Physik und Didaktik in Schule und Hochschule
PhyDid 1/15 (2016), S. 26-48

Messung experimenteller Kompetenz – ein computergestützter Experimentiertest –

Heike Theyßen*, Horst Schecker+, Knut Neumann++, Bodo Eickhorst+ & Martin Dickmann*

* Universität Duisburg-Essen, Didaktik der Physik,
+ Universität Bremen, Institut für Didaktik der Naturwissenschaften, Abt. Physikdidaktik,
++ Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik (IPN), Kiel,
heike.theyssen@uni-due.de, schecker@uni-bremen.de, neumann@ipn.uni-kiel.de,
bodoe@uni-bremen.de, martin.dickmann@uni-due.de
(Eingegangen: 21.10.2015; Angenommen: 25.10.2016)

Kurzfassung

Der Erwerb experimenteller Kompetenz gehört zu den zentralen Bildungszielen des Physikunterrichts. In wieweit dieses Ziel erreicht wird, wird jedoch bislang im Bildungsmonitoring nur unvollständig überprüft, da insbesondere für den Bereich der Durchführung von Experimenten geeignete Testinstrumente für den Large-Scale-Einsatz fehlen. In diesem Beitrag wird ein Experimentiertest vorgestellt, der diese Lücke schließen soll. Der Test wird vollständig am Computerbildschirm bearbeitet. Die Aufgaben aus den Themenbereichen Mechanik, Optik und Elektrizitätslehre beziehen sich auf typische Schülerexperimente der Sekundarstufe I. Die Realexperimente sind durch interaktive Simulationen zum Aufbau der Experimente und zur Durchführung von Messungen ersetzt. Die Bewertung der Schülerlösungen erfolgt computergestützt und teils automatisiert. Der Test hat sich in einer Large-Scale-Erprobung mit über 1.100 Schülerinnen und Schülern als praktikabel einsetzbar und auswertbar erwiesen. Über alle Items hinweg erweist sich die interne Konsistenz mindestens als gut und die empirisch gefundenen Itemschwierigkeiten lassen sich inhaltlich sinnvoll interpretieren.

Abstract

Teaching students the necessary skills to plan, perform and analyze experiments is a major aim of physics instruction. In large-scale-assessments, however, these skills are hardly assessed. A main reason is a lack of assessment instruments that can efficiently and reliably assess these skills, including the performance of experiments. In order to close this gap, in this paper, we present a completely computer-based test that assesses students' experimental skills in mechanics, electricity and optics, taking up typical student experiments in these areas. The test builds on interactive simulations that resemble the set-up and performance of real experiments. The scoring is computer-based and partially automated. Results of a large-scale-study with more than 1.100 students suggest that the test can efficiently be used in the large scale. Across all items, internal consistency was found to be at least good. The item difficulties can be interpreted plausibly.

1. Einleitung

Der Erwerb experimenteller Kompetenz ist national wie international ein zentrales Ziel des naturwissenschaftlichen Unterrichts (z.B. [1] und [2]). Gemäß den Bildungsstandards für den mittleren Schulabschluss sollen Schülerinnen und Schüler u. a. dazu in der Lage sein, einfache Experimente zu planen, durchzuführen und zu dokumentieren, gewonnene Daten auszuwerten und die Gültigkeit empirischer Ergebnisse zu beurteilen ([1], S. 11). Ob diese Bildungsziele erreicht werden, muss im Rahmen des Bildungsmonitorings überprüft werden, auch um der Wertigkeit dieser Bildungsziele gerecht zu werden.

