereiten (vgl. Abschnitt 12.2 auf Seite 111). Die Ergebnisse aus Studie G, in der das konsekutive Aufgabenformat des MeK-LSA Experimentiertests mit einem nicht-konsekutiven Aufgabenformat verglichen wurde (vgl. Abschnitt 12.3 auf Seite 116) sprechen dafür, dass die Schülerinnen und Schüler durch das konsekutive Aufgabenformat bei der Strukturierung des Bearbeitungsprozesses unterstützt werden. Trotz vorgegebenem Bearbeitungsleitfaden ändert sich die grundsätzliche experimentelle Vorgehensweise im Vergleich zur Bearbeitung einer Aufgabe ohne Bearbeitungsleitfaden (nicht-konsekutives Format) nicht bedeutsam. Darüber hinaus werden im konsekutiven Aufgabenformat Folgefehler vermieden (Eckloff, 2014). Einschränkungen ergeben sich durch die vergleichsweise kleinen Gelegenheitsstichproben und den Detailgrad der Datenerhebung (z. B. ist ein Lerneffekt in Studie G auf Grundlage der vorhandenen Datenbasis nicht auszuschließen; vgl. Abschnitt 12.4 auf Seite 119). Diese Einschränkungen schwächen auf der einen Seite zwar die Aussagefähigkeit der einzelnen Studienergebnisse, die daher entsprechend vorsichtig zu interpretieren sind. Auf der anderen Seite liefern die Studien trotz der vorhandenen Einschränkungen keine Evidenz, die gegen das Beibehalten von Annahme II.II (*Die Bearbeitung des Tests ermöglicht das Zeigen experimenteller Performanz*) spricht. In der Gesamtschau aller Studien sprechen die Ergebnisse dafür, dass die Bearbeitung des Tests das Zeigen experimenteller Performanz ermöglicht. Es bleibt allerdings kritisch zu diskutieren, ob die Berücksichtigung weiterer Indikatoren sinnvoll ist, um die Annahme auf einer breiteren empirischen Basis zu prüfen. Ein möglicher Indikator wären sogenannte Usability-Checks, die sich auch bei großen Stichproben effizient einsetzen lassen (z. B. Usability-Checks zur Softwarebedienung, vgl. Brooke, 1996).

Anteil experimentbezogener Überlegungen im on-screen und hands-on Format vergleichbar hoch? (Prüfung von Annahme II.III)

Auf kognitiver Ebene konnte durch die Untersuchung der Aufgabenbearbeitungsprozesse gezeigt werden, dass der Anteil experimentbezogener Überlegungen bei der Bearbeitung einer on-screen Aufgabe vergleichbar hoch ist wie bei der Bearbeitung einer inhaltlich identischen

hands-on Aufgabe (vgl. Abschnitt 13.1.2 auf Seite 122). Aufgrund der zugrundeliegenden Auswahlkriterien (vgl. Abschnitt 7.2 auf Seite 80; Stichwort: Worst-Case Abschätzung) stellt die Fokussierung auf ausgewählte Teilaufgaben der Aufgabe *Ausdehnung eines Gummibandes* keine bedrohende Einschränkung dar. Gleiches gilt für die zeitbasierte Analyse der Daten, die zwar qualitative Detailanalysen erschwert, aber dennoch eine zufriedenstellende Beurteilung der Gedankengänge ermöglicht (vgl. Abschnitt 13.2 auf Seite 123). Insgesamt zeigt sich bei der Prüfung von Annahme II.III (*Der Anteil experimentbezogener Überlegungen ist bei der Bearbeitung von on-screen Aufgaben vergleichbar hoch wie bei der Bearbeitung inhaltlich identischer hands-on Aufgaben*), dass das on-screen Format auf kognitiver Ebene mit dem als Referenzmaßstab geltenden hands-on Format vergleichbar ist.

Wahrgenommene kognitive Belastung zwischen on-screen und hands-on Format vergleichbar? (Prüfung von Annahme II.IV)

Die Ergebnisse zur Einschätzung der wahrgenommenen kognitiven Belastung zeigen, dass die wahrgenommene kognitive Belastung im on-screen Format vergleichbar hoch ist wie bei inhaltlich identischen hands-on Aufgaben (vgl. Kapitel 14 auf den Seiten 125-128). In beiden Formaten liegt der Anteil deutlich unterhalb einer neutralen Einschätzung, sodass plausibel anzunehmen ist, dass formatspezifische Eigenschaften weder bei der Bearbeitung der on-screen Aufgaben noch bei der Bearbeitung der hands-on Aufgaben zu einer Überschreitung der Verarbeitungskapazität des Arbeitsgedächtnisses führen. Die größte Einschränkung ergibt sich aufgrund der gewählten Gelegenheitsstichprobe in Studie D (Biologiestudierende). Eine ausführliche Diskussion dieser Einschränkung findet sich in Abschnitt 14.4 auf Seite 127. Dabei zeigt sich jedoch, dass es trotz dieser Einschränkung keine Hinweise gibt, die gegen das Beibehalten von Annahme II.IV (*Die wahrgenommene kognitive Belastung ist bei on-screen Aufgaben vergleichbar hoch wie bei inhaltlich identischen hands-on Aufgaben*) sprechen.

Zusammenfassende Bewertung der zweiten Aussage

Auch wenn eine noch breitere empirische Absicherung wünschenswert wäre, stützen die sich aus der Prüfung der Annahmen II.I bis II.IV (vgl. Tabelle 17.2 auf Seite 144) ergebenden Befunde die zweite übergeordnete Aussage (*Die beobachtete Performanz passt zur beabsichtigten Performanz).* Zusammenfassend lässt sich daher festhalten:

Die Aufgaben wurden angemessen in Performanz übersetzt.

Um noch stärkere Evidenz für das Beibehalten (bzw. die Ablehnung) der Annahmen bzw. der Aussage zu generieren, ist es sinnvoll, einzelne Studienergebnisse durch eine Erweiterung der empirischen Datenbasis umfangreicher als bisher herauszufordern. Das betrifft beispielsweise die Ergebnisse zum Vergleich von konsekutivem und nicht-konsekutivem Aufgabenformat (vgl. Abschnitt 12.3 auf Seite 116) oder die Ergebnisse zur wahrgenommenen kognitiven Belastung. Um stärkere Evidenz für eine grundsätzlich ähnliche experimentelle Vorgehensweise im konsekutiven und nicht-konsekutiven Aufgabenformat zu generieren, sollte eine kontrollierte Vergleichsstudie mit einer größeren Stichprobe und mindestens einer Aufgabe aus jedem Themengebiet (E-Lehre, Optik, Mechanik) durchgeführt werden.

Auf Basis der vorhandenen Ergebnisse zur wahrgenommenen kognitiven Belastung (vgl. Abschnitt 14.4 auf Seite 127) ist es zwar plausibel anzunehmen, dass auch bei Schülerinnen und Schülern die Verarbeitungskapazität im Arbeitsgedächtnis bei den Teilaufgabentypen *Versuchsplan entwerfen*, *Versuch aufbauen und testen* sowie *Messungen durchführen und dokumentieren* nicht überschritten wird. Bislang fehlen zu dieser Annahme allerdings belastbare empirische Daten, sodass die Gefahr einer *reification fallacy* (Stichwort: *vorschnelle Generalisierung*; vgl. Seite 65) nicht gänzlich ausgeschlossen werden kann.

17.3 Bewertung der dritten Aussage: Die beobachtete Performanz wird in geeigneter Art und Weise in Testwerte überführt (Teil III des INA).

In diesem Abschnitt wird bewertet, inwieweit die übergeordnete Aussage zum dritten Schritt (von der Performanz zu den Testwerten) bestätigt wird (vgl. Tabelle 17.3). Die Aussage gilt als bestätigt, wenn die zur Aussage gehörigen Annahmen, auf Basis der Prüfung in den Kapiteln 15 und 16, ohne bedrohende Einschränkungen beibehalten werden können.

Tabelle 17.3: Bewertung der dritten übergeordneten Aussage im INA

Die beobachtete Performanz wird in geeigneter Art und Weise in **Testwerte** überführt. (Teil III des INA)			
Annahme	**Annahme beibehalten?**	**Einschränkungen**	
		bedrohend	**nicht-bedrohend**
III.I: Der Bewertungsmaßstab spiegelt die beobachtete experimentelle Performanz wider.	✓	-	ausgewählte Aufgaben und Teilaufgabentypen; Detailgrad der Analyse; Gelegenheitsstichprobe
III.II: Bearbeitungen von on-screen Aufgaben werden ähnliche Testwerte zugewiesen wie inhaltlich identischen hands-on Aufgaben.	(✓)	Messwertediagramm	Gelegenheitsstichprobe; ausgewählte Aufgaben und Teilaufgabentypen

Der Vergleich zwischen der Regelbasiertheit experimenteller Überlegungen während der Aufgabenbearbeitung und der Zuweisung von Testwerten auf Basis des MeK-LSA Bewertungsmaßstabs zeigt, dass ungeeignete Lösungen den geringsten Anteil und geeignete Lösungen den höchsten Anteil regelbasierter Überlegungen aufweisen (vgl. Kapitel 15 auf den Seiten 129-133). Eine Differenzierung des Anteils regelbasierter Überlegungen zwischen teilweise geeigneten und geeigneten Lösungen liegt nicht vor. Hieraus ergibt sich allerdings keine bedrohende Einschränkung, da die Regelbasiertheit vermutlich eine Schwelle darstellt, die überschritten werden muss, um Bewertungsstufe 1 zu erreichen. Die explizit im Bewertungsmaßstab berücksichtigte fachliche Korrektheit ist erst danach entscheidend. Insgesamt zeigen die Befunde zumindest, dass der Bewertungsmaßstab die beobachtete experimentelle Performanz angemessen widerspiegelt.

Beim Vergleich der Testformate (vgl. Abschnitt 16.1 auf Seite 135) konnte gezeigt werden, dass Personen bei der Bearbeitung von on-screen gestellten Aufgaben in der Regel ähnliche Leistungen erzielen wie bei der Bearbeitung inhaltlich identischer hands-on Aufgaben (gemessen an den jeweiligen Testwerten). Als Einschränkung ergibt sich allerdings, dass

Studierende anstelle von Schülerinnen und Schüler getestet worden sind. Eine ausführliche Diskussion dieser Einschränkung findet sich in Abschnitt 16.3 auf Seite 137. Dabei zeigt sich jedoch, dass es aufgrund dieser Einschränkung keine Hinweise gibt, die gegen das Beibehalten von Annahme III.II (*Bearbeitungen von on-screen Aufgaben werden ähnliche Testwerte zugewiesen wie inhaltlich identischen hands-on Aufgaben.)* sprechen. Als bedrohende Einschränkung ergibt sich aus Studie E (Testformatvergleich: Anfertigen eines Messwertediagramms) die fehlende Vergleichbarkeit zwischen on-screen und hands-on Format für die Teilaufgabe zum Anfertigen eines Messwertediagramms. Qualitative Detailanalysen zeigen zwar strukturell ähnliche Lösungsansätze beim Erstellen eines Messwertediagramms (vgl. Jansen, 2014), allerdings bilden sich diese in der vorliegenden Studie nicht in vergleichbaren Testwerten ab. Es konnten jedoch Überarbeitungshinweise für das Diagramm-Tool abgeleitet werden (z. B. Hilfen für das Eintragen von Messwerten und die Achsenbeschriftung), mit denen die bestehende Einschränkung wahrscheinlich behoben werden kann.

Die sich aus der Prüfung der Annahmen III.I und III.II (vgl. Tabelle 17.3 auf Seite 147) ergebenden Befunde stützen in der Regel die dritte übergeordnete Aussage (*Die beobachtete Performanz wird in geeigneter Art und Weise in Testwerte überführt*). Eine Ausnahme ist die nicht vorhandene Vergleichbarkeit der Formate für die Teilaufgabe zum Anfertigen eines Messwertediagramms (gemessen an den jeweiligen Testwerten). Aufgrund dieser bestehenden Einschränkung ist eine weitere empirische Absicherung wünschenswert. Insgesamt überwiegen aber auch bei der vorliegenden Befundlage solche Hinweise, die für das Beibehalten der Annahmen sprechen. Zusammenfassend lässt sich daher unter Berücksichtigung der Einschränkung festhalten:

Die Performanz wurde angemessen in Testwerte übersetzt.

Um noch stärkere Evidenz für das Beibehalten (bzw. die Ablehnung) der Annahmen bzw. der Aussage (*Die beobachtete Performanz wird in geeigneter Art und Weise in Testwerte überführt*) zu generieren, ist es sinnvoll, die Studienergebnisse durch eine Erweiterung der empirischen Datenbasis umfangreicher als bisher herauszufordern. Das betrifft insbesondere die Ergebnisse zum Vergleich der Testformate.

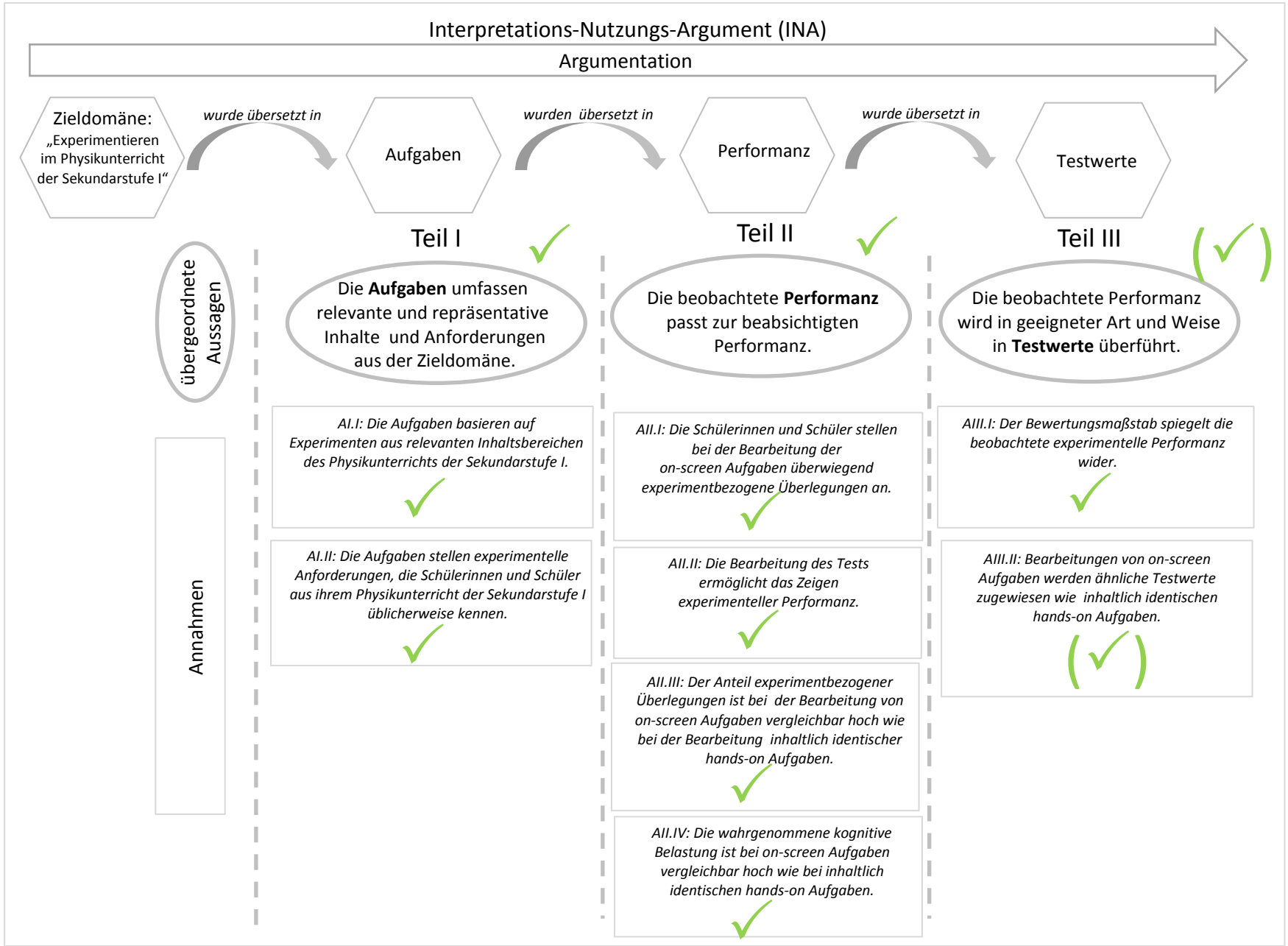
In einer Studie mit Schülerinnen und Schülern am Ende der Sekundarstufe I könnte geprüft werden, ob sich die vorliegenden Ergebnisse zur Vergleichbarkeit der Testformate für die Teilaufgabentypen *Versuchsplan entwerfen*, *Versuch aufbauen und testen* sowie *Messungen durchführen und dokumentieren* bestätigen lassen. Im Rahmen einer solchen Studie könnte auch ein modifiziertes Diagrammtool eingesetzt werden, um weitere Evidenz für bzw. gegen die Vergleichbarkeit der Testformate beim *Anfertigen eines Messwertediagramms* zu generieren. Eine solche Studie ist allerdings mit hohem materiellem und logistischem Aufwand verbunden. Vor der Durchführung einer solchen Studie sollte daher sorgfältig abgewogen werden, ob eine Herausforderung der vorliegenden Validitätsargumentation durch eine solche Studie tatsächlich notwendig und sinnvoll ist. Die vorliegenden Befunde zur Relevanz und Repräsentativität der Inhalte und Anforderungen (vgl. Abschnitt 17.1) und die Befunde zur beobachteten Performanz, insbesondere der hohe Anteil experimentbezogener Überlegungen

(vgl. Abschnitt 17.2), machen deutlich, dass der MeK-LSA Experimentiertest in der Lage ist, im on-screen Format Experimentierfähigkeiten auf inhaltlicher und kognitiver Ebene valide zu erfassen. Die Frage nach der Vergleichbarkeit des on-screen und hands-on Formats auf Ebene der Testwerte ist folglich für den MeK-LSA Experimentiertest nicht mehr zentral, da er bereits in der vorliegenden Form einen deutlichen Fortschritt zu bisherigen Large-Scale Experimentiertests darstellt, die nicht in der Lage sind, handlungsbezogene Experimentierfähigkeiten valide zu erfassen. Die Frage nach der Vergleichbarkeit des on-screen und hands-on Formats auf Ebene der Testwerte rückt erst dann wieder ins Zentrum des Interesses, wenn ein hands-on Test vorliegt, der bei großen Stichproben ähnlich effizient eingesetzt werden kann wie der MeK-LSA Experimentiertest.

17.4 Bewertung des INA: Zusammenfassung

Eine Übersicht über die Bewertung des INA ist in Abbildung 17.2 auf Seite 150 dargestellt. Die Übersicht zeigt die in den Abschnitten 17.1 bis 17.3 bewerteten übergeordneten Aussagen und Annahmen und veranschaulicht, dass die Ergebnisse der Aufgabenbearbeitungen und die daraus berechneten Testwerte (als Maß für die Lösungsqualität) – bis auf wenige Einschränkungen – valide als Ausdruck von Experimentierfähigkeiten interpretierbar sind.

Das INA fokussiert dabei bewusst auf den Prozess der Testentwicklung bis zur Zuweisung von Testwerten zu einzelnen experimentellen (Teil-)Aufgaben. Eine schlüssige Validitätsargumentation zu diesem Prozess liegt vor, sodass verlässliche Aussagen über die Experimentierfähigkeiten eines durchschnittlichen Schülers bzw. einer durchschnittlichen Schülerin für einzelne (Teil-)Aufgaben möglich sind. Damit ist eine notwendige Voraussetzung für eine inhaltlich sinnvolle Verwendung des MeK-LSA Experimentiertests erfüllt.



150 *Abbildung 17.2:* Übersicht über das INA (Die Argumentation wurde entlang der Schritte zu den (übergeordneten) Aussagen geführt; ✓: *Annahme bzw. Aussage kann beibehalten werden;* (✓)*: Annahme bzw. Aussage kann mit Einschränkungen beibehalten werden*

18 Schlussbemerkung und Ausblick

18.1 Schlussbemerkung

In der vorliegenden Dissertation wird der Prozess der Testentwicklung bis zur Zuweisung von Testwerten systematisch validiert. Durch die gewählte Vorgehensweise liegt für den MeK-LSA Experimentiertest eine transparente Prozessbeschreibung vor, die bei der praktischen Entwicklung von Testverfahren zur Kompetenzmessung bisher selten systematisch erfolgt und über die in der Folge – wenn überhaupt – nur bruchstückhaft berichtet wird.