Die zum Bildungsmonitoring bislang eingesetzten Large-Scale-Untersuchungen erfassen experimentelle Kompetenz aus Effizienzgründen meist mit rein schriftlichen Verfahren. Schülerinnen und Schüler experimentieren nicht selbst, sondern versetzen sich in eine experimentelle Situation und beantworten Fragen dazu (z. B. NAW-Test, vgl. z. B. [3]; IQB-Ländervergleich, vgl. z. B. [4], und im Hochschulsektor [5]). Der Fokus liegt dabei in der Regel auf der Planung von Experimenten und der Analyse von Messdaten. Mit rein schriftlichen Verfahren lassen sich Fähigkeiten zur Durchführung von Experimenten kaum erfassen. Erst Experimentiertests bringen

Schülerinnen und Schüler in Handlungssituationen mit realem Experimentiermaterial. In die gestellten Anforderungen kann daher neben der Planung und Auswertung auch die konkrete Durchführung von Experimenten einbezogen werden (Versuche aufbauen, Messen). Experimentiertests mit Realexperimenten sind jedoch sehr aufwändig in Administration und Auswertung [6].

In dem vorliegenden Beitrag wird ein Test vorgestellt, bei dem Realexperimente durch interaktive Simulationen mit möglichst ähnlichen Handlungsoptionen ersetzt wurden.

2. Stand der Forschung

2.1. Experimentelle Kompetenz

Das Experiment spielt in den Naturwissenschaften als Methode der Erkenntnisgewinnung eine zentrale Rolle (z. B. [7] und [8]). Fähigkeiten zur experimentellen Beantwortung bzw. Überprüfung naturwissenschaftlicher Fragestellungen und Hypothesen sind daher ein wesentlicher Bestandteil einer anschlussfähigen naturwissenschaftlichen Bildung. Sie werden in den Bildungsstandards sowie entsprechenden internationalen Dokumenten explizit berücksichtigt (siehe 1.).

Modelle experimenteller Kompetenz umfassen in der Regel die Fähigkeiten zur Planung und Durchführung von Experimenten sowie zur Auswertung der damit gewonnenen Daten, unterscheiden sich jedoch im Umfang und in der Gewichtung einzelner Fähigkeiten (z.B. [9] - [12]; für eine Übersicht siehe [13]). Ein Modell, das Lehrkräften für die Unterrichtsplanung Orientierungshilfen bietet, findet sich bei Nawrath, Maiseyenka und Schecker [14]. Dort werden folgende Fähigkeiten unterschieden: Fragestellung entwickeln, Hypothesen bilden, Experiment planen, Versuchsanordnung funktionsfähig aufbauen, Beobachten/ Messen, Daten aufbereiten und sachgerechte Schlüsse ziehen.

2.2. Messung experimenteller Kompetenz

Der in Bildungsplänen dokumentierte Stellenwert der Vermittlung experimenteller Kompetenz bildet sich im Bildungsmonitoring bisher noch nicht ab, nicht zuletzt wegen eines Mangels an Testverfahren zur reliablen und validen Erfassung aller Bereiche experimenteller Kompetenz auch in Large-Scale-Untersuchungen.

Die Erfassung experimenteller Kompetenz erfolgt bislang im Wesentlichen über zwei Verfahren, die sich entweder auf schriftliche Aufgaben oder auf Aufgaben mit Realexperimenten stützen. Verfahren mit schriftlichen Aufgaben fokussieren dabei häufig auf die Planung und Auswertung, wie z.B. bei Mayer, Grube und Möller [10]. Auch der ebenfalls auf schriftlichen Aufgaben basierende Naturwissenschaftliche-Arbeitsweisen (NAW)-Test (z.B. [3], [15] - [17]) erfasst neben der Fähigkeit zur Formulierung von Hypothesen und der Fähigkeit Schlussfolgerungen zu ziehen, lediglich, ob Schülerinnen und Schüler in der Lage sind, Vorschläge zur experimentellen Prüfung von Hypothesen auf ihre Richtigkeit zu prüfen [17].