Insgesamt fällt auf, dass die vorliegende Arbeit – aus dem Blickwinkel der Validität – ein positives Bild des MeK-LSA Experimentiertests zeichnet. Dieses positive Bild wird sogar noch verstärkt, wenn man weiterführende Ergebnisse aus der im Projekt MeK-LSA durchgeführten Large-Scale Studie berücksichtigt. Eine detaillierte Darstellung dieser Large-Scale Studie findet sich beispielsweise in Eickhorst (in Vorbereitung) und Theyßen et al. (2016b). Im Rahmen der Large-Scale Studie konnte gezeigt werden, dass mit dem MeK-LSA Experimentiertest Experimentierfähigkeiten von Schülerinnen und Schülern am Ende der Sekundarstufe I praktikabel, effizient und reliabel erfasst werden können. Darüber hinaus sind die empirisch gefundenen Itemschwierigkeiten inhaltlich plausibel erklärbar (Theyßen et al., 2016b). Insbesondere der letztgenannte Aspekt ist ein weiteres Argument für die Validität der Testwertinterpretation des MeK-LSA Experimentiertests.

Es ist allerdings nicht hinreichend, wenn man das positive Bild des MeK-LSA Experimentiertests alleine auf die systematische und sorgfältige Entwicklung und Validierung des Testverfahrens durch das Testentwicklungsteam zurückführt. Das positive Bild relativiert sich, wenn man bedenkt, dass die Aussagefähigkeit der Befunde durch die zugrundeliegende Kompetenzmodellierung (z. B. Definition experimenteller Kompetenz; Aufgabenentwicklungsmodell) beschränkt ist. Es bleibt daher abschließend zu diskutieren, welchen Beitrag der MeK-LSA Experimentiertest zum Assessment-Diskurs experimenteller Kompetenz (vgl. Gut, 2012) leistet (vgl. auch Problemaufriss in Abschnitt 1.2 auf Seite 12).

Mit dem MeK-LSA Experimentiertest liegt ein Test vor, der das Spektrum der in Large-Scale Assessments erfassbaren Experimentierfähigkeiten erweitert. Beispielsweise erfasst der Test Experimentierfähigkeiten über die tatsächliche Anwendung in repräsentativen – wenn auch on-screen präsentierten – Situationen (*Zeigen wie*). Der Schwerpunkt liegt dabei auf Experimentierfähigkeiten der Durchführung, da für Experimentierfähigkeiten der Planung und Auswertung bereits relativ vielversprechende schriftliche Testverfahren vorliegen (z. B. Glug, 2009; S. 221-237). In der Gesamtschau ermöglichen die Befunde zum MeK-LSA Experimentiertest die Initiierung eines theorie- und evidenzbasierten Diskurses, der im Kern auf die angemessene Erfassung der interessierenden Kompetenz fokussieren kann. Das Argument, sich bei der Erfassung von Experimentierfähigkeiten zur Durchführung aus ökonomischen und pragmatischen Gründen auf die Erfassung des Vorhandenseins von Wissen über experimentelles Vorgehen (*Wissen wie*) beschränken zu müssen, ist aufgrund der vorliegenden Befunde zum MeK-LSA Experimentiertest nicht mehr ohne weiteres haltbar.

Trotz des positiven Gesamtbilds des MeK-LSA Experimentiertests müssen für einen fairen und ergebnisoffenen Diskurs auch die inhaltlich kritischen Aspekte des Tests zusammenfassend benannt und diskutiert werden. Durch das konsekutive Aufgabenformat erfasst der MeK-LSA Experimentiertest keine Experimentierfähigkeiten, bei denen die Schülerinnen und Schüler beispielsweise die notwendigen experimentellen Teilschritte bzw. deren Abfolge eigenständig bestimmen können. Darüber hinaus kann mit dem MeK-LSA Experimentiertest nicht getestet werden, inwieweit Schülerinnen und Schüler ihre experimentelle Vorgehensweise während des Experimentierens verbessern können, da keine Rückschritte zu vorherigen experimentellen Teilschritten (z. B. von der Auswertung zur Messung) möglich sind. Daher erfasst der Test zwar, wie gut die einzelnen Experimentierfähigkeiten (z. B. Versuch aufbauen und testen) beherrscht werden, nicht aber, wie gut Schülerinnen und Schüler diese Fähigkeiten in Experimentiersituationen ohne Bearbeitungsleitfaden anwenden können. Folglich entspricht der MeK-LSA Experimentiertest sicherlich nicht einem fachdidaktischen Ideal von Experimentieren, da durch den Test nur ein eingeschränktes Bild des Experimentierens vermittelt wird. Andererseits deckt der Test zumindest relevante und repräsentative Anforderungen aus der Schulpraxis ab und stellt damit einen Fortschritt zu bisherigen Testverfahren dar, die im Rahmen des Bildungsmonitorings eingesetzt werden.

Auch die Nicht-Berücksichtigung rein qualitativer Aufgabenstellungen und die Nicht-Berücksichtigung von Aufgabenstellungen zur Erfassung weiterer – aus fachdidaktischer Perspektive – relevanter Experimentierfähigkeiten (z. B. *Aufstellen von Hypothesen*, *Entwickeln von Fragestellungen*, *Umgang mit Messunsicherheiten*, *Umgang mit Problemen und Fehlern*, *Reflexion der eigenen Vorgehensweise*, *Diskussion von Randbedingungen*) führen zu einem eingeschränkten Bild des Experimentierens. In der Tat beziehen sich die in den Aufgaben des MeK-LSA Experimentiertests gestellten Inhalte und Anforderungen ausschließlich auf Experimentierfähigkeiten, die mit dem experimentellen Prozess von der Entwicklung einer experimentellen Grundidee, zu einer gegebenen Fragestellung bis zur Interpretation von Messdaten bezüglich dieser Fragestellung verbunden sind. Es ist allerdings zu bedenken, dass das dem MeK-LSA Experimentiertest zugrundeliegende Aufgabenentwicklungsmodell anschlussfähig an bestehende Modelle experimenteller Kompetenz ist. Zusätzlich ermöglicht die durch das MeK-LSA Projektteam entwickelte Konstruktionsanleitung (vgl. Abschnitt 4.7.1 auf Seite 53) eine Umsetzung weiterer Aufgabenstellungen. Im Gegensatz zu den Einschränkungen, die sich durch das konsekutive Aufgabenformat ergeben, ist es daher potentiell möglich bisher nicht berücksichtigte Aspekte im Aufgabenformat des MeK-LSA Experimentiertests zu integrieren. Das kann entweder durch die Erweiterung bestehender Teilaufgabenstellungen (z. B. *Umgang mit Messunsicherheiten*) oder durch die Ergänzung weiterer Teilaufgabentypen (z. B. *Aufstellen von Hypothesen*) erfolgen. Ob eine solche Erweiterung sinnvoll ist und mit welchen Schwerpunktsetzungen die einzelnen Experimentierfähigkeiten dann getestet werden sollten, bleibt ein offener Diskussionspunkt im Assessment-Diskurs experimenteller Kompetenz.

Im Rahmen des Assessment-Diskurses sollte auch diskutiert werden, welches Bild vom Experimentieren in der Schulöffentlichkeit durch die Aufgabenstellungen des MeK-LSA Experimentiertests erzeugt wird. Die Aufgaben des MeK-LSA Experimentiertests bilden auf der

einen Seite zwar relevante und repräsentative Inhalte und Anforderungen aus der Schulpraxis ab. Auf der anderen Seite fehlen innovative und kreative Aufgabenstellungen beispielsweise mit sinnstiftenden Kontexten. Möchte man die Aufgabenkultur in der Schulpraxis durch den Einsatz von Testverfahren innovieren, zum Beispiel durch gezieltes *teaching to the test*, sind die Aufgaben des MeK-LSA Experimentiertests in der vorliegenden Form nicht geeignet. Bei dieser Diskussion ist allerdings zu bedenken, dass der Fokus bei der Entwicklung des MeK-LSA Experimentiertests bewusst zunächst auf der Umsetzung und Erprobung des innovativen konsekutiven Aufgabenformats lag. Die zusätzliche Einbindung innovativer Aufgabenstellungen hätte Schülerinnen und Schüler vermutlich eher überfordert. In der Folge wären keine verlässlichen Aussagen über die potentielle Eignung des konsekutiven Aufgabenformats des MeK-LSA Experimentiertests möglich gewesen. Inwieweit die Berücksichtigung kreativer und innovativer Aufgabenstellungen für zukünftige Verwendungen des MeK-LSA Experimentiertests (Stichwort: Innovation der Schulpraxis) sinnvoll und notwendig ist, bleibt zu diskutieren. Problematische Signale für die Schulpraxis ergeben sich allerdings erst dann, wenn sich notenrelevante Tests am MeK-LSA Experimentiertest orientieren würden. Im Hinblick auf eine valide Erfassung von Experimentierfähigkeiten kann der MeK-LSA Experimentiertest in jedem Fall bereits in der vorliegenden Form zu einer größeren Varianzaufklärung beitragen (Stichwort: Triangulation; vgl. Wendt & Bos, 2011, S. 17). Zusammenfassend lässt sich daher festhalten, dass der MeK-LSA Experimentiertest eine sinnvolle Ergänzung und Weiterentwicklung zu bisherigen Testverfahren zur Erfassung von Experimentierfähigkeiten darstellt.

18.2 Ausblick

Mit dem MeK-LSA Experimentiertest liegt ein Test vor, der experimentelle Kompetenz, operationalisiert durch das zugrunde liegende Aufgabenentwicklungsmodell, effizient, reliabel und valide messen kann. Das eröffnet zum einen die Möglichkeit, nationale Vergleichsstudien und internationale Schulleistungsstudien (z. B. PISA) durch den Einsatz des MeK-LSA Experimentiertests zu erweitern. Auf diese Weise könnten auch im Rahmen eines nationalen oder internationalen Bildungsmonitorings detailliertere Aussagen über die Ausprägung von Experimentierfähigkeiten getroffen werden. Der Einsatz in internationalen Studien setzt allerdings voraus, dass die Inhalte und Anforderungen des Tests auch international bedeutsam sind. Zumindest für den deutschsprachigen Raum (Österreich und Schweiz) liegen erste Hinweise in diese Richtung vor.[27] Einschränkend bleibt hier zu erwähnen, dass der Fokus bei den Untersuchungen im Projekt MeK-LSA bisher auf Schülerinnen und Schülern des Gymnasiums lag. Für die Entwicklung und Validierung war zunächst das am Gymnasium vorhandene Leistungsspektrum ausreichend. Für einen weiteren Einsatz des Testverfahrens im Rahmen des nationalen und internationalen Bildungsmonitorings sollte idealerweise auch noch die Eignung des Testverfahrens für Schülerinnen und Schüler anderer Schulformen

[27] Im Rahmen der Expertentagung (vgl. Abschnitt 4.5 auf Seite 44) wurden die Aufgabenskizzen mit einem Schweizer Experten und einer österreichischen Expertin aus der Physikdidaktik diskutiert. Beide bestätigen, dass die Aufgabenentwürfe bis auf wenige Ausnahmen (Österreich: 2; Schweiz: 1) auch für Schülerinnen und Schüler an österreichischen und Schweizer Gymnasien relevant sind.

geprüft werden. Das gilt insbesondere für Hauptschülerinnen und Hauptschüler, da diese in der Regel das untere Ende des Leistungsspektrums abbilden. In einer qualitativen Studie mit Hauptschülerinnen und Hauptschülern konnte Matusik (2013) im Rahmen ihrer Examensarbeit zumindest erste Hinweise finden, dass auch Hauptschülerinnen und Hauptschüler potentiell in der Lage sind, eine sprachlich vereinfachte Testaufgabe des MeK-LSA Experimentiertests sinnvoll zu bearbeiten.

Ein weiterer Verwendungszweck für den MeK-LSA Experimentiertest besteht nach Theyßen et al. (2016b) darin, die Lernwirksamkeit experimenteller Interventionen detailliert zu untersuchen. Hierzu ist jedoch vorab zu prüfen, ob der MeK-LSA Experimentiertest ausreichend empfindlich misst (ebenda).

Die in der Large-Scale Studie ermittelten Itemschwierigkeiten könnten darüber hinaus die Grundlage bilden, um mit dem MeK-LSA Experimentiertest Experimentierfähigkeiten adaptiv zu testen. Während die Anzahl der entwickelten Teilaufgaben für eine erste Erprobung einer adaptiven Testvariante ausreichend ist, besteht bezüglich der Datenauswertung noch Optimierungsbedarf. Bislang erfolgt zwar eine teilautomatisierte und in der Folge zeitökonomische Bewertung der Schülerlösungen (Aufwand für geschulte Rater: ca. drei Minuten für jede Testaufgabe mit sechs Teilaufgaben). Für eine adaptive Testversion müssten die Auswertungsroutinen allerdings vollständig automatisiert ablaufen. Zusätzlich wäre es für die Erstellung einer adaptiven Testversion empfehlenswert, die Itemschwierigkeiten durch eine Erhebung mit einer repräsentativen Stichprobe zunächst erneut zu bestimmen.

Neben dem Einsatz des MeK-LSA Experimentiertests als Testverfahren könnten die Testaufgaben (und mögliche nachentwickelte Aufgaben) auch als (Selbst-)Lerneinheiten verwendet werden. Hat ein Schüler beispielsweise Schwierigkeiten beim Planen von Experimenten in der Optik, werden ihm weitere Aufgaben dieses Typs zum Üben zur Verfügung gestellt. Unter der Voraussetzung, dass die Auswertung vollständig automatisiert ablaufen kann, erfolgt die Auswahl der nächsten Aufgabe anhand der Qualität der Schülerlösung. Die Verwendung als (Selbst-)Lerneinheit ist aber nicht zwingend an eine vollständig automatisierte Auswertung geknüpft. Zum einen könnte die Lehrkraft in den Auswahlprozess eingebunden werden, wobei unterrichtstaugliche Auswahlkriterien (z. B. globale Sichtprüfung durch Lehrkraft) diskutiert werden müssten. Zum anderen könnten auch Schülerselbstbeurteilungen (vgl. Schreiber & Theyßen, 2015) die Grundlage für die Auswahl der nächsten Aufgabe sein.

Aus dem Blickwinkel der Validität stellt sich abschließend die Frage, wie umfangreich die Validierungsstudien für weitere Verwendungszwecke gestaltet werden müssen. Aus theoretischer Perspektive kann die Frage nach der Validität immer nur vor dem Hintergrund des jeweiligen Verwendungszwecks beantwortet werden (vgl. Abschnitt 5.1 auf Seite 55). Weiterführende Verwendungszwecke des MeK-LSA Experimentiertests müssen sich folglich immer auch einer angemessenen und umfangreichen Validitätsargumentation stellen. Für eine solche Argumentation sind die in der vorliegenden Dissertation generierten Befunde jedoch eine fundierte Grundlage. Das führt dazu, dass der Aufwand für Validitätsbetrachtungen weiterer Verwendungszwecke reduziert wird, da sich notwendige Annahmen bereits auf Basis der vorhandene Befunde plausibel erklären lassen. Dabei ist aber auf die Vermeidung der

reification fallacy (Stichwort: *vorschnelle Generalisierung*; vgl. Seite 65) durch eine kritische Prüfung der vorhandenen Befunde zu achten. Sollen mit dem MeK-LSA Experimentiertest beispielsweise Aussagen auf der Ebene einzelner Schülerinnen und Schüler getroffen werden, müsste der Detailgrad vieler Analysen (z. B. Aufgabenbearbeitungsprozesse) erweitert werden, um mögliche Bedrohungen für die Validitätsargumentation zu erkennen. Folgt man dieser Auffassung konsequent, wäre allerdings bereits die Validierung der Testwertinterpretation des MeK-LSA Experimentiertests ein niemals endender Prozess. Aus praktischer Perspektive erscheint diese Sichtweise weder sinnvoll noch praktikabel, da nur ein begrenzter Validierungsaufwand tatsächlich leistbar ist.

Zusammenfassend lässt sich festhalten, dass der vorliegende MeK-LSA Experimentiertest eine fruchtbare Grundlage für weiterführende Verwendungen liefert.

Literaturverzeichnis

Abrahams, I., Reiss, M. J., and Sharpe, R. M. (2013). The assessment of practical work in school science. *Studies in Science Education*, 49(2), 209–251.

Adamina, M., Labudde, P., Gingins, F., Nidegger, C., Bazzigher, L., Bringold, B., ... Zeyer, A. (2009). *Naturwissenschaften: Wissenschaftlicher Kurzbericht und Kompetenzmodell.* Bern: HarmoS Konsortium Naturwissenschaften+.

AERA, APA & NCME (2014). *Standards for educational and psychological testing*. Washington D.C.: American Educational Research Association.

Allalouf, A., & Ben-Shakhar, G. (1998). The effect of coaching on the predictive validity of scholastic aptitude tests. *Journal of Educational Measurement*, *35*(1), 31-47.

Anastasi, A. (1986). Evolving concepts of test validation. *Annual Review of Psychology*, 37(1*)*, 1-15.

Artelt, C., Baumert, J., Klieme, E., Neubrand, M., Prenzel, M., Schiefele, U., ... & Weiß, M. (2001). *PISA 2000: Zusammenfassung zentraler Befunde*. Berlin: Max-Planck-Institut für Bildungsforschung. *Abgerufen unter https://www.mpib-berlin.mpg.de/Pisa/ergebnisse.pdf (Datum: 11.03.16).*

Aryadoust, V. (2013). *Building a validity argument for a listening test of academic proficiency*. Newcastle: Cambridge Scholars Publishing.

Aschbacher, P. R. (1991). Performance assessment: State activity, interest, and concerns. *Applied measurement in Education*, *4*(4), 275-288.

Aufschnaiter, C. v., & Rogge, C. (2010). Wie lassen sich Verläufe der Entwicklung von Kompetenz modellieren. *Zeitschrift für Didaktik der Naturwissenschaften*, *16*, 95-114.

Bannert, M. (2007). *Metakognition beim Lernen mit Hypermedien*. Münster: Waxmann.

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

Baumert, J., Klieme, E., Lehrke, M., & Savelsbergh, E. (2000). Konzeption und Aussagekraft der TIMSS-Leistungstests. Zur Diskussion um TIMSS-Aufgaben aus der Mittelstufenphysik (preprint). *Abgerufen unter https://www.mpib-berlin.mpg.de/volltexte/institut/dok/full/Baumert/bjkuads_/Hagemeister.pdf (Datum: 11.03.16).*

Baxter, G. P., Shavelson, R. J., Goldman, S. R. & Pine, J. (1992). Evaluation of Procedure-Based Scoring for Hands-On Science Assessment. *Journal of Educational Measurement, 29*(1), 1-17.

Baxter, G. P., & Shavelson, R. J. (1994). Science performance assessments: Benchmarks and surrogates. *International Journal of Educational Research*, *21* (3), 279-298.

Blömeke, S. (2013). Validierung als Aufgabe im Forschungsprogramm „Kompetenzmodellierung und Kompetenzerfassung im Hochschulsektor“ (KoKoHs Working Papers, 2). Berlin & Mainz: Humboldt-Universität & Johannes Gutenberg-Universität.

Borawski, H., Theyßen, H. & Heinke, H. (2005). Entwicklung eines Physikpraktikums für Studierende der Biologie. *Didaktik der Physik–Beiträge zur DPG-Frühjahrstagung 2005 in Berlin*, 1-6.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111(4)*, 1061–1071.

Borsboom, D. & Mellenbergh, G. J. (2007). Test Validity in Cognitive Assessment. In J. P. Leighton & M. J. Gierl (Hrsg.), *Cognitive diagnostic assessment for education: theory and applications.* Cambridge: Cambridge University Press, 85-118.

Borsboom, D. & Markus, K. A. (2013). Truth and Evidence in Validity Theory. *Journal of Educational Measurement*, *50(1)*, 110–114.

Bortz, J. & Döring, N. (2006). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler.* Springer: Heidelberg.

Breakspear, S. (2012). *The policy impact of PISA: an exploration of the normative effects of international benchmarking in school system performance* (No. 71). OECD Publishing.

Brennan, R. L. (2011). *Generalizability Theory. Statistics for social and behavorial sciences*. New York: Springer.

Brooke, J. (1996). SUS-A quick and dirty usability scale. *Usability Evaluation in Industry*, *189*(194), 4-7.

Brunken, R., Plass, J. L., & Leutner, D. (2003). Direct measurement of cognitive load in multimedia learning. *Educational Psychologist*, *38*(1), 53-61.

Brunner, M., Artelt, C., Krauss, S., & Baumert, J. (2007). Coaching for the PISA test. *Learning and Instruction*, *17*(2), 111-122.

Buber, R. (2007). Denke-Laut-Protokolle. In R. Buber & H. H. Holzmüller (Hrsg.), *Qualitative Marktforschung Konzepte – Methoden – Analysen*. Göttingen: Hogrefe, 557-568.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56(2)*, 81–105.

Chapelle, C. A., Enright, M. K. & Jamieson, J. M. (2008). *Building a Validity Argument for the Test of English as a Foreign Language*. New York: Routledge.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Hrsg.), *Educational measurement* (2.Ausgabe), Washington D.C.: American Council on Education, 443-507.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Hrsg.), *Test validity*. Hillsdale, NJ: Lawrence Erlbaum, 3-17.

Crooks, T. J., Kane, M. T., & Cohen, A. S. (1996). Threats to the valid use of assessments. *Assessment in Education*, *3*(3), 265-286.

DfE, Department for Education (Hrsg.) (2013). *National Curriculum England. Science programmes of study: key stage 3*. London: Department for Education.

DfEE, Department for Education and Employment (Hrsg.) (1999). *Science – The national Curriculum for England*. London: Department for Education and Employment.

Dickmann, M., & Theyßen, H. (2013). Curriculare Validität von Units zur Messung experimenteller Kompetenz. In S. Bernholt (Hrsg.), *Inquiry-based Learning – Forschendes Lernen. Gesellschaft für Didaktik der Chemie und Physik, Jahrestagung in Hannover 2012*. Kiel: IPN, 587–589.

Eckloff, B. (2014). *Wie gut können Schüler experimentieren? Vergleich zweier Aufgabenformate in Simulationen zur Messung experimenteller Kompetenz (Masterarbeit)*. Universität Bremen.

Eickhorst, B., Dickmann, M., Schecker, H., Theyßen, H. & Neumann, K. (2015). Messung experimenteller Kompetenz im Large Scale: Bewertung experimenteller Aufgaben. In S. Bernholt (Hrsg.), *Heterogenität und Diversität – Vielfalt der Voraussetzungen im naturwissenschaftlichen Unterricht. Gesellschaft für Didaktik der Chemie und Physik, Jahrestagung in Bremen 2014*. Kiel: IPN, 169-171.

Eickhorst, B. (in Vorbereitung). *Experimentelle Kompetenz - Zur Validierung eines Fähigkeitskonstrukts (Dissertation)*. Universität Bremen.

Emden, M. (2011). *Prozessorientierte Leistungsmessung des naturwissenschaftlich-experimentellen Arbeitens: eine vergleichende Studie zu Diagnoseinstrumenten zu Beginn der Sekundarstufe I* (Studien zum Physik- und Chemielernen, Band 118). Berlin: Logos.

Emden, M., & Sumfleth, E. (2012). Prozessorientierte Leistungsbewertung. Zur Eignung einer Protokollmethode für die Bewertung von Experimentierprozessen. *Der mathematische und naturwissenschaftliche Unterricht*, *65*(2), 68-75.

Erb, R., Neumann, K. & Härtig, H. (2015). Bestandteile Experimenteller Kompetenzen. Eine Expertenbefragung. *Vortragsfolien zum Vortrag auf der DPG Frühjahrstagung 2015 in Wuppertal.*

Ericsson, K. A. & Simon, H. (1993). *Protocol Analysis. Verbal Reports as Data*. London: The MIT Press.

Gärtner, H. & Pant, H. E. (2011a). Validierungsstrategien für Verfahren und Ergebnisse von Schulinspektion. In S. Müller, M. Pietsch, & W. Bos (Hrsg.), *Schulinspektion in Deutschland. Eine Zwischenbilanz aus empirischer Sicht*. Münster: Waxmann, 9-32.

Gärtner, H., & Pant, H. A. (2011b). How valid are school inspections? Problems and strategies for validating processes and results. *Studies in educational evaluation*, *37*(2), 85-93.

Gass, S. M. & Mackey, A. (2000). *Stimulated Recall Methodology in Second Language Research*. New York: Routledge.

Gierl, M. J., Leighton, J. P., & Hunka, S. M. (2007). Using the attribute hierarchy method to make diagnostic inferences about examinees' cognitive skills. In J. P. Leighton & M. J. Gierl (Hrsg.), Cognitive diagnostic assessment for education: theory and applications. Cambridge: Cambridge University Press, 242-274.

Glug, I. (2009). *Entwicklung und Validierung eines Multiple-Choice-Tests zur Erfassung prozessbezogener naturwissenschaftlicher Grundbildung (Dissertation)*. Christian-Albrechts Universität Kiel.

Gott, R., & Duggan, S. (2002). Problems with the assessment of performance in practical science: which way now?. *Cambridge Journal of Education*, *32*(2), 183-201.

Gut, C. (2012). *Modellierung und Messung experimenteller Kompetenz: Analyse eines large-scale Experimentiertests* (Studien zum Physik- und Chemielernen, Band 134). Berlin: Logos.

Gut, C., Hild, P., Metzger, S. & Tardent, J. (2014a). Projekt ExKoNawi: Modell für hands-on Assessments experimenteller Kompetenzen. In S. Bernholt (Hrsg.), *Naturwissenschaftliche Bildung zwischen Science- und Fachunterricht. Gesellschaft für Didaktik der Chemie und Physik, Jahrestagung in München 2013*. Kiel: IPN, 171-173.

Gut, C., Metzger, S., Hild, P., & Tardent, J. (2014b). Problemtypenbasierte Modellierung und Messung experimenteller Kompetenzen von 12- bis 15-jährigen Jugendlichen. *PhyDid B, Didaktik der Physik, Beiträge zur DPG-Frühjahrstagung 2014 in Frankfurt*, 1-9.

Häder, M. (2014). *Delphi-Befragungen. Ein Arbeitsbuch* (3. Auflage). Wiesbaden: Springer.

Hamilton, L. S. (1994). *Validating Hands-On Science Assessments through an Investigation of Response Processes.* New Orleans: Paper presented at the 1994 Annual Meeting of the American Educational Research Association.

Hammann, M. (2004). Kompetenzentwicklungsmodelle. Merkmale und ihre Bedeutung - dargestellt anhand von Kompetenzen beim Experimentieren. *Der mathematische und naturwissenschaftliche Unterricht*, *57(4)*, 196-203.

Hartig, J., & Jude, N. (2007). Empirische Erfassung von Kompetenzen und psychometrische Kompetenzmodelle. In J. Hartig & E. Klieme (Hrsg.), *Möglichkeiten und Voraussetzungen technologiebasierter Kompetenzdiagnostik. Eine Expertise im Auftrag des BMBF.* Bonn: Bundesministerium für Bildung und Forschung, 17-36.

Hartig, J., Frey, A. & Jude, N. (2012). Validität. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (2. Auflage). Berlin: Springer, 143-171.

Härtig, H. (2010). *Sachstrukturen von Physikschulbüchern als Grundlage zur Bestimmung der Inhaltsvalidität eines Tests* (Studien zum Physik- und Chemielernen , Band 101). Berlin: Logos.

Härtig, H., Kauertz, A. & Fischer, H. E. (2012). Das Schulbuch im Physikunterricht. Nutzung von Schulbüchern zur Unterrichtsvorbereitung in Physik. *Der mathematische und naturwissenschaftliche Unterricht, 65*(4), 197-200.

Haslam, C. Y., & Hamilton, R. J. (2010). Investigating the use of integrated instructions to reduce the cognitive load associated with doing practical work in secondary school science. *International Journal of Science Education*, *32*(13), 1715-1737.

Hild, P., Tardent, J., Gut, C. & Metzger, S. (2015). Projekt ExKoNawi: Typenspezifische Kompetenzprogressionen bei hands-on Testaufgaben. In S. Bernholt (Hrsg.), *Heterogenität und Diversität - Vielfalt der Voraussetzungen im naturwissenschaftlichen Unterricht. Gesellschaft für Didaktik der Chemie und Physik, Jahrestagung in Bremen 2014.* Kiel: IPN, 145-147.

Höttecke, D., & Rieß, F. (2015). Naturwissenschaftliches Experimentieren im Lichte der jüngeren Wissenschaftsforschung – Auf der Suche nach einem authentischen Experimentbegriff der Fachdidaktik. *Zeitschrift für Didaktik der Naturwissenschaften*, *21*(1), 127-139.

International Test Commission (2001). International guidelines for test use. *International Journal of Testing*, *1*(2), 93-114.

IQB (2013). Kompetenzstufenmodelle zu den Bildungsstandards im Fach Physik für den Mittleren Schulabschluss. Kompetenzbereiche "Fachwissen" und "Erkenntnisgewinnung" - Entwurf. *Abgerufen unter: https://www.iqb.hu-berlin.de/bista/ksm/KSM_Physik.pdf (Datum 11.3.2016).*

Jansen, F. (2014). *Diagrammerstellung am PC bzw. mit Papier und Bleistift ein Vergleich von Schülerleistungen (Bachelorarbeit)*. Universität Duisburg-Essen.

Jaschinski, T. (2013). *Entwicklung und Evaluation einer Trainingsunit für einen Online-Experimentiertest (Staatsexamensarbeit).* Universität Duisburg-Essen.

Jenßen, L., Dunekacke, S., & Blömeke, S. (2015). Qualitätssicherung in der Kompetenzforschung: Empfehlungen für den Nachweis von Validität in Testentwicklung und Veröffentlichungspraxis. *Zeitschrift für Pädagogik*, *61 (Beiheft)*, 11-31.

Johnson, S. (1989). *National Assessment: The APU Science Approach*. London: HM Stationery Office.

Jurecka, A., & Hartig, J. (2007). Computer- und netzwerkbasiertes Assessment. In J. Hartig & E. Klieme (Hrsg.), *Möglichkeiten und Voraussetzungen technologiebasierter Kompetenzdiagnostik. Eine Expertise im Auftrag des BMBF.* Bonn: Bundesministerium für Bildung und Forschung, 37-48.

Kalyuga, S., Chandler, P., & Sweller, J. (1999). Managing split-attention and redundancy in multimedia instruction. *Applied Cognitive Psychology*, *13*(4), 351-371.

Kane, M. T. (1992). An argument-based approach to validation. *Psychological Bulletin, 112(3),* 527–535.

Kane, M. T., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational measurement: issues and practice*, *18*(2), 5-17.

Kane, M. T. (2006). Validation. In R. Brennan (Hrsg.), *Educational measurement* (4. Ausgabe). Westport, CT: American Council on Education and Praeger, 17-64.

Kane, M. (2012). Validating score interpretations and uses. *Language Testing*, *29*(1), 3-17.

Kane, M. T. (2013). Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement*, *50(1)*, 1–73.

Karaböcek, F. & Erb, R. (2015). Survey Experimente – Der Einsatz von Experimenten im Physikunterricht. In S. Bernholt (Hrsg.), *Heterogenität und Diversität - Vielfalt der Voraussetzungen im naturwissenschaftlichen Unterricht. Gesellschaft für Didaktik der Chemie und Physik, Jahrestagung in Bremen 2014*. Kiel: IPN, 399-401.

Kelle, U., & Kluge, S. (2010). *Vom Einzelfall zum Typus: Fallvergleich und Fallkontrastierung in der qualitativen Sozialforschung* (2. *Auflage).* Wiesbaden: VS Verlag für Sozialwissenschaften.

Kempa, R. (1986). *Assessment in science*. Cambridge Science Education Series. Cambridge: Cambridge University Press.

Klieme, E., Avenarius, H., Blum, W., Döbrich, P., Gruber, H., Prenzel, M., Reiss, K., Riquarts, K., Rost, J., Tenorth, H. E. & Vollmer, H. J. (2003). *Zur Entwicklung nationaler Bildungsstandards*. *Eine Expertise*. Bonn: Bundesministerium für Bildung und Forschung.

Klieme, E., & Leutner, D. (2006). Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen. Beschreibung eines neu eingerichteten Schwerpunktprogramms der DFG. *Zeitschrift für Pädagogik*, *52*(6), 876-903.

Klieme, E., Maag-Merki, K., & Hartig, J. (2007). Kompetenzbegriff und Bedeutung von Kompetenzen im Bildungswesen. In J. Hartig, & E. Klieme, E. (Hrsg.), *Möglichkeiten und Voraussetzungen technologiebasierter Kompetenzdiagnostik. Eine Expertise im Auftrag des BMBF.* Bonn: Bundesministerium für Bildung und Forschung, 5-15.

KMK, Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (Hrsg.) (2005a). *Bildungsstandards im Fach Biologie für den Mittleren Schulabschluss*. München: Luchterhand.

KMK, Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (Hrsg.) (2005b). *Bildungsstandards im Fach Chemie für den Mittleren Schulabschluss*. München: Luchterhand.

KMK, Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (Hrsg.) (2005c). *Bildungsstandards im Fach Physik für den Mittleren Schulabschluss*. München: Luchterhand.

Köller, O. & Schöps, K. (2013). Die deutsche Schule im Lichte internationaler Schulleistungsuntersuchungen (TIMSS, PISA, PIRLS/IGLU, DESI und TEDS M). In L. Haag, S. Rahm, H. J. Apel & W. Sacher(Hrsg.), *Studienbuch Schulpädagogik.* Bad Heilbrunn: Klinkhardt *UTB,* 72-96.

Konrad, K. (2010). Lautes Denken. In G. Mey, & K. Mruck. *Handbuch Qualitative Forschung in der Psychologie*. Wiesbaden: VS Verlag für Sozialwissenschaften, 476-490.

Kuhn, W. (Hrsg.) (2008). *Physik 1*. Braunschweig: Westermann.

Lachmayer, S. (2008). *Entwicklung und Überprüfung eines Strukturmodells der Diagrammkompetenz für den Biologieunterricht (Dissertation)*. Christian-Albrechts Universität Kiel.

Leuders, T. (2014). Modellierungen mathematischer Kompetenzen–Kriterien für eine Validitätsprüfung aus fachdidaktischer Sicht. *Journal für Mathematik-Didaktik*, *35*(1), 7-48.

Leutner, D., Hartig, J. & Jude, N. (2008). Measuring Competencies: Introduction to Concepts and Questions of Assessment in Education. In J. Hartig, E. Klieme & D. Leutner (Hrsg.), *Assessment of Competencies in Educational Contexts.* Göttingen: Hogrefe, 177-192.

Lissitz, R. W., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher*, *36*(8), 437-448.

Lunetta, V. N. (2003). The School Science Laboratory: Historical Perspectives and Contexts for Contemporary Teaching. In B. J. Fraser & K. G. Tobin (Hrsg.), *International Handbook of Science Education. Special Paperback Edition.* Dordrecht: Kluwer, 249-262.

Maiseyenka, V., Schecker, H., & Nawrath, D. (2013). Kompetenzorientierung des naturwissenschaftlichen Unterrichts - Symbiotische Kooperation bei der Entwicklung eines Modells experimenteller Kompetenz. *Physik und Didaktik in Schule und Hochschule*, *1*(12), 1-17.

Maiseyenka, V. (2014). *Modellbasiertes Experimentieren im Unterricht: Praxistauglichkeit und Lernwirkungen* (Studien zum Physik- und Chemielernen, Band 166). Berlin: Logos.

Marcus, N., Cooper, M., & Sweller, J. (1996). Understanding instructions. *Journal of educational psychology*, *88*(1), 49-63.

Matters, G. (2009). A problematic leap in the use of test data: From performance to inference. In C. Wyatt-Smith (Hrsg.), *Educational Assessment in the 21st Century*. Springer: Netherlands, 209-225.

Matusik. M. D. (2013). *Messung experimenteller Fähigkeiten bei Hauptschülerinnen und –schülern. Adaption und Erprobung von Testaufgaben (Staatsexamensarbeit)*. Universität Duisburg-Essen.

Mayer, J. (2007). Erkenntnisgewinnung als wissenschaftliches Problemlösen. In H. Vogt, & D. Krüger (Hrsg.), *Theorien in der biologiedidaktischen Forschung: Ein Handbuch für Lehramtsstudenten und Doktoranden*. Springer: Berlin, 177-186.

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin*, *114*(3), 449-458.

Meier, M. & Mayer, J. (2012). Experimentierkompetenz praktisch erfassen – Entwicklung und Validierung eines anwendungsbezogenen Aufgabendesigns. In U. Harms & F. X. Bogner (Hrsg.), *Lehr- und Lernforschung in der Biologiedidaktik, Band 5*. Innsbruck: Studienverlag, 81-98.

Meier, M. (2014). Wie lassen sich Experimentierfähigkeiten von Schülerinnen und Schülern diagnostizieren und beschreiben? In A. Fischer, C. Hößle, S. Jahnke-Klein, H. Kipper, M. Komorek, J. Michaelis, V. Niesel & J. Sjuts (Hrsg.), *Diagnostik für lernwirksamen Unterricht*. Baltmannsweiler: Schneider Verlag Hohengehren, 127-143.

Meier, M. & Mayer, J. (2014). Selbständiges Experimentieren: Entwicklung und Einsatz eines anwendungsbezogenen Aufgabendesigns. *Der mathematische und naturwissenschaftliche Unterricht, 67*(1), 4-10.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35(11),* 1012-1027.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. Braun (Hrsg.), *Test validity.* Hillsdale, NJ: Lawrence Erlbaum, 33-45.

Messick, S. (1989). Validity. In R. L. Linn (Hrsg.), *Educational measurement* (3. Ausgabe). New York: American Council on Education, 13-103.

Messick, S. (1995). Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American psychologist*, *50*(9), 741-749.

Messick, S. (1996). Validity of Performance Assessments. In G. W. Phillips (Hrsg.), *Technical Issues in Large-Scale Performance Assessment.* Washington D.C.: US Government Printing Office, 1-18.

Mikelskis, H. F. (2006). *Physik-Didaktik*. Berlin: Cornelsen Scriptor.

Miller, G. E. (1990). The Assessment of Clinical Skills/Competence/Performance. *Academic Medicine. Journal of the Association of American Medical Colleges*, *65 (9),* 63–67.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, *1*(1), 3-67.

MSW NRW, Ministerium für Schule und Weiterbildung des Landes Nordrhein-Westfalen (2008). *Kernlehrplan für das Gymnasium der Sekundarstufe I in Nordrhein-Westfalen.* Düsseldorf: Ministerium für Schule und Weiterbildung des Landes Nordrhein-Westfalen.

NAGB, National Assessment Governing Board (2008). Science Framework for the 2009 National Assessment of Educational Progress. Washington D.C.: U.S. Government Printing Office.

National Research Council (NRC) (1996). *National Science Education Standards*. Washington D.C.: National Academies Press.

National Research Council (NRC) (2012). *Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*. Washington D.C.: National Academies Press.

Nawrath, D., Maiseyenka, V., & Schecker, H. (2011). Experimentelle Kompetenz–Ein Modell für die Unterrichtspraxis. *Praxis der Naturwissenschaften–Physik in der Schule*, *60*(6), 42-49.

NCES, National Center for Education Statistics (2012). The Nation's Report Card: Science in Action: Hands-On and Interactive Computer Tasks From the 2009 Science Assessment. (NCES 2012-468). Washington D. C.: Institute of Education Sciences, U.S. Department of Education.

Nehring, A., Nowak, K. H., Upmeier zu Belzen, A. & Tiemann, R. (2012). "VerE-Studie": Aufgabenentwicklung für eine modellbasierte Erfassung von Schülerkompetenzen im Bereich der Erkenntnisgewinnung des Chemie- und Biologieunterrichts. In S. Bernholt (Hrsg.), *Konzepte fachdidaktischer Strukturierung für den Unterricht.* Münster: LIT-Verlag, 301-303.

Nehring, A. (2014). *Wissenschaftliche Denk-und Arbeitsweisen im Fach Chemie: eine kompetenzorientierte Modell-und Testentwicklung für den Bereich der Erkenntnisgewinnung* (Studien zum Physik- und Chemielernen, Band 177). Berlin: Logos.

Neumann, K. (2004). *Didaktische Rekonstruktion eines physikalischen Praktikums für Physiker* (Studien zum Physik- und Chemielernen, Band 38). Berlin: Logos.

Neumann, K. (2013). Mit welchem Auflösungsgrad können Kompetenzen modelliert werden? In welcher Beziehung stehen Modelle zueinander, die Kompetenz in einer Domäne mit unterschiedlichem Auflösungsgrad beschreiben?. *Zeitschrift für Erziehungswissenschaft*, *16*(1), 35-39.