Während Tests mit rein schriftlichen Aufgaben mittels einer großen Anzahl zu bearbeitender Aufgaben durchaus hohe Reliabilität erzielen können, führt die Vernachlässigung der konkreten Durchführung von Experimenten zu einer unvollständigen Repräsentation des Konstruktes experimenteller Kompetenz. Hinzu kommt, dass empirische Untersuchungen nur moderate Zusammenhänge zwischen Tests mit schriftlichen Aufgaben und Tests auf der Grundlage experimenteller Aufgaben nachweisen können (z.B. [18] – [20]). Beide Aspekte stellen die Validität der Kompetenzmessung mit rein schriftlichen Aufgaben in Frage. Bezogen auf den Validitätsbegriff nach Messick [21] betrifft das den inhaltlichen und den externen (hier: konvergenten) Aspekt von Validität.

Die Erfassung experimenteller Kompetenz mit Hilfe von Aufgaben, bei denen ein Experiment tatsächlich auch durchgeführt werden muss, geht zurück auf die – vor allem im anglo-amerikanischen Raum angesiedelte – Forschung zu Performance Tests (z.B. [19], [22] – [25]). Tests mit Realexperimenten werden – sicherlich nicht zuletzt aufgrund des hohen organisatorischen, zeitlichen und finanziellen Aufwands – nur vereinzelt in Large-Scale-Untersuchungen eingesetzt, z. B. im Rahmen der Third International Mathematics and Science (TIMS) Studie [26] und im Rahmen des Projekts zur Harmonisierung der obligatorischen Schule (HarmoS) in der Schweiz [27]. Es ist schwer, mit solchen Messungen zufriedenstellende Reliabilitäten zu erreichen, da ein Proband jeweils nur eine geringe Zahl von Aufgaben bearbeiten kann (z.B. [28] und [29]).

Als weiteres Testverfahren für experimentelle Kompetenz wurden vereinzelt Tests mit interaktiven Simulationen entwickelt (z.B. [20], [30] und [31]). Sie sind Tests mit Realexperimenten hinsichtlich der Praktikabilität überlegen und ermöglichen, mehr Aufgaben innerhalb einer gegebenen Testzeit zu bearbeiten. Im Vergleich zu Tests mit rein schriftlichen Aufgaben kann durch den Einsatz interaktiver Simulationen die Bewältigung konkreter (wenn auch virtueller) Handlungssituationen im Bereich der Durchführung von Experimenten mit erfasst werden. Dabei erhalten die Probandinnen und Probanden eine dem Realexperiment vergleichbare Reaktion des Experimentiermaterials auf Manipulationen, z. B. können sie unmittelbar erkennen, ob eine Glühlampe in dem von ihnen aufgebauten Stromkreis leuchtet, und können darauf reagieren. Vor diesem Hintergrund kommen Shavelson und Kollegen [30] zu dem Schluss, dass in Performance Assessments Aufgaben mit interaktiven Simulationen potenziell einen geeigneten Ersatz für Aufgaben mit Realexperimenten darstellen – obwohl sich empirisch keine hohen Korrelationen zwischen den Schülerleistungen in beiden Aufgabenformaten finden. Ein auf

diesem Ansatz aufbauendes Testverfahren für experimentelle Kompetenz, das hinsichtlich Reliabilität und Validität zufriedenstellend abgesichert ist, liegt jedoch bislang nicht vor.

3. Der MeK-LSA-Test

Ausgehend von den bisherigen Erkenntnissen zur Eignung vorliegender Testverfahren wurde in dem Verbundprojekt „Messung experimenteller Kompetenz in Large-Scale Assessments“ (MeK-LSA)[1] ein Testverfahren entwickelt, mit dem Schülerfähigkeiten in allen drei Bereichen des Experimentierens gemessen werden können: Planung, Durchführung und Auswertung. Zielgruppe sind Schülerinnen und Schüler am Ende der Sekundarstufe I. Über diese große Zielgruppe wird die Anschlussfähigkeit an Large-Scale-Erhebungen zum Bildungsmonitoring gesichert (z. B. PISA: [32]; IQB Ländervergleich: [4]). Der Test fokussiert auf die Kernbereiche experimenteller Fähigkeiten. Es geht nicht um die Messung allgemeiner Problemlösekompetenz, sondern um die Fähigkeiten, die notwendig sind, um experimentelle Verfahren für die Beantwortung physikalischer Fragen heranziehen zu können.