Newton, P. E. (2012a). Clarifying the consensus definition of validity. *Measurement: Interdisciplinary Research and Perspectives, 10(1-2)*, 1-29.

Newton, P. E. (2012b). Questioning the consensus definition of validity. *Measurement: Interdisciplinary Research and Perspectives, 10(1-2)*, 110-122.

Newton, P. E., & Shaw, S. (2014). *Validity in Educational and Psychological Assessment*. London: Sage.

Nitko, A. J., & Brookhart, S. M. (2007). *Educational Assessment of Students* (5. Auflage). Upper Saddle River: Pearson.

Oetinger, B. (2013). Förderung der Experimentierkompetenz im integrierten Naturwissenschaftsunterricht im Jahrgang 8, Unterrichtseinheit ”Bewegung”. In H. Schecker, D. Nawrath, H. Elvers, J. Borgstädt, S. Einfeldt & V. Maiseyenka (Hrsg.), *Modelle und Lernarrangements für die Förderung naturwissenschaftlicher Kompetenzen*. Hamburg: Landesinstitut für Lehrerbildung und Schulentwicklung, 81-88.

Paas, F. G. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *Journal of Educational Psychology*, *84*(4), 429-434.

Pant, H. A., Stanat, P., Pöhlmann, C., & Böhme, K. (2013). Die Bildungsstandards im allgemeinbildenden Schulsystem. In H. A. Pant, P. Stanat, U. Schroeders, A. Roppelt, T. Siegle, & C. Pöhlmann (Hrsg.), *IQB-Ländervergleich 2012: mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I*. Münster: Waxmann, 13-22.

Pant, H. A., Böhme, K. & Köller, O. (2013). Das Kompetenzkonzept der Bildungsstandards und die Entwicklung von Kompetenzstufenmodellen. In H. A. Pant, P. Stanat, U. Schroeders, A. Roppelt, T. Siegle, & C. Pöhlmann (Hrsg.), *IQB-Ländervergleich 2012: mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I*. Münster: Waxmann, 53-60.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). Knowing what students know. *The Science and Design of Educational Assessment.* Washington D.C.: National Academy Press.

Quellmalz, E. S., Timms, M. J., & Buckley, B. (2010). The promise of simulation-based science assessment: The Calipers project. *International Journal of Learning Technology*, *5*(3), 243-263.

Quellmalz, E. S., Silberglitt, M. D., & Timms, M. J. (2011). How can simulations be components of balanced state science assessment systems. *Abgerufen unter http://simscientist.org/downloads/SimScientistsPolicyBrief.pdf (Datum:11.03.16).*

Reynolds, C. R., Livingston, R. & Willson, V. (2010). *Measurement and assessment in education* (2.Auflage). London: Pearson Education International.

Rogge, C. (2010). *Entwicklung physikalischer Konzepte in aufgabenbasierten Lernumgebungen* (Studien zum Physik- und Chemielernen, Band 106). Berlin: Logos.

Ropohl, M., Sumfleth, E., & Walpuski, M. (2014). Lehrpläne, Kerncurricula, Bildungspläne usw.. Gibt es eine Einheit in der Vielfalt inhaltlicher Vorgaben für das Fach Chemie? *CHEMKON*, *21*(1), 7-14.

Rossa, H. (2012). *Mentale Prozesse beim Hörverstehen in der Fremdsprache. Eine Studie zur Validität der Messung sprachlicher Kompetenzen*. Frankfurt am Main: Peter Lang.

Rubin, D. B. (1988). Discussion. In H. Wainer & H. I. Braun (Hrsg.), *Test Validity*. Hillsdale, NJ: Lawrence Erlbaum, 241-256.

Rupp, A. A., & Mislevy, R. J. (2007). Cognitive foundations of structured item response theory models. In J. Leighton & M. Gierl (Hrsg.), *Cognitive diagnostic assessment in education: theory and applications*. Cambridge: Cambridge University Press, 205-241.

Sandmann, A. (2014). Lautes Denken – die Analyse von Denk-, Lern- und Problemlöseprozessen. In D. Krüger, I. Parchmann & H. Schecker (Hrsg.), *Methoden in der naturwissenschaftsdidaktischen Forschung*. Berlin, Heidelberg: Springer, 179-188.

Schecker, H., Nawrath, D., Elvers, H., Borgstädt, J., Einfeldt, S., Maiseyenka, V. (Hrsg.) (2013). *Modelle und Lernarrangements für die Förderung naturwissenschaftlicher Kompetenzen*. Hamburg: Landesinstitut für Lehrerbildung und Schulentwicklung.

Schecker, H., Neumann, K., Theyßen, H., Eickhorst, B. & Dickmann, M. (im Druck). Stufen experimenteller Kompetenz. *Zeitschrift für Didaktik der Naturwissenschaften,* im Druck.

Schmiemann, P., & Lücken, M. (2014). Validität – Misst mein Test, was er soll?. In D. Krüger, I. Parchmann & H. Schecker (Hrsg.), *Methoden in der naturwissenschaftsdidaktischen Forschung.* Berlin, Heidelberg: Springer, 107-118.

Schreiber, N., Theyßen, H., & Schecker, H. (2009). Experimentelle Kompetenz messen?!. *Physik und Didaktik in Schule und Hochschule*, *8 (*3), 92-101.

Schreiber, N. (2012). *Diagnostik experimenteller Kompetenz: Validierung technologiegestützter Testverfahren im Rahmen eines Kompetenzstrukturmodells* (Studien zum Physik- und Chemielernen, Band 139). Berlin: Logos.

Schreiber, N., Theyßen, H., & Schecker, H. (2014). Diagnostik experimenteller Kompetenz: Kann man Realexperimente durch Simulationen ersetzen?. *Zeitschrift für Didaktik der Naturwissenschaften*, *20*(1), 161-173.

Schreiber, N. & Theyßen, H. (2015). Experimentelle Fähigkeiten unterstützt durch Schülerselbstbeurteilungen diagnostizieren?. In S. Bernholt (Hrsg.), *Heterogenität und Diversität - Vielfalt der Voraussetzungen im naturwissenschaftlichen Unterricht. Gesellschaft für Didaktik der Chemie und Physik, Jahrestagung in Bremen 2014*. Kiel: IPN, 654-656.

Schreier, M. (2010). Fallauswahl. In G. Mey & K. Mruck. *Handbuch Qualitative Forschung in der Psychologie*. Wiesbaden: VS Verlag für Sozialwissenschaften, 238-251.

Schwamborn, A., Thillmann, H., Opfermann, M., & Leutner, D. (2011). Cognitive load and instructionally supported learning with provided and learner-generated visualizations. *Computers in Human Behavior*, *27*(1), 89-93.

Schwarz, I., Effertz, C., & Heinke, H. (2013). Entwicklung eines Physikpraktikums für Biologiestudierende-der Umgang mit Messunsicherheiten. *PhyDid B, Didaktik der Physik, Beiträge zur DPG-Frühjahrstagung 2013 in Jena*, 1-5.

Shavelson, R. J., Baxter, G., Pine, J. (1991). Performance Assessment in Science. *Applied Measurement in Education, 4(4)*, 347-362.

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., 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. (2010). On the measurement of competency. *Empirical Research in Vocational Education and Training*, *2*(1), 41-63.

Shepard, L. A. (1993). Evaluating test validity. *Review of research in education*, *19*, 405-450.

Sireci, S. G. (1998). The construct of content validity. *Social indicators research*, *45*(1-3), 83-117.

Stebler, R., Reusser, K., & Ramseier, E. (1998). Praktische Anwendungsaufgaben zur integrierten Förderung formaler und materialer Kompetenzen: Erträge aus dem TIMSS-Experimentiertest. *Bildungsforschung und Bildungspraxis*, *20* (1), 28-54.

Stecher, B. M., & Klein, S. P. (1997). The cost of science performance assessments in large-scale testing programs. *Educational Evaluation and Policy Analysis*, *19*(1), 1-14.

Stobart, G. (2009). Determining validity in national curriculum assessments. *Educational Research*, *51*(2), 161-179.

Sweller, J., Ayres, J. & Kalyuga, S. (2011). *Cognitive Load Theory*. New York: Springer.

Tesch, M. & Duit, R. (2004). Experimentieren im Physikunterricht – Ergebnisse einer Videostudie. *Zeitschrift für Didaktik der Naturwissenschaften, 10*, 51-69.

Theyßen, H., Schecker, H., Neumann, K., Dickmann, M., & Eickhorst, B. (2013). Messung experimenteller Kompetenz in Large Scale Assessments. In S. Bernholt (Hrsg.), *Inquiry-based Learning – Forschendes Lernen. Gesellschaft für Didaktik der Chemie und Physik, Jahrestagung in Hannover 2012.* Kiel: IPN, 596-598.

Theyßen, H., Schecker, H., Dickmann, M., Eickhorst, B. und Neumann, K. (2016a). Messung experimenteller Kompetenz in Large-Scale-Assessments (MeK-LSA). In Bundesministerium für Bildung und Forschung (Hrsg.), *Forschungsvorhaben in Ankopplung an Large-Scale-Assessments.* Berlin: Bundesministerium für Bildung und Forschung, Bildungsforschung Band 44, 83-96.

Theyßen, H., Schecker, H., Neumann, K., Eickhorst, B. & Dickmann, M. (2016b). Messung experimenteller Kompetenz – ein computergestützter Experimentiertest. *Physik und Didaktik in Schule und Hochschule*, *15 (*1).

Türck, V. (2014). *Womit beschäftigen sich Schülerinnen und Schüler bei der Bearbeitung einer Experimentieraufgabe? (Bachelorarbeit)*. Universität Duisburg-Essen.

van Gog, T., & Jarodzka, H. (2013). Eye tracking as a tool to study and enhance cognitive and metacognitive processes in computer-based learning environments. In R. Azevedo & V. Aleven (Hrsg.), *International handbook of metacognition and learning technologies*. New York: Springer, 143-156.

van Gog, T., Kirschner, F., Kester, L., & Paas, F. (2012). Timing and frequency of mental effort measurement: Evidence in favour of repeated measures. *Applied Cognitive Psychology*, *26*(6), 833-839.

Vollstädt, W., Tillmann, K. J., Rauin, U., Höhmann, K. & Tebrügge, A. (1999). *Lehrpläne im Schulalltag. Eine empirische Studie zur Akzeptanz und Wirkung von Lehrplänen in der Sekundarstufe I.* Opladen: Leske + Budrich.

Walpuski, M. (2006). *Optimierung von experimenteller Kleingruppenarbeit durch Strukturierungshilfen und Feedback: eine empirische Studie* (Studien zum Physik- und Chemielernen, Band 49). Berlin: Logos.

Weir, C. J. (2005). *Language testing and validation*. UK: Macmillan.

Wellnitz, N. (2012). *Kompetenzstruktur und -niveaus von Methoden naturwissenschaftlicher Erkenntnisgewinnung*. Berlin: Logos.

Wellnitz, N., Fischer, H. E., Kauertz, A., Mayer, J., Neumann, I., Pant, H. A., Sumfleth, E. & Walpuski, M. (2012). Evaluation der Bildungsstandards – eine fächerübergreifende Testkonzeption für den Kompetenzbereich Erkenntnisgewinnung. *Zeitschrift für Didaktik der Naturwissenschaften*, 18, 261-291.

Wendt, H., & Bos, W. (2011). Fachdidaktik und Bildungsforschung – von der Notwendigkeit zur Kooperation im Zeitalter globalisierter Kompetenzen. In K. O. Bauer & N. Logemann (Hrsg.), *Unterrichtsqualität und fachdidaktische Forschung – Modelle und Instrumente zur Messung fachspezifischer Lernbedingungen und Kompetenzen*. Münster: Waxmann, 11-34.

Wilhelm, O., & Kunina, O. (2009). Pädagogisch-psychologische Diagnostik. In E. Wild & J. Möller (Hrsg.), *Pädagogische Psychologie.* Heidelberg: Springer, 307-331.

Wirtz, M. A., & Caspar, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität: Methoden zur Bestimmung und Verbesserung der Zuverlässigkeit von Einschätzungen mittels Kategoriensystemen und Ratingskalen*. Göttingen: Hogrefe.

Wolming, S., & Wikström, C. (2010). The concept of validity in theory and practice. *Assessment in Education: Principles, Policy & Practice*, *17*(2), 117-132.

Woolfolk, A. (2008). *Pädagogische Psychologie* (10. Auflage). München: Pearson.

Zeitler, S., Asbrand, B., & Heller, N. (2013). Steuerung durch Bildungsstandards– Bildungsstandards als Innovation zwischen Implementation und Rezeption. In I. Bormann & M. Rürup (Hrsg.), *Innovationen im Bildungswesen*. Springer: Wiesbaden, 127-147.

Zirwes, S. (2014). *Womit beschäftigen sich Schülerinnen und Schüler bei der Bearbeitung eines Online-Experimentiertests? (Staatsexamensarbeit)*. Universität Duisburg-Essen.

Abbildungsverzeichnis

Tabellenverzeichnis

Anhang

A.1 Testaufgabe zur Ausdehnung eines Gummibandes

Die folgenden Seiten zeigen Bildschirmkopien der Aufgabe zur Ausdehnung eines Gummibandes.[28] Die Aufgabenstellung lautet:

> "Alina und Bodo wollen untersuchen, wie sich ein Gummiband ausdehnt, wenn man verschiedene Gewichte daran hängt. Die beiden erwarten, dass die Ausdehnung des Gummibandes im gleichen Maße zunimmt, wie das angehängte Gewicht.
>
> Physikalisch könnte man ihre Vermutung so formulieren: "Die Ausdehnung des Gummibands ist proportional zum angehängten Gewicht.
>
> Du sollst jetzt Alina und Bodo dabei helfen ihre Vermutung zu überprüfen! Alina und Bodo führen das Experiment ebenfalls durch. Du wirst zwischendurch sehen, wie sie dabei vorgehen."

Die Schüler müssen

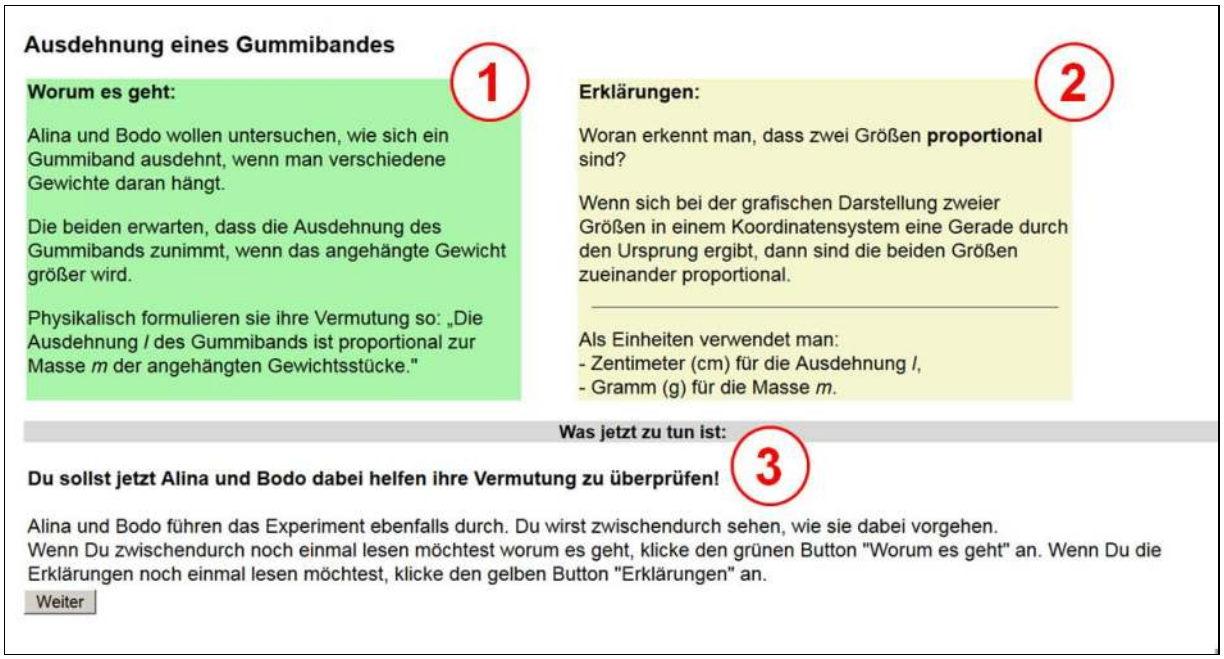
a) die zu messenden Größen benennen
b) **den Versuch planen (Geräte auswählen, eine Versuchsskizze erstellen, die Durchführung stichwortartig beschreiben**
c) **den Versuch funktionsfertig aufbauen**
d) **Messungen durchführen und protokollieren**
e) die Vorgehensweise bei der Datenauswertung beschreiben
f) **die Daten auswerten (grafisch darstellen)**
g) eine Schlussfolgerung mit Bezug auf die zu prüfenden Hypothese ziehen

   (Es zeigt sich, dass man durch die Messpunkte keine Gerade legen kann. Der Zusammenhang zwischen Zugkraft und Dehnung ist demnach weder proportional noch linear.)

Die in **Fettdruck** hervorgehobenen Teilaufgaben sind in der vorliegenden Dissertation bei den Analysen berücksichtigt worden.

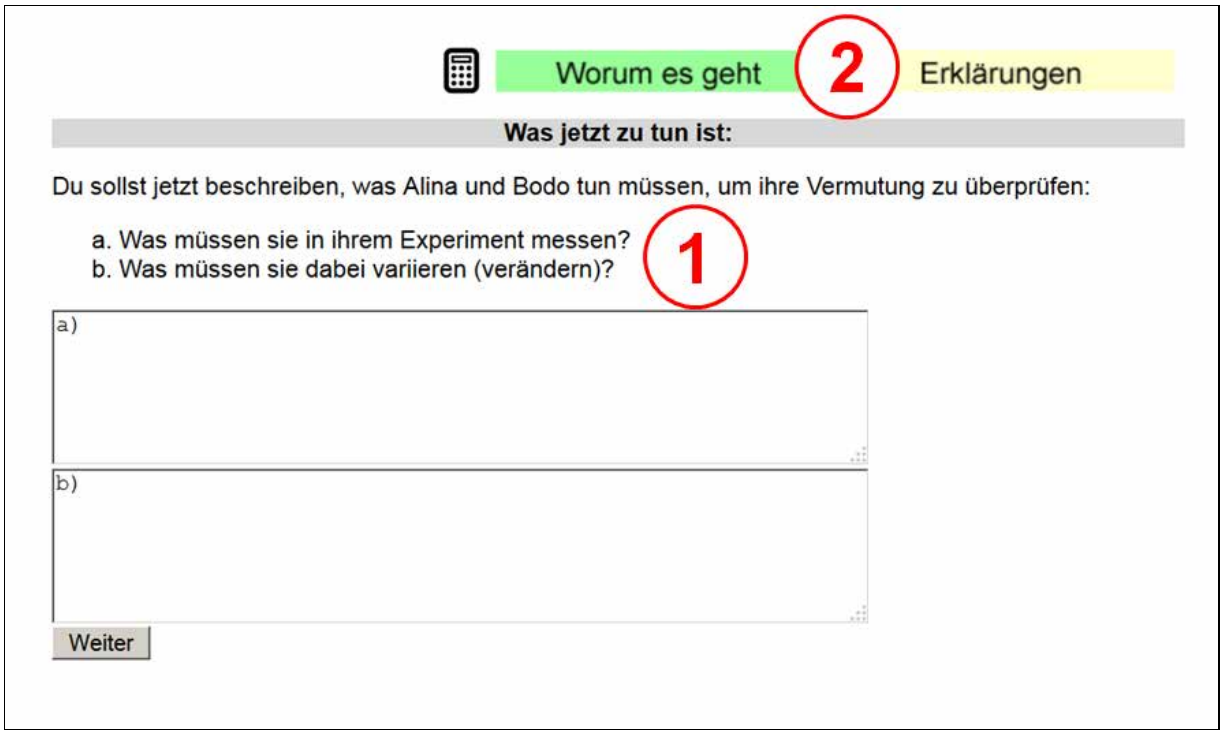
[28] Die Bildschirmkopien sind dem Abschlussbericht zum Projekt *Messung experimenteller Kompetenz in Large-Scale Assessments* entnommen.

Seite 1: Einführung in die Aufgabe:
Vorgegeben sind (1) übergeordnete Aufgabenstellung, (2) fachliche Erklärung, (3) Einführung von „Alina und Bodo“.



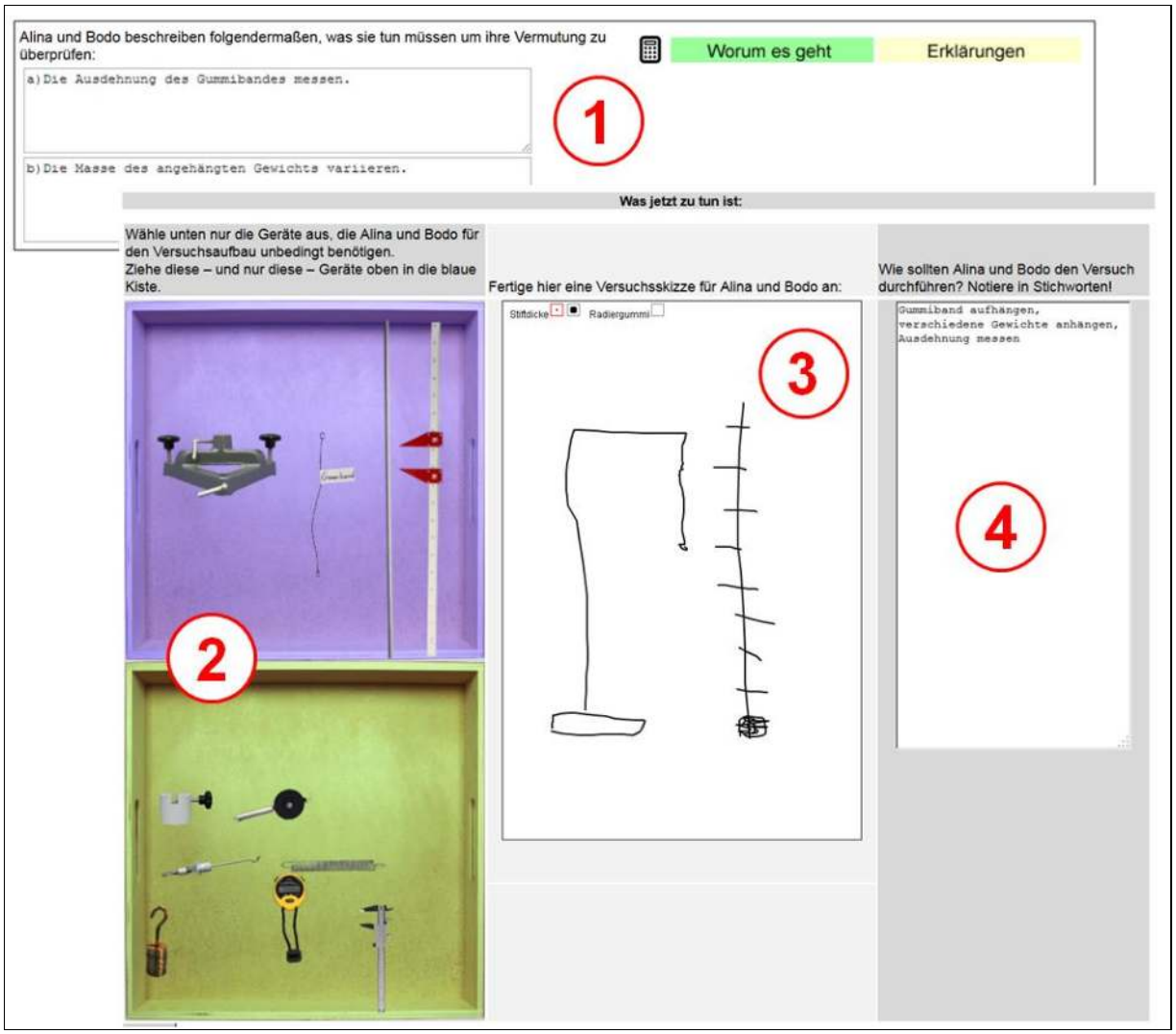
Seite 2: Teilaufgabe „Grundidee skizzieren“

Die Schülerinnen und Schüler sollen angeben, was Alina und Bodo messen müssen und was sie dabei konstant halten sollen (1). Übergeordnete Aufgabe und fachliche Erklärungen können dabei abgerufen werden (2).



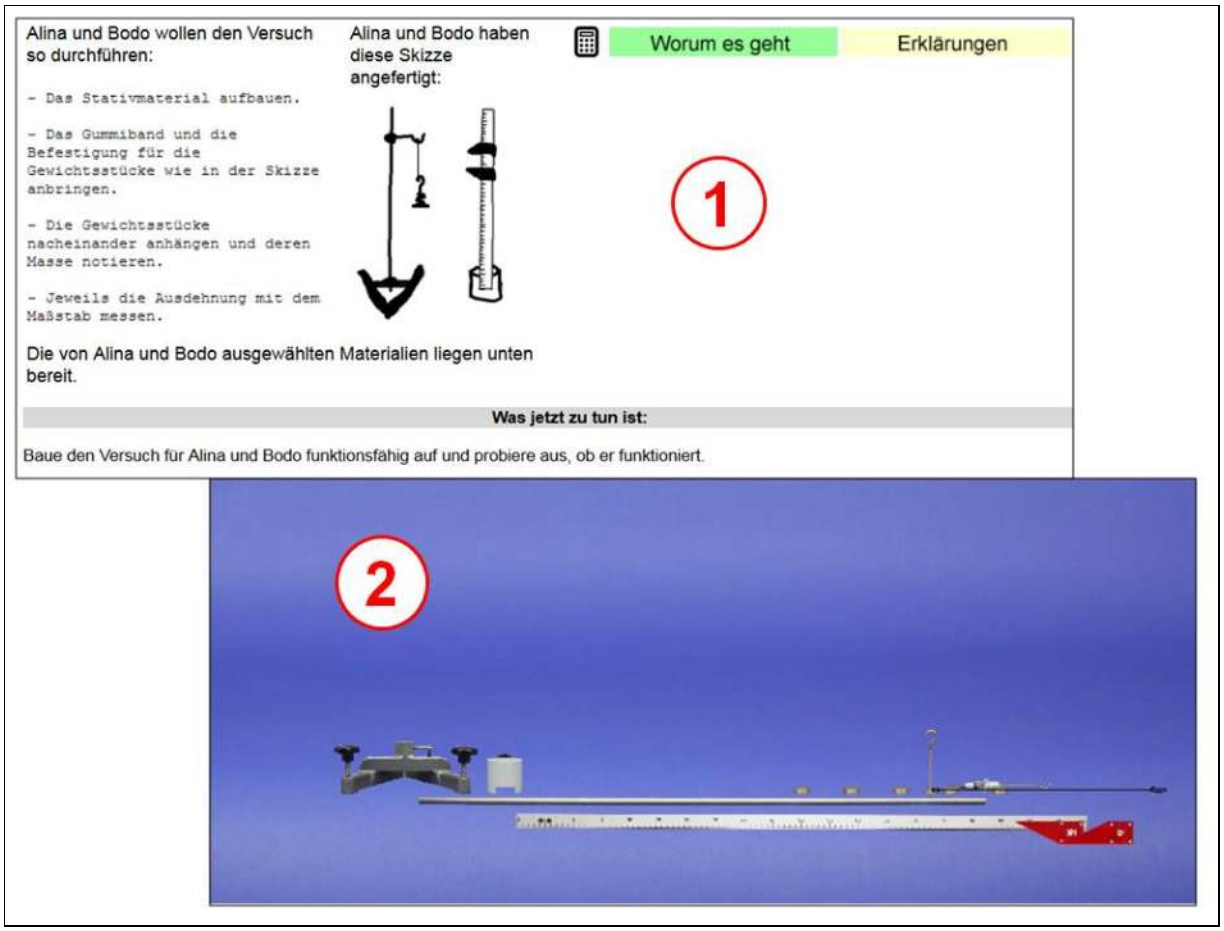
Seite 3: Teilaufgabe „Versuchsplan entwerfen“

Vorgegeben ist die Grundidee (1). Die Schülerinnen und Schüler sollen das benötigte Material auswählen (2), die Aufbau skizzieren (3) und die Vorgehensweise in Stichworten beschreiben (4).



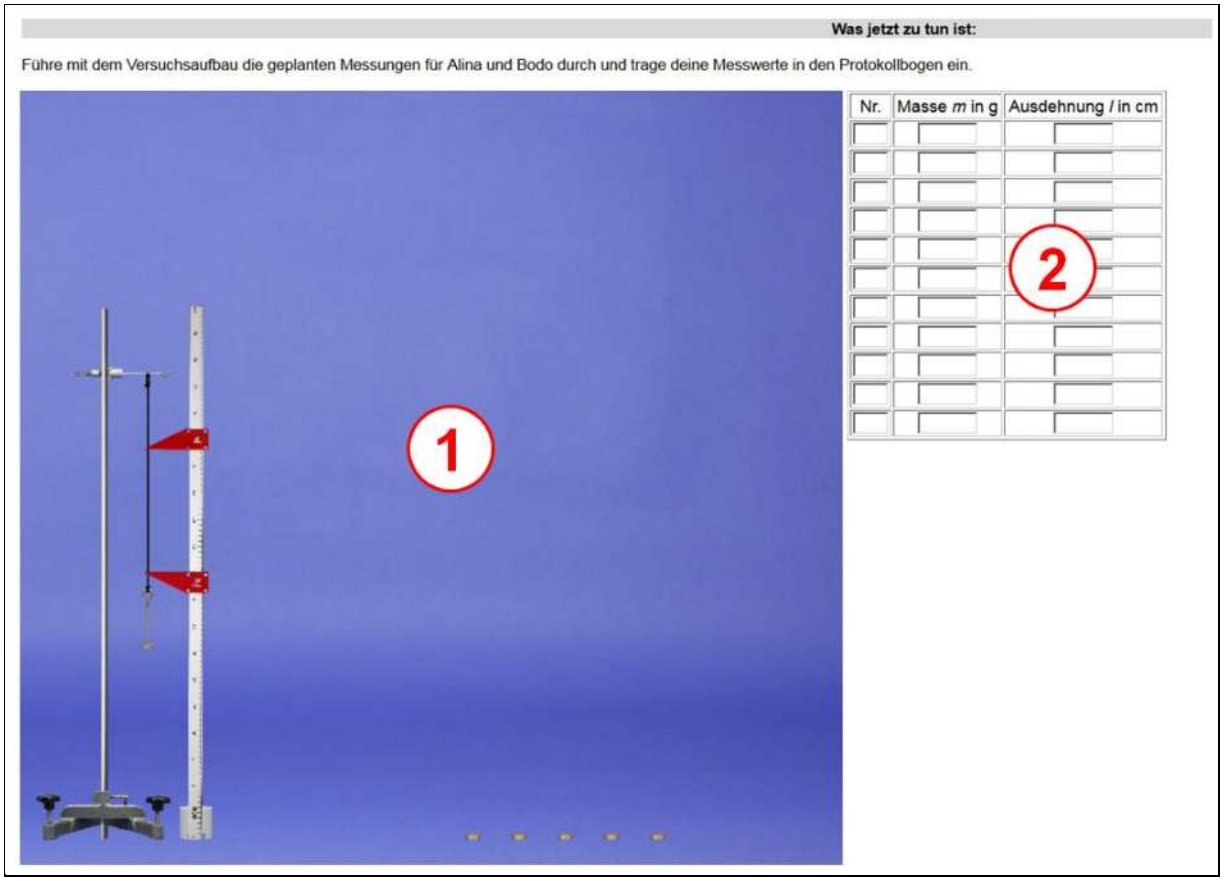
Seite 4: Teilaufgabe „Versuch aufbauen und testen“

Die Geräte, eine Skizze und die Beschreibung der Vorgehensweise sind vorgegeben (1). Die Schülerinnen und Schüler sollen nun den Versuch aufbauen und testen (2).



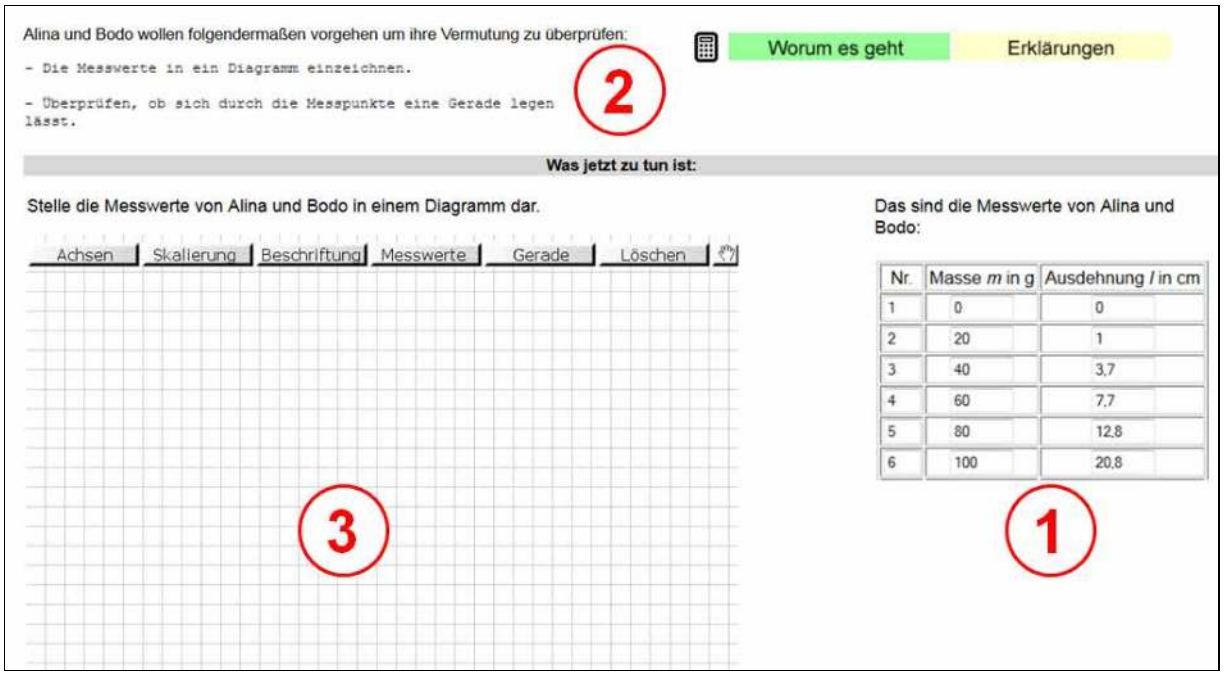
Seite 5: Teilaufgabe „Messung durchführen und dokumentieren“

Aufbau (1) und Tabelle (2) sind für die Messung vorgegeben.



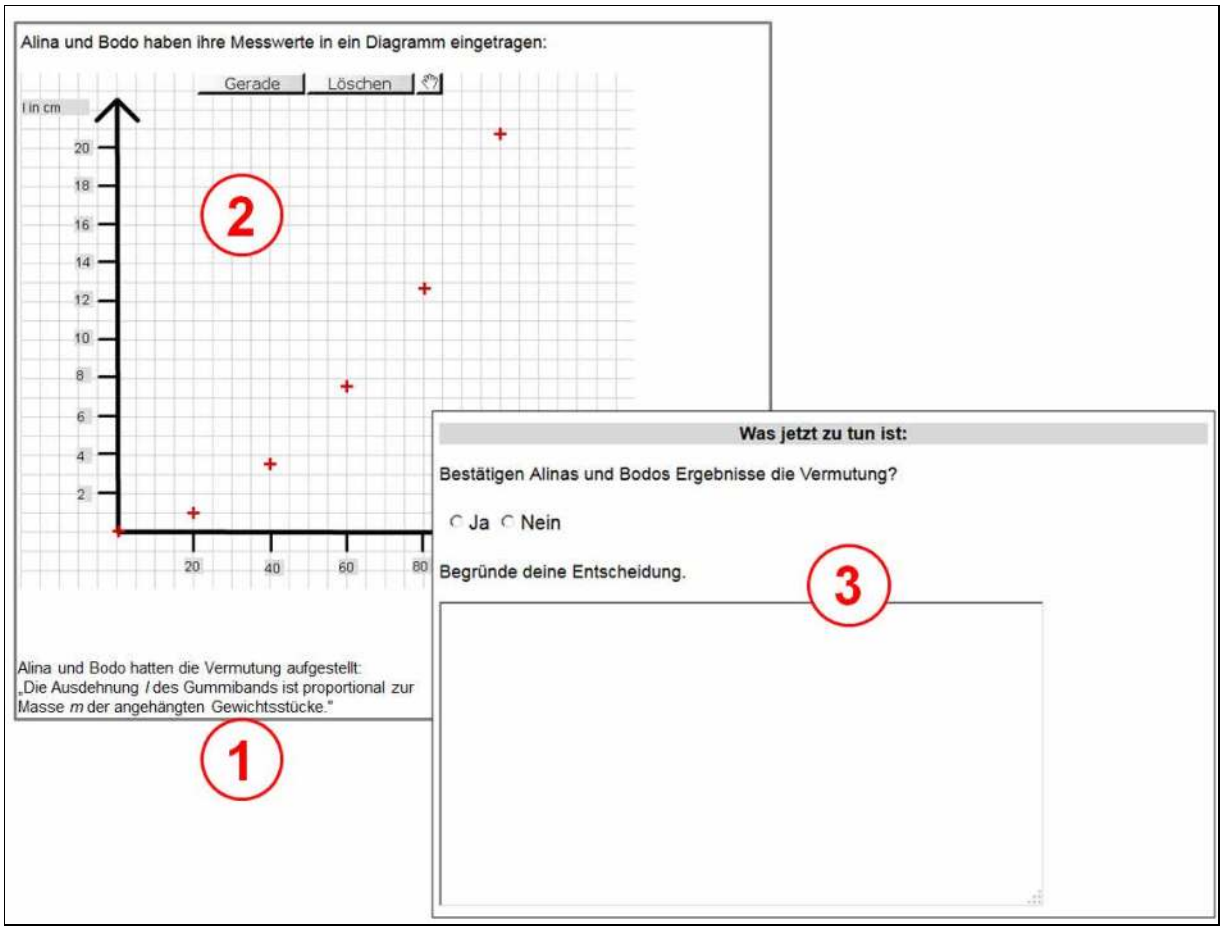
Seite 6: Teilaufgabe „Datenauswertung durchführen“

Die Schülerinnen und Schüler sollen die vorgegebenen Messwerte (1) auf vorgegebene Weise (2) auswerten, indem sie ein Diagramm erstellen (3).



Seite 7: Teilaufgabe „Schlüsse ziehen“

Die Vermutung aus dem Aufgabenstamm (1) und das Diagramm (2) sind vorgegeben. Die Schülerinnen und Schüler sollen entscheiden und begründen, ob die Ergebnisse die Vermutung stützen oder nicht (3).