Im Folgenden wird zunächst die Konzeption des Tests erläutert, bevor die zur Verfügung stehenden Testmaterialien und das Verfahren zur Bewertung der Schülerleistungen vorgestellt werden.

3.1. Konzeption des Tests

Der Test ist vollständig am Bildschirm zu bearbeiten (on-screen-Test) und enthält anstelle von Realexperimenten interaktive Simulationen zum Aufbau von Experimenten und zur Durchführung von Messungen. Dies lässt gegenüber rein schriftlichen Verfahren eine größere inhaltliche Validität der Ergebnisse erwarten, da das Konstrukt vollständiger abgedeckt werden kann (vgl. [21]). Gleichzeitig wird ein effektiver Einsatz in (inzwischen ebenfalls meist on-screen durchgeführten) Large-Scale-Untersuchungen (z. B. PISA 2015, vgl. [33] und NAEP, vgl. [34]) ermöglicht. Beim Einsatz interaktiver Simulationen wird gegenüber einem Einsatz von Realexperimenten Zeit für Umbau und Wechsel von Experimentiermaterial eingespart. Dadurch können mehr Aufgaben eingesetzt werden. Zudem sind die Themen weder von der apparativen Ausstattung der Einzelschule abhängig noch müssen standardisierte Geräte auf die Schulen verteilt werden.

Grundlage der Testentwicklung war ein an vorliegenden Modellierungen experimenteller Kompetenz (vgl. 2.1; insb. [14]) anschließendes Aufgabenentwicklungsmodell, das drei Bereiche des Experimentierprozesses mit insgesamt acht Komponenten umfasst (Abb. 1).

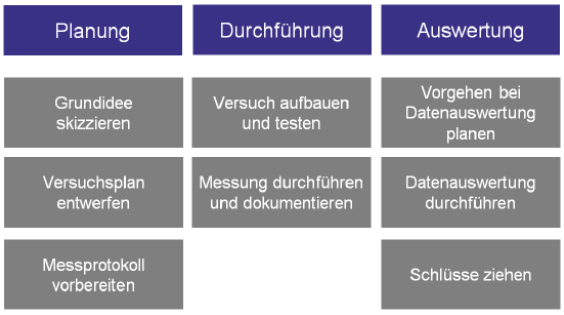


Abb. 1: Aufgabenentwicklungsmodell

Eine Testaufgabe (Unit) berücksichtigt jeweils sechs Komponenten, die in aufeinander aufbauenden Teilaufgaben (Items) umgesetzt sind. Bei jedem neuen Item wird eine Zwischenlösung des vorangegangenen Items angegeben, mit der die Probandin oder der Proband weiterarbeiten soll. Bei dem Item *Versuch aufbauen und testen* wird den Probandinnen und Probanden z. B. eine geeignete Versuchsplanung in Form der für den Aufbau benötigten Materialien, einer Skizze des Versuchsaufbaus sowie einer kurzen Beschreibung der Vorgehensweise als Zwischenlösung vorgegeben.

Die Unterteilung der Unit in einzelne Items und die Bereitstellung von Zwischenlösungen berücksichtigen den Befund von Schreiber ([31], S. 136), dass eine durchgängige eigenständige Bearbeitung einer experimentellen Aufgabe bei vielen Schülerinnen und Schülern zu Folgefehlern und Abbruch der Bearbeitung führt. Somit können z. B. Fähigkeiten zur Durchführung einer Messung nicht beurteilt werden, wenn Probandinnen oder Probanden nicht in der Lage sind, einen funktionsfähigen Aufbau zu erstellen. Durch die Zwischenlösungen werden hingegen Folgefehler abgefangen und die lokale statistische Unabhängigkeit der Items sichergestellt.

Die Angabe der Zwischenlösungen ist in eine gedachte Rahmenhandlung integriert: Die Schülerinnen und Schüler verfolgen parallel zu ihrer eigenen Testbearbeitung ein fiktives Schülerpaar („Bodo und Alina“), das die Aufgabe ebenfalls bearbeitet (Abb. 2). Die Zwischenlösungen werden eingeführt als „Bodo und Alina wollen ...“ (vgl. Abb. 3). Die Evaluation ergab, dass die Schülerinnen und Schüler diese Rahmung für Zwischenlösungen akzeptieren und dadurch nicht im eigenen Vorgehen demotiviert werden.

Die Items zur Durchführung (*Versuch aufbauen und testen* und *Messung durchführen und dokumentieren*) sind in jeder Unit enthalten. Sie werden umgeben von je zwei Items zur Planung und Auswertung. Das Item *Versuchsplan entwerfen* ist ebenfalls in jeder Unit enthalten und umfasst die Auswahl der benötigten Geräte aus einer virtuellen Materialbox, die Erstellung einer Versuchsskizze und eine kurze Beschreibung der geplanten Vorgehensweise beim

[1] Das Projekt der Universitäten Duisburg-Essen und Bremen sowie des IPN Kiel wurden vom BMBF im Rahmen des Programms zur Förderung von Forschungsvorhaben in Ankopplung an Large-Scale-Assessments gefördert (FKZ 01LSA005).

Experiment. Die Tabelle in Anhang A zeigt im Detail, welche Units welche Items enthalten. Im Anhang B ist eine Unit zum Zusammenhang von Stromstärke und Spannung bei einer Glühlampe (Unit „Kennlinie“) ausführlich dargestellt. Der Aufgabenstamm einer Unit enthält neben der übergeordneten Aufgabenstellung auch die zur Bearbeitung wichtigsten Fachinformationen, z. B. zur Bedeutung von „proportional“ bei Zusammenhängen physikalischer Größen. Abbildung 2 zeigt als Beispiel den Aufgabenstamm der Unit „Kennlinie“. Die übergeordnete Aufgabenstellung (1) und die Fachinformation (2) können während der Bearbeitung der nachfolgenden Items jederzeit wieder aufgerufen werden.

Abbildung 3 zeigt das Item *Versuch aufbauen und testen* der hier als Beispiel gewählten Unit. Im oberen Bereich sind die Zwischenlösung (1) und die Buttons zum Aufruf der übergeordneten Aufgabenstellung und der Fachinformation (2) zu erkennen. Darunter befindet sich die konkrete Aufgabenstellung für das Item und die interaktive Simulation zur Bearbeitung der Aufgabenstellung (3).

Die interaktiven Simulationen von Realexperimenten kommen in den Items zur Durchführung zum Einsatz. Im Item *Versuch aufbauen und testen* werden die benötigten Versuchsmaterialien (in der Optik z.B. Lampe, Linsen, Millimeterpapier, Lineal) zur Verfügung gestellt. Die Probandinnen und Probanden sollen damit selbständig die Versuchsanordnung aufbauen und deren Funktionsfähigkeit testen. Die Materialien können mit der Maus frei auf der virtuellen Arbeitsfläche platziert und konfiguriert werden. Beim Einschalten der Lampe (Beispiel Optik) werden die Lichtbündel gezeigt, die sich nach den Brechungs- und Reflexionsgesetzen ergeben. Die Simulationen bilden die Geräte von Realexperimenten und deren Funktionsweisen weitgehend realistisch nach und erlauben auch Fehlbedienung, z. B. die Zerstörung von Glühlampen in der Elektrizitätslehre.