A.2 Übersicht über die analysierten Lehrpläne

Bundesland	Bezeichnung	analysierte Fächer	analysierte Schulform
Baden-Württemberg	Bildungsplan	Naturphänomene Physik	Gymnasium Gymnasium
Bayern	Lehrplan	Natur und Technik Physik	Gymnasium Gymnasium
Berlin	Rahmenlehrplan	Naturwissenschaften Physik	Grundschule Gymnasium
Brandenburg	Rahmenlehrplan	Naturwissenschaften Physik	Grundschule Gymnasium
Bremen	Bildungsplan	Naturwissenschaften Physik	Gymnasium Gymnasium
Hamburg	Bildungsplan	Naturwissenschaften/Technik Physik	Gymnasium Gymnasium
Hessen	Lehrplan Kerncurriculum	Physik	Gymnasium
Mecklenburg-Vorpommern	Rahmenlehrplan	Physik	Gymnasium Int. Gesamtschule[29]
Niedersachsen	Kerncurriculum	Physik	Gymnasium
Nordrhein-Westfalen	Kernlehrplan	Physik	Gymnasium
Rheinland-Pfalz	Rahmenlehrplan/ Lehrplan	Naturwissenschaften Physik	HRGyGe[30] Gymnasium
Saarland	Lehrplan	Naturwissenschaften Physik	Gymnasium Gymnasium
Sachsen	Lehrplan	Physik	Gymnasium
Sachsen-Anhalt	Rahmenrichtlinien	Physik	Gymnasium
Schleswig-Holstein	Lehrplan Orientierungshilfen G8	Physik	Haupt- und Realschule sowie Gymnasium
Thüringen	Lehrplan	Mensch-Natur-Technik Physik	Gymnasium Gymnasium

[29] Integrierte Gesamtschule
[30] HRGyGe: Hauptschule, Realschule, Gymnasium und Gesamtschule

A.3 identifizierte Unterrichtsthemen

Inhaltsbereich	Unterrichtsthemen	Inhaltsbereich	Unterrichtsthemen
Mechanik	**Kraft und Verformung**	Optik	**Brechungsgesetz**
	Kraftwandler		**Reflexionsgesetz**
	Hebelgesetz		Totalreflexion
	Flaschenzug		**Abbildungen an Linsen**
	schiefe Ebene		Eigenschaften von Spiegelbildern
	Schweredruck in Flüssigkeiten		Lochkamera
	Zusammensetzung von Kräften		Zerlegung von weißem Licht
	hydrostatisches Paradoxon		Farbmischung
	Reibungskraft	Elektrizitätslehre	**elektrischer Widerstand**
	Auftrieb in Flüssigkeiten		**Reihenschaltung**
	Dichtebestimmung		**Parallelschaltung**
	Bewegungen		**elektrische Leistung**
	freier Fall		**Leiter und Nichtleiter**
	mechanische Arbeit	Akustik	Schallausbreitung
Wärmelehre	**Aggregatzustandsänderungen**		Schallentstehung
	Wärmeübertragung		Schallarten
	Wärmeleitung		Schallquellen
	Wärmestrahlung		Laustärke
	Konvektion		Tonhöhe
	Wärmeausdehnung		Schallreflexion
	Volumenausdehnung		Schallgeschwindigkeit
	Längenausdehnung		Hörbereich
	spezifische Wärmekapazität		Ultraschall und Infraschall

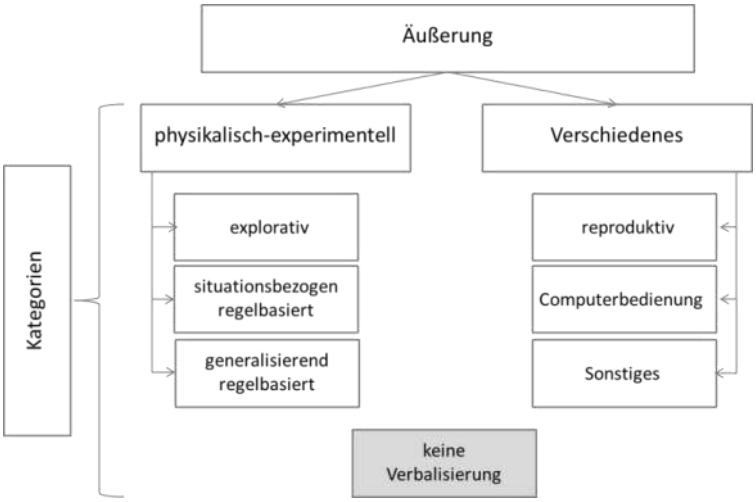
Die in **Fettdruck** hervorgehobenen Unterrichtsthemen wurden als relevant identifiziert (vgl. Tabelle 4.2 auf Seite 40).

A.4 Ergebnisse der Lehrkräftebefragung (tabellarisch)

Kurzbeschreibung	Bekanntheit				Anforderungen					
	Gesehen?		Durchgeführt?		Planbarkeit?		Durchführbarkeit?		Auswertbarkeit?	
	MW	SD	MW	SD	MW	SD	MW	SD	MW	SD
U-I Kennlinie einer Glühlampe	3,2	0,9	3,2	0,8	2,9	0,8	3,3	0,7	3,3	0,7
Reihenschaltung einer Glühlampe	3,3	0,9	3,0	0,7	3,2	0,7	3,3	0,7	3,4	0,6
Parallelschaltung einer Glühlampe	3,2	0,8	3,0	0,8	3,0	0,9	3,1	0,7	3,2	0,7
Widerstand eines Drahtstückes (Abh. Länge)	3,0	1,0	2,5	0,8	2,7	0,7	3,1	0,7	3,0	0,7
Widerstand eines Drahtstückes (Abh. Material)	2,6	0,8	2,2	0,7	2,7	0,8	2,9	0,7	2,9	0,7
Leistung von Glühlampen	2,8	0,9	2,8	0,9	2,7	0,8	2,8	0,6	2,8	0,6
Widerstand eines Drahtstückes (Abh. Durchmesser)	2,3	0,8	2,0	0,8	2,4	0,8	2,8	0,7	2,6	0,9
Reflexion am Plexiglasblock	3,0	1,1	2,9	1,1	3,2	0,8	3,4	0,7	3,2	0,9
Brennweite einer Linse bestimmen	3,2	0,9	2,9	0,9	3,0	0,7	3,2	0,7	3,3	0,5
Lichtbrechung am Halbkreisblock	3,0	1,0	3,0	1,1	3,0	0,9	3,0	0,8	2,9	0,8
Abbildungsmaßstab	2,8	1,0	2,7	1,0	2,7	0,7	2,9	0,8	3,0	0,7
Abbildungsgleichung	2,9	0,9	2,7	1,0	2,4	1,0	2,8	0,8	2,4	0,8
Bild- und Gegenstandsweite	3,1	0,9	2,8	1,0	2,7	0,9	2,8	0,8	2,9	0,7
Reflexion am Wölbspiegel	2,0	0,8	1,9	0,9	2,4	0,9	2,7	0,7	2,6	0,9
Lochkamera	2,4	1,0	2,3	1,1	2,4	0,9	2,7	0,8	2,6	0,8
Ausdehnung eines Gummibandes	2,7	1,0	2,3	0,9	3,0	0,7	3,3	0,5	2,8	0,7
Dichtebestimmung	2,7	1,2	2,7	1,0	2,8	0,9	3,1	0,7	3,2	0,6
Auftriebskraft in Wasser	2,5	1,0	2,4	1,0	2,6	0,8	3,0	0,7	2,7	0,8
Bewegung auf der schiefen Ebene	2,8	1,0	2,2	0,9	2,9	0,8	2,9	0,9	2,7	0,9
Spezifische Wärmekapazität	3,0	1,0	2,6	1,0	2,8	0,7	3,3	0,6	2,8	0,6
Wärmestrahlung	2,2	1,1	2,0	1,0	2,9	0,8	3,3	0,5	3,2	0,6
Wärmeleitfähigkeit	2,0	0,9	1,6	0,8	2,5	1,0	3,0	0,7	2,8	0,7

A.5 Kodiermanual zur Kategorisierung der Daten aus dem begleitenden Think-Aloud

Im Folgenden wird nur der prinzipielle Aufbau des Kodiermanuals veranschaulicht. Das vollständige Kodiermanual kann auf Anfrage beim Autor dieser Dissertation (Kontakt: martin.dickmann@uni-due.de) angefordert werden.



Die Einteilung in die Kategorien erfolgt zeitbasiert, d. h. in **10 Sekunden-Intervallen**. Treten in einem Zeitintervall mehrere Kategorien auf die kodiert werden müssen, so wird die Kategorie kodiert, die den größeren zeitlichen Anteil in dem Zeitintervall einnimmt. Ausnahme: Die Kategorie *keine Verbalisierung* wird nur zugewiesen, wenn in einem kompletten Intervall keine Schüleräußerung vorliegt.

Bei den **physikalisch-experimentellen Äußerungen** werden in Anlehnung an von Aufschnaiter & Rogge (2010) drei Kategorien (Niveaus) unterschieden:

Kategorie	Beschreibung: Äußerungen…
explorativ	sind gekennzeichnet durch eine fehlende Sicherheit bzw. eine fehlende Systematik.
situationsbezogen regelbasiert	beziehen sich konkret auf die vorliegende (Experimentier-) Situation. Vorhandenes Wissen wird bezogen auf diese Situation angewendet.
generalisierend regelbasiert	beziehen sich nicht auf die konkret vorliegende (Experimentier-) Situation. Vorhandenes Wissen wird oder Erkenntnisse aus dem konkreten Experiment werden generalisiert.

Die Oberkategorie **Verschiedenes** umfasst alle Äußerungen, die keiner physikalisch-experimentellen Äußerung zuzuordnen sind:

Kategorie	Beschreibung: Äußerungen…
reproduktiv	wie das Vorlesen/Wiedergeben der Aufgabenstellung oder Mitsprechen der Antworten bei der Eingabe.
Computerbedienung	zur Bedienung der on-screen Testumgebung (z. B. interaktive Simulationen)
Sonstiges	die keiner anderen Kategorie eindeutig zuzuordnen sind.

Bei der Kodierung der Zeitintervalle sind in jedem Fall die ausführlich beschriebenen Indikatoren und Ankerbeispiele zu berücksichtigen.[31]

[31] Die ausführlich beschriebenen Indikatoren und Ankerbeispiele sind hier nicht aufgeführt, werden aber auf Anfrage (siehe oben) zur Verfügung gestellt.

A.6 Skript zur Trainingsaufgabe

Bei Interesse am Skript zur Trainingsaufgabe wenden Sie sich bitte per Mail an:

Martin Dickmann

Mail: martin.dickmann@uni-due.de

Danksagung

Für die Ermöglichung und die sehr hilfreiche Unterstützung bei der Erstellung meiner Arbeit möchte ich mich bei allen Beteiligten ganz herzlich bedanken. Mein besonderer Dank gilt:

Frau Prof. Dr. Heike Theyßen für die Begutachtung meiner Arbeit, die herausfordernde Themenstellung und die herausragende Betreuung während meiner Promotion.

Herrn Prof. Dr. Horst Schecker für die Begutachtung meiner Arbeit und die hervorragende Beratung und Unterstützung während meiner Promotion.

Herrn Prof. Dr. Roger Erb für die Begutachtung meiner Arbeit.

Herrn Prof. Dr. Philipp Schmiemann für die Mitgliedschaft in der Prüfungskommission.

Herrn Prof. Dr. Andreas Wucher für die Übernahme des Prüfungsvorsitzes.

Herrn Prof. Dr. Knut Neumann für die wertvollen Ratschläge und Anmerkungen zu meiner Arbeit während und außerhalb unserer Projekttreffen.

Bodo Eickhorst für die intensive und gemeinsame Arbeit an unserem Projekt. Ohne unsere enge Zusammenarbeit wäre das Entstehen meiner Arbeit nicht möglich gewesen.

Dem BMBF für die Finanzierung des Forschungsvorhabens.

Nico Schreiber für die sehr hilfreiche Beratung in allen Phasen der Arbeit. Deine Ideen und Anmerkungen haben mir immer wieder innovierende Impulse für meine Arbeit gegeben.

Britta Kalthoff und Alexander Pusch für eure Unterstützung während meiner Promotion und die zahlreichen fachlichen und überfachlichen Gespräche die meine Arbeit kontinuierlich vorangebracht haben.

Allen studentischen Hilfskräften für die Unterstützung während der Datenerhebung und Datenauswertung.

Allen Studierenden, die mit Ihrer Abschlussarbeit einen wertvollen Beitrag zum Gelingen meiner Arbeit geleistet haben.

Den teilnehmenden Schülerinnen und Schülern sowie Studierenden, die sich intensiv mit den Experimentieraufgaben auseinandergesetzt haben.

Meinen Eltern und meiner Schwester mit ihrer Familie, die mich zu jeder Zeit auf meinem Weg uneingeschränkt unterstützt haben.

Meiner Frau Julia für die liebevolle, motivierende Unterstützung und die Bereitschaft auf sehr viel gemeinsame Zeit während meiner Promotion zu verzichten.

Bisher erschienene Bände der Reihe *„Studien zum Physik- und Chemielernen“*

ISSN 1614-8967 (vormals *Studien zum Physiklernen* ISSN 1435-5280)

1 Helmut Fischler, Jochen Peuckert (Hrsg.): Concept Mapping in fachdidaktischen Forschungsprojekten der Physik und Chemie
ISBN 978-3-89722-256-4 40.50 EUR

2 Anja Schoster: Bedeutungsentwicklungsprozesse beim Lösen algorithmischer Physikaufgaben. *Eine Fallstudie zu Lernprozessen von Schülern im Physiknachhilfeunterricht während der Bearbeitung algorithmischer Physikaufgaben*
ISBN 978-3-89722-045-4 40.50 EUR

3 Claudia von Aufschnaiter: Bedeutungsentwicklungen, Interaktionen und situatives Erleben beim Bearbeiten physikalischer Aufgaben
ISBN 978-3-89722-143-7 40.50 EUR

4 Susanne Haeberlen: Lernprozesse im Unterricht mit Wasserstromkreisen. *Eine Fallstudie in der Sekundarstufe I*
ISBN 978-3-89722-172-7 40.50 EUR

5 Kerstin Haller: Über den Zusammenhang von Handlungen und Zielen. *Eine empirische Untersuchung zu Lernprozessen im physikalischen Praktikum*
ISBN 978-3-89722-242-7 40.50 EUR

6 Michaela Horstendahl: Motivationale Orientierungen im Physikunterricht
ISBN 978-3-89722-227-4 50.00 EUR

7 Stefan Deylitz: Lernergebnisse in der Quanten-Atomphysik. *Evaluation des Bremer Unterrichtskonzepts*
ISBN 978-3-89722-291-5 40.50 EUR

8 Lorenz Hucke: Handlungsregulation und Wissenserwerb in traditionellen und computergestützten Experimenten des physikalischen Praktikums
ISBN 978-3-89722-316-5 50.00 EUR

9 Heike Theyßen: Ein Physikpraktikum für Studierende der Medizin. *Darstellung der Entwicklung und Evaluation eines adressatenspezifischen Praktikums nach dem Modell der Didaktischen Rekonstruktion*
ISBN 978-3-89722-334-9 40.50 EUR

10 Annette Schick: Der Einfluß von Interesse und anderen selbstbezogenen Kognitionen auf Handlungen im Physikunterricht. *Fallstudien zu Interessenhandlungen im Physikunterricht*
ISBN 978-3-89722-380-6 40.50 EUR

11 Roland Berger: Moderne bildgebende Verfahren der medizinischen Diagnostik. *Ein Weg zu interessanterem Physikunterricht*
ISBN 978-3-89722-445-2 40.50 EUR

12 Johannes Werner: Vom Licht zum Atom. *Ein Unterrichtskonzept zur Quantenphysik unter Nutzung des Zeigermodells*
ISBN 978-3-89722-471-1 40.50 EUR

13 Florian Sander: Verbindung von Theorie und Experiment im physikalischen Praktikum. *Eine empirische Untersuchung zum handlungsbezogenen Vorverständnis und dem Einsatz grafikorientierter Modellbildung im Praktikum*
ISBN 978-3-89722-482-7 40.50 EUR

14 Jörn Gerdes: Der Begriff der physikalischen Kompetenz. *Zur Validierung eines Konstruktes*
ISBN 978-3-89722-510-7 40.50 EUR

15 Malte Meyer-Arndt: Interaktionen im Physikpraktikum zwischen Studierenden und Betreuern. *Feldstudie zu Bedeutungsentwicklungsprozessen im physikalischen Praktikum*
ISBN 978-3-89722-541-1 40.50 EUR

16 Dietmar Höttecke: Die Natur der Naturwissenschaften historisch verstehen. *Fachdidaktische und wissenschaftshistorische Untersuchungen*
ISBN 978-3-89722-607-4 40.50 EUR

17 Gil Gabriel Mavanga: Entwicklung und Evaluation eines experimentell- und phänomenorientierten Optikcurriculums. *Untersuchung zu Schülervorstellungen in der Sekundarstufe I in Mosambik und Deutschland*
ISBN 978-3-89722-721-7 40.50 EUR

18 Meike Ute Zastrow: Interaktive Experimentieranleitungen. *Entwicklung und Evaluation eines Konzeptes zur Vorbereitung auf das Experimentieren mit Messgeräten im Physikalischen Praktikum*
ISBN 978-3-89722-802-3 40.50 EUR

19 Gunnar Friege: Wissen und Problemlösen. *Eine empirische Untersuchung des wissenszentrierten Problemlösens im Gebiet der Elektrizitätslehre auf der Grundlage des Experten-Novizen-Vergleichs*
ISBN 978-3-89722-809-2 40.50 EUR

20 Erich Starauschek: Physikunterricht nach dem Karlsruher Physikkurs. *Ergebnisse einer Evaluationsstudie*
ISBN 978-3-89722-823-8 40.50 EUR

21 Roland Paatz: Charakteristika analogiebasierten Denkens. *Vergleich von Lernprozessen in Basis- und Zielbereich*
ISBN 978-3-89722-944-0 40.50 EUR

22 Silke Mikelskis-Seifert: Die Entwicklung von Metakonzepten zur Teilchenvorstellung bei Schülern. *Untersuchung eines Unterrichts über Modelle mithilfe eines Systems multipler Repräsentationsebenen*
ISBN 978-3-8325-0013-9 40.50 EUR

23 Brunhild Landwehr: Distanzen von Lehrkräften und Studierenden des Sachunterrichts zur Physik. *Eine qualitativ-empirische Studie zu den Ursachen*
ISBN 978-3-8325-0044-3 40.50 EUR

24 Lydia Murmann: Physiklernen zu Licht, Schatten und Sehen. *Eine phänomenografische Untersuchung in der Primarstufe*
ISBN 978-3-8325-0060-3 40.50 EUR

25 Thorsten Bell: Strukturprinzipien der Selbstregulation. *Komplexe Systeme, Elementarisierungen und Lernprozessstudien für den Unterricht der Sekundarstufe II*
ISBN 978-3-8325-0134-1 40.50 EUR

26 Rainer Müller: Quantenphysik in der Schule
ISBN 978-3-8325-0186-0 40.50 EUR

27 Jutta Roth: Bedeutungsentwicklungsprozesse von Physikerinnen und Physikern in den Dimensionen Komplexität, Zeit und Inhalt
ISBN 978-3-8325-0183-9 40.50 EUR

28 Andreas Saniter: Spezifika der Verhaltensmuster fortgeschrittener Studierender der Physik
ISBN 978-3-8325-0292-8 40.50 EUR

29 Thomas Weber: Kumulatives Lernen im Physikunterricht. *Eine vergleichende Untersuchung in Unterrichtsgängen zur geometrischen Optik*
ISBN 978-3-8325-0316-1 40.50 EUR

30 Markus Rehm: Über die Chancen und Grenzen moralischer Erziehung im naturwissenschaftlichen Unterricht
ISBN 978-3-8325-0368-0 40.50 EUR

31 Marion Budde: Lernwirkungen in der Quanten-Atom-Physik. *Fallstudien über Resonanzen zwischen Lernangeboten und SchülerInnen-Vorstellungen*
ISBN 978-3-8325-0483-0 40.50 EUR

32 Thomas Reyer: Oberflächenmerkmale und Tiefenstrukturen im Unterricht. *Exemplarische Analysen im Physikunterricht der gymnasialen Sekundarstufe*
ISBN 978-3-8325-0488-5 40.50 EUR

33 Christoph Thomas Müller: Subjektive Theorien und handlungsleitende Kognitionen von Lehrern als Determinanten schulischer Lehr-Lern-Prozesse im Physikunterricht
ISBN 978-3-8325-0543-1 40.50 EUR

34 Gabriela Jonas-Ahrend: Physiklehrervorstellungen zum Experiment im Physikunterricht
ISBN 978-3-8325-0576-9 40.50 EUR

35 Dimitrios Stavrou: Das Zusammenspiel von Zufall und Gesetzmäßigkeiten in der nichtlinearen Dynamik. *Didaktische Analyse und Lernprozesse*
ISBN 978-3-8325-0609-4 40.50 EUR

36 Katrin Engeln: Schülerlabors: authentische, aktivierende Lernumgebungen als Möglichkeit, Interesse an Naturwissenschaften und Technik zu wecken
ISBN 978-3-8325-0689-6 40.50 EUR

37 Susann Hartmann: Erklärungsvielfalt
ISBN 978-3-8325-0730-5 40.50 EUR

38 Knut Neumann: Didaktische Rekonstruktion eines physikalischen Praktikums für Physiker
ISBN 978-3-8325-0762-6 40.50 EUR

39 Michael Späth: Kontextbedingungen für Physikunterricht an der Hauptschule. *Möglichkeiten und Ansatzpunkte für einen fachübergreifenden, handlungsorientierten und berufsorientierten Unterricht*
ISBN 978-3-8325-0827-2 40.50 EUR

40 Jörg Hirsch: Interesse, Handlungen und situatives Erleben von Schülerinnen und Schülern beim Bearbeiten physikalischer Aufgaben
ISBN 978-3-8325-0875-3 40.50 EUR

41 Monika Hüther: Evaluation einer hypermedialen Lernumgebung zum Thema Gasgesetze. *Eine Studie im Rahmen des Physikpraktikums für Studierende der Medizin*
ISBN 978-3-8325-0911-8 40.50 EUR

42 Maike Tesch: Das Experiment im Physikunterricht. *Didaktische Konzepte und Ergebnisse einer Videostudie*
ISBN 978-3-8325-0975-0 40.50 EUR

43 Nina Nicolai: Skriptgeleitete Eltern-Kind-Interaktion bei Chemiehausaufgaben. *Eine Evaluationsstudie im Themenbereich Säure-Base*
ISBN 978-3-8325-1013-8 40.50 EUR

44 Antje Leisner: Entwicklung von Modellkompetenz im Physikunterricht
ISBN 978-3-8325-1020-6 40.50 EUR

45 Stefan Rumann: Evaluation einer Interventionsstudie zur Säure-Base-Thematik
ISBN 978-3-8325-1027-5 40.50 EUR

46 Thomas Wilhelm: Konzeption und Evaluation eines Kinematik/Dynamik-Lehrgangs zur Veränderung von Schülervorstellungen mit Hilfe dynamisch ikonischer Repräsentationen und graphischer Modellbildung – mit CD-ROM
ISBN 978-3-8325-1046-6 45.50 EUR

47 Andrea Maier-Richter: Computerunterstütztes Lernen mit Lösungsbeispielen in der Chemie. *Eine Evaluationsstudie im Themenbereich Löslichkeit*
ISBN 978-3-8325-1046-6 40.50 EUR

48 Jochen Peuckert: Stabilität und Ausprägung kognitiver Strukturen zum Atombegriff
ISBN 978-3-8325-1104-3 40.50 EUR

49 Maik Walpuski: Optimierung von experimenteller Kleingruppenarbeit durch Strukturierungshilfen und Feedback
ISBN 978-3-8325-1184-5 40.50 EUR

50 Helmut Fischler, Christiane S. Reiners (Hrsg.): Die Teilchenstruktur der Materie im Physik- und Chemieunterricht
ISBN 978-3-8325-1225-5 34.90 EUR

51 Claudia Eysel: Interdisziplinäres Lehren und Lernen in der Lehrerbildung. *Eine empirische Studie zum Kompetenzerwerb in einer komplexen Lernumgebung*
ISBN 978-3-8325-1238-5 40.50 EUR

52 Johannes Günther: Lehrerfortbildung über die Natur der Naturwissenschaften. *Studien über das Wissenschaftsverständnis von Grundschullehrkräften*
ISBN 978-3-8325-1287-3 40.50 EUR

53 Christoph Neugebauer: Lernen mit Simulationen und der Einfluss auf das Problemlösen in der Physik
ISBN 978-3-8325-1300-9 40.50 EUR

54 Andreas Schnirch: Gendergerechte Interessen- und Motivationsförderung im Kontext naturwissenschaftlicher Grundbildung. *Konzeption, Entwicklung und Evaluation einer multimedial unterstützten Lernumgebung*
ISBN 978-3-8325-1334-4 40.50 EUR

55 Hilde Köster: Freies Explorieren und Experimentieren. *Eine Untersuchung zur selbstbestimmten Gewinnung von Erfahrungen mit physikalischen Phänomenen im Sachunterricht*
ISBN 978-3-8325-1348-1 40.50 EUR

56 Eva Heran-Dörr: Entwicklung und Evaluation einer Lehrerfortbildung zur Förderung der physikdidaktischen Kompetenz von Sachunterrichtslehrkräften
ISBN 978-3-8325-1377-1 40.50 EUR

57 Agnes Szabone Varnai: Unterstützung des Problemlösens in Physik durch den Einsatz von Simulationen und die Vorgabe eines strukturierten Kooperationsformats
ISBN 978-3-8325-1403-7 40.50 EUR

58 Johannes Rethfeld: Aufgabenbasierte Lernprozesse in selbstorganisationsoffenem Unterricht der Sekundarstufe I zum Themengebiet ELEKTROSTATIK. *Eine Feldstudie in vier 10. Klassen zu einer kartenbasierten Lernumgebung mit Aufgaben aus der Elektrostatik*
ISBN 978-3-8325-1416-7 40.50 EUR

59 Christian Henke: Experimentell-naturwissenschaftliche Arbeitsweisen in der Oberstufe. *Untersuchung am Beispiel des HIGHSEA-Projekts in Bremerhaven*
ISBN 978-3-8325-1515-7 40.50 EUR

60 Lutz Kasper: Diskursiv-narrative Elemente für den Physikunterricht. *Entwicklung und Evaluation einer multimedialen Lernumgebung zum Erdmagnetismus*
ISBN 978-3-8325-1537-9 40.50 EUR

61 Thorid Rabe: Textgestaltung und Aufforderung zu Selbsterklärungen beim Physiklernen mit Multimedia
ISBN 978-3-8325-1539-3 40.50 EUR

62 Ina Glemnitz: Vertikale Vernetzung im Chemieunterricht. *Ein Vergleich von traditionellem Unterricht mit Unterricht nach Chemie im Kontext*
ISBN 978-3-8325-1628-4 40.50 EUR

63 Erik Einhaus: Schülerkompetenzen im Bereich Wärmelehre. *Entwicklung eines Testinstruments zur Überprüfung und Weiterentwicklung eines normativen Modells fachbezogener Kompetenzen*
ISBN 978-3-8325-1630-7 40.50 EUR

64 Jasmin Neuroth: Concept Mapping als Lernstrategie. *Eine Interventionsstudie zum Chemielernen aus Texten*
ISBN 978-3-8325-1659-8 40.50 EUR

65 Hans Gerd Hegeler-Burkhart: Zur Kommunikation von Hauptschülerinnen und Hauptschülern in einem handlungsorientierten und fächerübergreifenden Unterricht mit physikalischen und technischen Inhalten
ISBN 978-3-8325-1667-3 40.50 EUR

66 Karsten Rincke: Sprachentwicklung und Fachlernen im Mechanikunterricht. *Sprache und Kommunikation bei der Einführung in den Kraftbegriff*
ISBN 978-3-8325-1699-4 40.50 EUR

67 Nina Strehle: Das Ion im Chemieunterricht. *Alternative Schülervorstellungen und curriculare Konsequenzen*
ISBN 978-3-8325-1710-6 40.50 EUR

68 Martin Hopf: Problemorientierte Schülerexperimente
ISBN 978-3-8325-1711-3 40.50 EUR

69 Anne Beerenwinkel: Fostering conceptual change in chemistry classes using expository texts
ISBN 978-3-8325-1721-2 40.50 EUR

70 Roland Berger: Das Gruppenpuzzle im Physikunterricht der Sekundarstufe II. *Eine empirische Untersuchung auf der Grundlage der Selbstbestimmungstheorie der Motivation*
ISBN 978-3-8325-1732-8 40.50 EUR

71 Giuseppe Colicchia: Physikunterricht im Kontext von Medizin und Biologie. *Entwicklung und Erprobung von Unterrichtseinheiten*
ISBN 978-3-8325-1746-5 40.50 EUR

72 Sandra Winheller: Geschlechtsspezifische Auswirkungen der Lehrer-Schüler-Interaktion im Chemieanfangsunterricht
ISBN 978-3-8325-1757-1 40.50 EUR

73 Isabel Wahser: Training von naturwissenschaftlichen Arbeitsweisen zur Unterstützung experimenteller Kleingruppenarbeit im Fach Chemie
ISBN 978-3-8325-1815-8 40.50 EUR

74 Claus Brell: Lernmedien und Lernerfolg - reale und virtuelle Materialien im Physikunterricht. *Empirische Untersuchungen in achten Klassen an Gymnasien (Laborstudie) zum Computereinsatz mit Simulation und IBE*
ISBN 978-3-8325-1829-5 40.50 EUR

75 Rainer Wackermann: Überprüfung der Wirksamkeit eines Basismodell-Trainings für Physiklehrer
ISBN 978-3-8325-1882-0 40.50 EUR

76 Oliver Tepner: Effektivität von Aufgaben im Chemieunterricht der Sekundarstufe I
ISBN 978-3-8325-1919-3 40.50 EUR

77 Claudia Geyer: Museums- und Science-Center-Besuche im naturwissenschaftlichen Unterricht aus einer motivationalen Perspektive. *Die Sicht von Lehrkräften und Schülerinnen und Schülern*
ISBN 978-3-8325-1922-3 40.50 EUR

78 Tobias Leonhard: Professionalisierung in der Lehrerbildung. *Eine explorative Studie zur Entwicklung professioneller Kompetenzen in der Lehrererstausbildung*
ISBN 978-3-8325-1924-7 40.50 EUR

79 Alexander Kauertz: Schwierigkeitserzeugende Merkmale physikalischer Leistungstestaufgaben
ISBN 978-3-8325-1925-4 40.50 EUR

80 Regina Hübinger: Schüler auf Weltreise. *Entwicklung und Evaluation von Lehr-/ Lernmaterialien zur Förderung experimentell-naturwissenschaftlicher Kompetenzen für die Jahrgangsstufen 5 und 6*
ISBN 978-3-8325-1932-2 40.50 EUR

81 Christine Waltner: Physik lernen im Deutschen Museum
ISBN 978-3-8325-1933-9 40.50 EUR

82 Torsten Fischer: Handlungsmuster von Physiklehrkräften beim Einsatz neuer Medien. *Fallstudien zur Unterrichtspraxis*
ISBN 978-3-8325-1948-3 42.00 EUR

83 Corinna Kieren: Chemiehausaufgaben in der Sekundarstufe I des Gymnasiums. *Fragebogenerhebung zur gegenwärtigen Praxis und Entwicklung eines optimierten Hausaufgabendesigns im Themenbereich Säure-Base*
978-3-8325-1975-9 37.00 EUR

84 Marco Thiele: Modelle der Thermohalinen Zirkulation im Unterricht. *Eine empirische Studie zur Förderung des Modellverständnisses*
ISBN 978-3-8325-1982-7 40.50 EUR

85 Bernd Zinn: Physik lernen, um Physik zu lehren. *Eine Möglichkeit für interessanteren Physikunterricht*
ISBN 978-3-8325-1995-7 39.50 EUR

86 Esther Klaes: Außerschulische Lernorte im naturwissenschaftlichen Unterricht. *Die Perspektive der Lehrkraft*
ISBN 978-3-8325-2006-9 43.00 EUR

87 Marita Schmidt: Kompetenzmodellierung und -diagnostik im Themengebiet Energie der Sekundarstufe I. *Entwicklung und Erprobung eines Testinventars*
ISBN 978-3-8325-2024-3 37.00 EUR

88 Gudrun Franke-Braun: Aufgaben mit gestuften Lernhilfen. *Ein Aufgabenformat zur Förderung der sachbezogenen Kommunikation und Lernleistung für den naturwissenschaftlichen Unterricht*
ISBN 978-3-8325-2026-7 38.00 EUR

89 Silke Klos: Kompetenzförderung im naturwissenschaftlichen Anfangsunterricht. *Der Einfluss eines integrierten Unterrichtskonzepts*
ISBN 978-3-8325-2133-2 37.00 EUR

90 Ulrike Elisabeth Burkard: Quantenphysik in der Schule. *Bestandsaufnahme, Perspektiven und Weiterentwicklungsmöglichkeiten durch die Implementation eines Medienservers*
ISBN 978-3-8325-2215-5 43.00 EUR

91 Ulrike Gromadecki: Argumente in physikalischen Kontexten. *Welche Geltungsgründe halten Physikanfänger für überzeugend?*
ISBN 978-3-8325-2250-6 41.50 EUR

92 Jürgen Bruns: Auf dem Weg zur Förderung naturwissenschaftsspezifischer Vorstellungen von zukünftigen Chemie-Lehrenden
ISBN 978-3-8325-2257-5 43.50 EUR

93 Cornelius Marsch: Räumliche Atomvorstellung. *Entwicklung und Erprobung eines Unterrichtskonzeptes mit Hilfe des Computers*
ISBN 978-3-8325-2293-3 82.50 EUR

94 Maja Brückmann: Sachstrukturen im Physikunterricht. *Ergebnisse einer Videostudie*
ISBN 978-3-8325-2272-8 39.50 EUR

95 Sabine Fechner: Effects of Context-oriented Learning on Student Interest and Achievement in Chemistry Education
ISBN 978-3-8325-2343-5 36.50 EUR

96 Clemens Nagel: eLearning im Physikalischen Anfängerpraktikum
ISBN 978-3-8325-2355-8 39.50 EUR

97 Josef Riese: Professionelles Wissen und professionelle Handlungskompetenz von (angehenden) Physiklehrkräften
ISBN 978-3-8325-2376-3 39.00 EUR

98 Sascha Bernholt: Kompetenzmodellierung in der Chemie. *Theoretische und empirische Reflexion am Beispiel des Modells hierarchischer Komplexität*
ISBN 978-3-8325-2447-0 40.00 EUR

99 Holger Christoph Stawitz: Auswirkung unterschiedlicher Aufgabenprofile auf die Schülerleistung. *Vergleich von Naturwissenschafts- und Problemlöseaufgaben der PISA 2003-Studie*
ISBN 978-3-8325-2451-7 37.50 EUR

100 Hans Ernst Fischer, Elke Sumfleth (Hrsg.): nwu-essen – 10 Jahre Essener Forschung zum naturwissenschaftlichen Unterricht
ISBN 978-3-8325-3331-1 40.00 EUR

101 Hendrik Härtig: Sachstrukturen von Physikschulbüchern als Grundlage zur Bestimmung der Inhaltsvalidität eines Tests
ISBN 978-3-8325-2512-5 34.00 EUR

102 Thomas Grüß-Niehaus: Zum Verständnis des Löslichkeitskonzeptes im Chemieunterricht. *Der Effekt von Methoden progressiver und kollaborativer Reflexion*
ISBN 978-3-8325-2537-8 40.50 EUR

103 Patrick Bronner: Quantenoptische Experimente als Grundlage eines Curriculums zur Quantenphysik des Photons
ISBN 978-3-8325-2540-8 36.00 EUR

104 Adrian Voßkühler: Blickbewegungsmessung an Versuchsaufbauten. *Studien zur Wahrnehmung, Verarbeitung und Usability von physikbezogenen Experimenten am Bildschirm und in der Realität*
ISBN 978-3-8325-2548-4 47.50 EUR

105 Verena Tobias: Newton'sche Mechanik im Anfangsunterricht. *Die Wirksamkeit einer Einführung über die zweidimensionale Dynamik auf das Lehren und Lernen*
ISBN 978-3-8325-2558-3 54.00 EUR

106 Christian Rogge: Entwicklung physikalischer Konzepte in aufgabenbasierten Lernumgebungen
ISBN 978-3-8325-2574-3 45.00 EUR

107 Mathias Ropohl: Modellierung von Schülerkompetenzen im Basiskonzept Chemische Reaktion. *Entwicklung und Analyse von Testaufgaben*
ISBN 978-3-8325-2609-2 36.50 EUR

108 Christoph Kulgemeyer: Physikalische Kommunikationskompetenz. *Modellierung und Diagnostik*
ISBN 978-3-8325-2674-0 44.50 EUR

109 Jennifer Olszewski: The Impact of Physics Teachers' Pedagogical Content Knowledge on Teacher Actions and Student Outcomes
ISBN 978-3-8325-2680-1 33.50 EUR

110 Annika Ohle: Primary School Teachers' Content Knowledge in Physics and its Impact on Teaching and Students' Achievement
ISBN 978-3-8325-2684-9 36.50 EUR

111 Susanne Mannel: Assessing scientific inquiry. *Development and evaluation of a test for the low-performing stage*
ISBN 978-3-8325-2761-7 40.00 EUR

112 Michael Plomer: Physik physiologisch passend praktiziert. *Eine Studie zur Lernwirksamkeit von traditionellen und adressatenspezifischen Physikpraktika für die Physiologie*
ISBN 978-3-8325-2804-1 34.50 EUR

113 Alexandra Schulz: Experimentierspezifische Qualitätsmerkmale im Chemieunterricht. *Eine Videostudie*
ISBN 978-3-8325-2817-1 40.00 EUR

114 Franz Boczianowski: Eine empirische Untersuchung zu Vektoren im Physikunterricht der Mittelstufe
ISBN 978-3-8325-2843-0 39.50 EUR

115 Maria Ploog: Internetbasiertes Lernen durch Textproduktion im Fach Physik
ISBN 978-3-8325-2853-9 39.50 EUR

116 Anja Dhein: Lernen in Explorier- und Experimentiersituationen. *Eine explorative Studie zu Bedeutungsentwicklungsprozessen bei Kindern im Alter zwischen 4 und 6 Jahren*
ISBN 978-3-8325-2859-1 45.50 EUR

117 Irene Neumann: Beyond Physics Content Knowledge. *Modeling Competence Regarding Nature of Scientific Inquiry and Nature of Scientific Knowledge*
ISBN 978-3-8325-2880-5 37.00 EUR

118 Markus Emden: Prozessorientierte Leistungsmessung des naturwissenschaftlich-experimentellen Arbeitens. *Eine vergleichende Studie zu Diagnoseinstrumenten zu Beginn der Sekundarstufe I*
ISBN 978-3-8325-2867-6 38.00 EUR

119 Birgit Hofmann: Analyse von Blickbewegungen von Schülern beim Lesen von physikbezogenen Texten mit Bildern. *Eye Tracking als Methodenwerkzeug in der physikdidaktischen Forschung*
ISBN 978-3-8325-2925-3 59.00 EUR

120 Rebecca Knobloch: Analyse der fachinhaltlichen Qualität von Schüleräußerungen und deren Einfluss auf den Lernerfolg. *Eine Videostudie zu kooperativer Kleingruppenarbeit*
ISBN 978-3-8325-3006-8 36.50 EUR

121 Julia Hostenbach: Entwicklung und Prüfung eines Modells zur Beschreibung der Bewertungskompetenz im Chemieunterricht
ISBN 978-3-8325-3013-6 38.00 EUR

122 Anna Windt: Naturwissenschaftliches Experimentieren im Elementarbereich. *Evaluation verschiedener Lernsituationen*
ISBN 978-3-8325-3020-4 43.50 EUR

123 Eva Kölbach: Kontexteinflüsse beim Lernen mit Lösungsbeispielen
ISBN 978-3-8325-3025-9 38.50 EUR

124 Anna Lau: Passung und vertikale Vernetzung im Chemie- und Physikunterricht
ISBN 978-3-8325-3021-1 36.00 EUR

125 Jan Lamprecht: Ausbildungswege und Komponenten professioneller Handlungskompetenz. *Vergleich von Quereinsteigern mit Lehramtsabsolventen für Gymnasien im Fach Physik*
ISBN 978-3-8325-3035-8 38.50 EUR

126 Ulrike Böhm: Förderung von Verstehensprozessen unter Einsatz von Modellen
ISBN 978-3-8325-3042-6 41.00 EUR

127 Sabrina Dollny: Entwicklung und Evaluation eines Testinstruments zur Erfassung des fachspezifischen Professionswissens von Chemielehrkräften
ISBN 978-3-8325-3046-4 37.00 EUR

128 Monika Zimmermann: Naturwissenschaftliche Bildung im Kindergarten. *Eine integrative Längsschnittstudie zur Kompetenzentwicklung von Erzieherinnen*
ISBN 978-3-8325-3053-2 54.00 EUR

129 Ulf Saballus: Über das Schlussfolgern von Schülerinnen und Schülern zu öffentlichen Kontroversen mit naturwissenschaftlichem Hintergrund. *Eine Fallstudie*
ISBN 978-3-8325-3086-0 39.50 EUR

130 Olaf Krey: Zur Rolle der Mathematik in der Physik. *Wissenschaftstheoretische Aspekte und Vorstellungen Physiklernender*
ISBN 978-3-8325-3101-0 46.00 EUR

131 Angelika Wolf: Zusammenhänge zwischen der Eigenständigkeit im Physikunterricht, der Motivation, den Grundbedürfnissen und dem Lernerfolg von Schülern
ISBN 978-3-8325-3161-4 45.00 EUR

132 Johannes Börlin: Das Experiment als Lerngelegenheit. *Vom interkulturellen Vergleich des Physikunterrichts zu Merkmalen seiner Qualität*
ISBN 978-3-8325-3170-6 45.00 EUR

133 Olaf Uhden: Mathematisches Denken im Physikunterricht. *Theorieentwicklung und Problemanalyse*
ISBN 978-3-8325-3170-6 45.00 EUR