Bei dem Item *Messung durchführen und dokumentieren* stehen als Zwischenlösung eine im Grundaufbau fertig konfigurierte Versuchsanordnung (zur Vermeidung von Folgefehlern aus einem möglicherweise ungeeigneten eigenen Aufbau) sowie ein Messprotokoll zur Verfügung. An dem Aufbau können die erforderlichen Einstellungen vorgenommen und die Messwerte abgelesen werden.

Die Items zur Planung und Auswertung enthalten ebenfalls interaktive Komponenten. So können die Probandinnen und Probanden bei der Auswahl von Versuchsmaterialien diese virtuell per Maus greifen und im Raum rotieren, um sie von verschiedenen Seiten zu betrachten. In das Item zur Datenauswertung ist ein Tool zur Erstellung von Diagrammen integriert und die Versuchsskizze wird mit einem einfachen Freihand-Grafiktool erstellt. Während der Bearbeitung eines Items werden alle eingetragenen Texte und Messwerte sowie die Interaktionen mit den interaktiven Simulationen mit Nutzerkennung und Zeitpunkt in einer Datenbank gespeichert. Damit stehen alle Itembearbeitungen für eine (zeitaufgelöste) datenbankbasierte Auswertung zur Verfügung (vgl. 3.3).

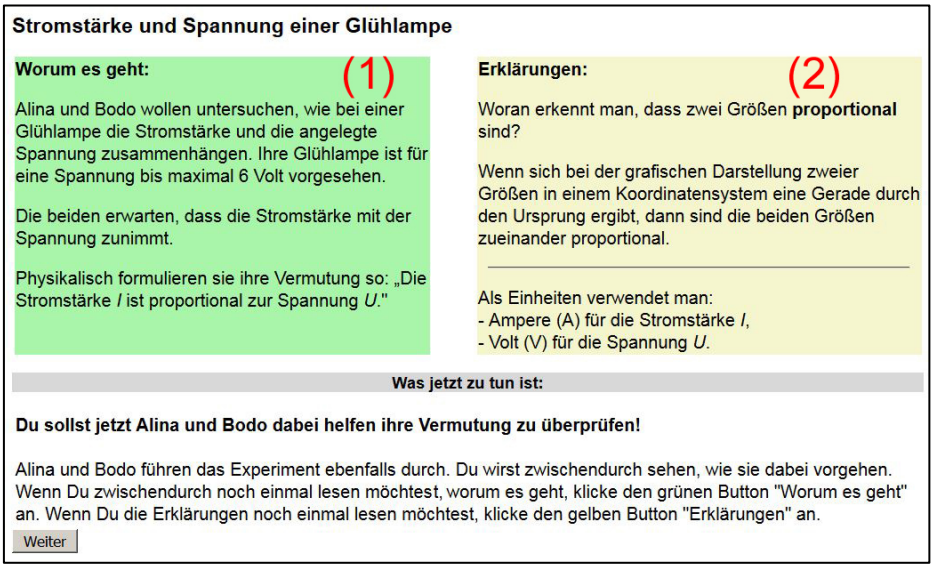


Abb. 2: Aufgabenstamm der Unit „Kennlinie“. (1) übergeordnete Aufgabenstellung, (2) Fachinformation