134 Christoph Gut: Modellierung und Messung experimenteller Kompetenz. *Analyse eines large-scale Experimentiertests*
ISBN 978-3-8325-3213-0 40.00 EUR

135 Antonio Rueda: Lernen mit ExploMultimedial in kolumbianischen Schulen. *Analyse von kurzzeitigen Lernprozessen und der Motivation beim länderübergreifenden Einsatz einer deutschen computergestützten multimedialen Lernumgebung für den naturwissenschaftlichen Unterricht*
ISBN 978-3-8325-3218-5 45.50 EUR

136 Krisztina Berger: Bilder, Animationen und Notizen. *Empirische Untersuchung zur Wirkung einfacher visueller Repräsentationen und Notizen auf den Wissenserwerb in der Optik*
ISBN 978-3-8325-3238-3 41.50 EUR

137 Antony Crossley: Untersuchung des Einflusses unterschiedlicher physikalischer Konzepte auf den Wissenserwerb in der Thermodynamik der Sekundarstufe I
ISBN 978-3-8325-3275-8 40.00 EUR

138 Tobias Viering: Entwicklung physikalischer Kompetenz in der Sekundarstufe I. *Validierung eines Kompetenzentwicklungsmodells für das Energiekonzept im Bereich Fachwissen*
ISBN 978-3-8325-3277-2 37.00 EUR

139 Nico Schreiber: Diagnostik experimenteller Kompetenz. *Validierung technologiegestützter Testverfahren im Rahmen eines Kompetenzstrukturmodells*
ISBN 978-3-8325-3284-0 39.00 EUR

140 Sarah Hundertmark: Einblicke in kollaborative Lernprozesse. *Eine Fallstudie zur reflektierenden Zusammenarbeit unterstützt durch die Methoden Concept Mapping und Lernbegleitbogen*
ISBN 978-3-8325-3251-2 43.00 EUR

141 Ronny Scherer: Analyse der Struktur, Messinvarianz und Ausprägung komplexer Problemlösekompetenz im Fach Chemie. *Eine Querschnittstudie in der Sekundarstufe I und am Übergang zur Sekundarstufe II*
ISBN 978-3-8325-3312-0 43.00 EUR

142 Patricia Heitmann: Bewertungskompetenz im Rahmen naturwissenschaftlicher Problemlöseprozesse. *Modellierung und Diagnose der Kompetenzen Bewertung und analytisches Problemlösen für das Fach Chemie*
ISBN 978-3-8325-3314-4 37.00 EUR

143 Jan Fleischhauer: Wissenschaftliches Argumentieren und Entwicklung von Konzepten beim Lernen von Physik
ISBN 978-3-8325-3325-0 35.00 EUR

144 Nermin Özcan: Zum Einfluss der Fachsprache auf die Leistung im Fach Chemie. *Eine Förderstudie zur Fachsprache im Chemieunterricht*
ISBN 978-3-8325-3328-1 36.50 EUR

145 Helena van Vorst: Kontextmerkmale und ihr Einfluss auf das Schülerinteresse im Fach Chemie
ISBN 978-3-8325-3321-2 38.50 EUR

146 Janine Cappell: Fachspezifische Diagnosekompetenz angehender Physiklehrkräfte in der ersten Ausbildungsphase
ISBN 978-3-8325-3356-4 38.50 EUR

147 Susanne Bley: Förderung von Transferprozessen im Chemieunterricht
ISBN 978-3-8325-3407-3 40.50 EUR

148 Cathrin Blaes: Die übungsgestützte Lehrerpräsentation im Chemieunterricht der Sekundarstufe I. *Evaluation der Effektivität*
ISBN 978-3-8325-3409-7 43.50 EUR

149 Julia Suckut: Die Wirksamkeit von piko-OWL als Lehrerfortbildung. Eine Evaluation zum Projekt *Physik im Kontext* in Fallstudien
ISBN 978-3-8325-3440-0 45.00 EUR

150 Alexandra Dorschu: Die Wirkung von Kontexten in Physikkompetenztestaufgaben
ISBN 978-3-8325-3446-2 37.00 EUR

151 Jochen Scheid: Multiple Repräsentationen, Verständnis physikalischer Experimente und kognitive Aktivierung: *Ein Beitrag zur Entwicklung der Aufgabenkultur*
ISBN 978-3-8325-3449-3 49.00 EUR

152 Tim Plasa: Die Wahrnehmung von Schülerlaboren und Schülerforschungszentren
ISBN 978-3-8325-3483-7 35.50 EUR

153 Felix Schoppmeier: Physikkompetenz in der gymnasialen Oberstufe.*Entwicklung und Validierung eines Kompetenzstrukturmodells für den Kompetenzbereich Umgang mit Fachwissen*
ISBN 978-3-8325-3502-5 36.00 EUR

154 Katharina Groß: Experimente alternativ dokumentieren. *Eine qualitative Studie zur Förderung der Diagnose- und Differenzierungskompetenz in der Chemielehrerbildung*
ISBN 978-3-8325-3508-7 43.50 EUR

155 Barbara Hank: Konzeptwandelprozesse im Anfangsunterricht Chemie. *Eine quasiexperimentelle Längsschnittstudie*
ISBN 978-3-8325-3519-3 38.50 EUR

156 Katja Freyer: Zum Einfluss von Studieneingangsvoraussetzungen auf den Studienerfolg Erstsemesterstudierender im Fach Chemie
ISBN 978-3-8325-3544-5 38.00 EUR

157 Alexander Rachel: Auswirkungen instruktionaler Hilfen bei der Einführung des (Ferro-)Magnetismus. *Eine Vergleichsstudie in der Primar- und Sekundarstufe*
ISBN 978-3-8325-3548-3 43.50 EUR

158 Sebastian Ritter: Einfluss des Lerninhalts Nanogrößeneffekte auf Teilchen- und Teilchenmodellvorstellungen von Schülerinnen und Schülern
ISBN 978-3-8325-3558-2 36.00 EUR

159 Andrea Harbach: Problemorientierung und Vernetzung in kontextbasierten Lernaufgaben
ISBN 978-3-8325-3564-3 39.00 EUR

160 David Obst: Interaktive Tafeln im Physikunterricht. *Entwicklung und Evaluation einer Lehrerfortbildung*
ISBN 978-3-8325-3582-7 40.50 EUR

161 Sophie Kirschner: Modellierung und Analyse des Professionswissens von Physiklehrkräften
ISBN 978-3-8325-3601-5 35.00 EUR

162 Katja Stief: Selbstregulationsprozesse und Hausaufgabenmotivation im Chemieunterricht
ISBN 978-3-8325-3631-2 34.00 EUR

163 Nicola Meschede: Professionelle Wahrnehmung der inhaltlichen Strukturierung im naturwissenschaftlichen Grundschulunterricht. *Theoretische Beschreibung und empirische Erfassung*
ISBN 978-3-8325-3668-8 37.00 EUR

164 Johannes Maximilian Barth: Experimentieren im Physikunterricht der gymnasialen Oberstufe. *Eine Rekonstruktion übergeordneter Einbettungsstrategien*
ISBN 978-3-8325-3681-7 39.00 EUR

165 Sandra Lein: Das Betriebspraktikum in der Lehrerbildung. *Eine Untersuchung zur Förderung der Wissenschafts- und Technikbildung im allgemeinbildenden Unterricht*
ISBN 978-3-8325-3698-5 40.00 EUR

166 Veranika Maiseyenka: Modellbasiertes Experimentieren im Unterricht. *Praxistauglichkeit und Lernwirkungen*
ISBN 978-3-8325-3708-1 38.00 EUR

167 Christoph Stolzenberger: Der Einfluss der didaktischen Lernumgebung auf das Erreichen geforderter Bildungsziele am Beispiel der W- und P-Seminare im Fach Physik
ISBN 978-3-8325-3708-1 38.00 EUR

168 Pia Altenburger: Mehrebenenregressionsanalysen zum Physiklernen im Sachunterricht der Primarstufe. *Ergebnisse einer Evaluationsstudie.*
ISBN 978-3-8325-3717-3 37.50 EUR

169 Nora Ferber: Entwicklung und Validierung eines Testinstruments zur Erfassung von Kompetenzentwicklung im Fach Chemie in der Sekundarstufe I
ISBN 978-3-8325-3727-2 39.50 EUR

170 Anita Stender: Unterrichtsplanung: Vom Wissen zum Handeln.
Theoretische Entwicklung und empirische Überprüfung des Transformationsmodells der Unterrichtsplanung
ISBN 978-3-8325-3750-0 41.50 EUR

171 Jenna Koenen: Entwicklung und Evaluation von experimentunterstützten Lösungsbeispielen zur Förderung naturwissenschaftlich-experimenteller Arbeitsweisen
ISBN 978-3-8325-3785-2 43.00 EUR

172 Teresa Henning: Empirische Untersuchung kontextorientierter Lernumgebungen in der Hochschuldidaktik. *Entwicklung und Evaluation kontextorientierter Aufgaben in der Studieneingangsphase für Fach- und Nebenfachstudierende der Physik*
ISBN 978-3-8325-3801-9 43.00 EUR

173 Alexander Pusch: Fachspezifische Instrumente zur Diagnose und individuellen Förderung von Lehramtsstudierenden der Physik
ISBN 978-3-8325-3829-3 38.00 EUR

174 Christoph Vogelsang: Validierung eines Instruments zur Erfassung der professionellen Handlungskompetenz von (angehenden) Physiklehrkräften. *Zusammenhangsanalysen zwischen Lehrerkompetenz und Lehrerperformanz*
ISBN 978-3-8325-3846-0 50.50 EUR

175 Ingo Brebeck: Selbstreguliertes Lernen in der Studieneingangsphase im Fach Chemie
ISBN 978-3-8325-3859-0 37.00 EUR

176 Axel Eghtessad: Merkmale und Strukturen von Professionalisierungsprozessen in der ersten und zweiten Phase der Chemielehrerbildung. *Eine empirisch-qualitative Studie mit niedersächsischen Fachleiter_innen der Sekundarstufenlehrämter*
ISBN 978-3-8325-3861-3 45.00 EUR

177 Andreas Nehring: Wissenschaftliche Denk- und Arbeitsweisen im Fach Chemie. Eine kompetenzorientierte Modell- und Testentwicklung für den Bereich der Erkenntnisgewinnung
ISBN 978-3-8325-3872-9 39.50 EUR

178 Maike Schmidt: Professionswissen von Sachunterrichtslehrkräften. Zusammenhangsanalyse zur Wirkung von Ausbildungshintergrund und Unterrichtserfahrung auf das fachspezifische Professionswissen im Unterrichtsinhalt „Verbrennung“
ISBN 978-3-8325-3907-8 38.50 EUR

179 Jan Winkelmann: Auswirkungen auf den Fachwissenszuwachs und auf affektive Schülermerkmale durch Schüler- und Demonstrationsexperimente im Physikunterricht
ISBN 978-3-8325-3915-3 41.00 EUR

180 Iwen Kobow: Entwicklung und Validierung eines Testinstrumentes zur Erfassung der Kommunikationskompetenz im Fach Chemie
ISBN 978-3-8325-3927-6 34.50 EUR

181 Yvonne Gramzow: Fachdidaktisches Wissen von Lehramtsstudierenden im Fach Physik. Modellierung und Testkonstruktion
ISBN 978-3-8325-3931-3 42.50 EUR

182 Evelin Schröter: Entwicklung der Kompetenzerwartung durch Lösen physikalischer Aufgaben einer multimedialen Lernumgebung
ISBN 978-3-8325-3975-7 54.50 EUR

183 Inga Kallweit: Effektivität des Einsatzes von Selbsteinschätzungsbögen im Chemieunterricht der Sekundarstufe I. *Individuelle Förderung durch selbstreguliertes Lernen*
ISBN 978-3-8325-3965-8 44.00 EUR

184 Andrea Schumacher: Paving the way towards authentic chemistry teaching. *A contribution to teachers' professional development*
ISBN 978-3-8325-3976-4 48.50 EUR

185 David Woitkowski: Fachliches Wissen Physik in der Hochschulausbildung. *Konzeptualisierung, Messung, Niveaubildung*
ISBN 978-3-8325-3988-7 53.00 EUR

186 Marianne Korner: Cross-Age Peer Tutoring in Physik. *Evaluation einer Unterrichtsmethode*
ISBN 978-3-8325-3979-5 38.50 EUR

187 Simone Nakoinz: Untersuchung zur Verknüpfung submikroskopischer und makroskopischer Konzepte im Fach Chemie
ISBN 978-3-8325-4057-9 38.50 EUR

188 Sandra Anus: Evaluation individueller Förderung im Chemieunterricht.*Adaptivität von Lerninhalten an das Vorwissen von Lernenden am Beispiel des Basiskonzeptes Chemische Reaktion*
ISBN 978-3-8325-4059-3 43.50 EUR

189 Thomas Roßbegalle: Fachdidaktische Entwicklungsforschung zum besseren Verständnis atmosphärischer Phänomene. *Treibhauseffekt, saurer Regen und stratosphärischer Ozonabbau als Kontexte zur Vermittlung von Basiskonzepten der Chemie*
ISBN 978-3-8325-4059-3 45.50 EUR

190 Kathrin Steckenmesser-Sander: Gemeinsamkeiten und Unterschiede physikbezogener Handlungs-, Denk- und Lernprozesse von Mädchen und Jungen
ISBN 978-3-8325-4066-1 38.50 EUR

191 Cornelia Geller: Lernprozessorientierte Sequenzierung des Physikunterrichts im Zusammenhang mit Fachwissenserwerb. *Eine Videostudie in Finnland, Deutschland und der Schweiz*
ISBN 978-3-8325-4082-1 35.50 EUR

192 Jan Hofmann: Untersuchung des Kompetenzaufbaus von Physiklehrkräften während einer Fortbildungsmaßnahme
ISBN 978-3-8325-4104-0 38.50 EUR

193 Andreas Dickhäuser: Chemiespezifischer Humor. *Theoriebildung, Materialentwicklung, Evaluation*
ISBN 978-3-8325-4108-8 37.00 EUR

194 Stefan Korte: Die Grenzen der Naturwissenschaft als Thema des Physikunterrichts
ISBN 978-3-8325-4112-5 57.50 EUR

195 Carolin Hülsmann: Kurswahlmotive im Fach Chemie. Eine Studie zum Wahlverhalten und Erfolg von Schülerinnen und Schülern in der gymnasialen Oberstufe
ISBN 978-3-8325-4144-6 49.00 EUR

196 Caroline Körbs: Mindeststandards im Fach Chemie am Ende der Pflichtschulzeit
ISBN 978-3-8325-4148-4 34.00 EUR

197 Andreas Vorholzer: Wie lassen sich Kompetenzen des experimentellen Denkens und Arbeitens fördern? *Eine empirische Untersuchung der Wirkung eines expliziten und eines impliziten Instruktionsansatzes*
ISBN 978-3-8325-4194-1 37.50 EUR

198 Anna Katharina Schmitt: Entwicklung und Evaluation einer Chemielehrerfortbildung zum Kompetenzbereich Erkenntnisgewinnung
ISBN 978-3-8325-4228-3 39.50 EUR

199 Christian Maurer: Strukturierung von Lehr-Lern-Sequenzen
ISBN 978-3-8325-4247-4 36.50 EUR

201 Simon Zander: Lehrerfortbildung zu Basismodellen und Zusammenhänge zum Fachwissen
ISBN 978-3-8325-4248-1 35.00 EUR

202 Kerstin Arndt: Experimentierkompetenz erfassen. *Analyse von Prozessen und Mustern am Beispiel von Lehramtsstudierenden der Chemie*
ISBN 978-3-8325-4266-5 45.00 EUR

203 Christian Lang: Kompetenzorientierung im Rahmen experimentalchemischer Praktika
ISBN 978-3-8325-4268-9 42.50 EUR

204 Eva Cauet: Testen wir relevantes Wissen? *Zusammenhang zwischen dem Professionswissen von Physiklehrkräften und gutem und erfolgreichem Unterrichten*
ISBN 978-3-8325-4276-4 39.50 EUR

205 Patrick Löffler: Modellanwendung in Problemlöseaufgaben. *Wie wirkt Kontext?*
ISBN 978-3-8325-4303-7 35.00 EUR

206 Carina Gehlen: Kompetenzstruktur naturwissenschaftlicher Erkenntnisgewinnung im Fach Chemie
ISBN 978-3-8325-4318-1 43.00 EUR

208 Jennifer Petersen: Zum Einfluss des Merkmals Humor auf die Gesundheitsförderung im Chemieunterricht der Sekundarstufe I. *Eine Interventionsstudie zum Thema Sonnenschutz*
ISBN 978-3-8325-4348-8 40.00 EUR

209 Philipp Straube: Modellierung und Erfassung von Kompetenzen naturwissenschaftlicher Erkenntnisgewinnung bei (Lehramts-) Studierenden im Fach Physik
ISBN 978-3-8325-4351-8 35.50 EUR

210 Martin Dickmann: Messung von Experimentierfähigkeiten. *Validierungsstudien zur Qualität eines computerbasierten Testverfahrens*
ISBN 978-3-8325-4356-3 41.00 EUR

In Vorbereitung

207 Lars Oettinghaus Lehrerüberzeugungen und physikbezogenes Professionswissen
ISBN 978-3-8325-4319-8 38.50 EUR

Studien zum Physik- und Chemielernen

Herausgegeben von Hans Niedderer, Helmut Fischler und Elke Sumfleth

Die Reihe umfasst inzwischen eine große Zahl von wissenschaftlichen Arbeiten aus vielen Arbeitsgruppen der Physik- und Chemiedidaktik und zeichnet damit ein gültiges Bild der empirischen physik- und chemiedidaktischen Forschung in Deutschland.
Die Herausgeber laden daher Interessenten zu neuen Beiträgen ein und bitten sie, sich im Bedarfsfall an den Logos-Verlag oder an ein Mitglied des Herausgeberteams zu wenden.

Kontaktadressen:

Prof. Dr. Hans Niedderer
Institut für Didaktik der Naturwissenschaften,
Abt. Physikdidaktik, FB Physik/Elektrotechnik,
Universität Bremen,
Postfach 33 04 40, 28334 Bremen
Tel. 0421-218 2484/4695, e-mail:
niedderer@physik.uni-bremen.de

Prof. Dr. Helmut Fischler
Didaktik der Physik, FB Physik, Freie Universität Berlin,
Arnimallee 14, 14195 Berlin
Tel. 030-838 56712/55966, e-mail:
fischler@physik.fu-berlin.de

Prof. Dr. Elke Sumfleth
Didaktik der Chemie,
Fachbereich Chemie,
Universität Duisburg-Essen,
Schützenbahn 70, 45127 Essen
Tel. 0201-183 3757/3761, e-mail:
elke.sumfleth@uni-essen.de

Die vorliegende Arbeit beschäftigt sich mit der Validierung eines neuen Tests zur Messung von Experimentierfähigkeit in Large-Scale-Assessments. Die Testaufgaben erfordern die Planung, Durchführung und Auswertung typischer physikalischer Schülerexperimente der Sekundarstufe I. Anstelle von Realexperimenten werden interaktive Simulationen zum Aufbau von Experimenten und zur Durchführung von Messungen eingesetzt. Die Bearbeitung der Testaufgaben erfolgt vollständig am Computerbildschirm. Diesen Bearbeitungen werden in der Auswertung Punkte (Testwerte) zugewiesen.

In der heute gültigen Vorstellung ist Validierung im Wesentlichen als ein theorie- und evidenzbasierter Argumentationsprozess aufzufassen. Nach der Nennung und Begründung notwendiger Validierungsschritte erfolgt eine schlüssige Validitätsargumentation. Ein Validierungsschritt kann z. B. darin bestehen, zu überprüfen, in wieweit zur erfolgreichen Bearbeitung der Testaufgaben Experimentierfähigkeiten eingesetzt werden.

Die Ergebnisse der Validierungsschritte lassen sich wie folgt zusammenfassen: die Aufgabeninhalte und -anforderungen des Tests passen zum Physikunterricht der Sekundarstufe I; die Schüler denken und handeln bei der Bearbeitung der Testaufgaben experimentell; der verwendete Bewertungsmaßstab misst Experimentierfähigkeit; die erreichten Testwerte korrelieren mit den Testwerten, die bei inhaltlich identischen Realexperimenten erreicht werden. Zusammenfassend ist festzuhalten, dass der computerbasierte Test das Spektrum der in Large-Scale Assessments erfassbaren Experimentierfähigkeiten erweitert.

Logos Verlag Berlin

ISBN 978-3-8325-4356-3