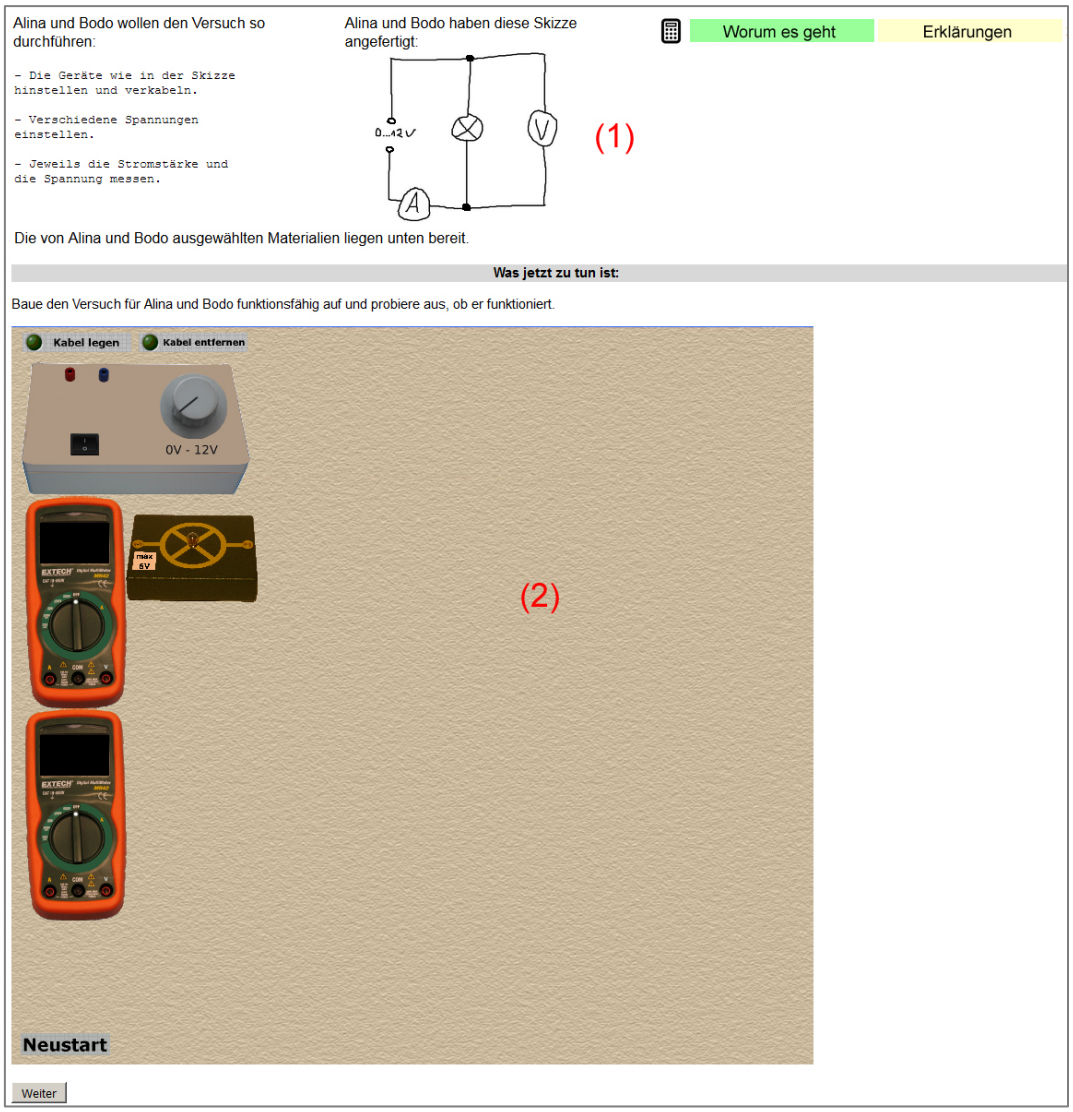


Abb. 3: Item Versuch aufbauen und testen der Unit „Kennlinie“. (1) Zwischenlösung, (2) Buttons zum Aufruf von Aufgabenstamm und Fachinformation, (3) interaktive Simulation für den Aufbau

3.2. Testmaterialien

Die Testmaterialien beinhalten zwölf Units zu experimentellen Aufgabenstellungen aus drei physikalischen Themenbereichen: geometrische Optik, Mechanik und Elektrizitätslehre (Stromkreise). Die Aufgabenstellungen orientieren sich an typischen Schülerexperimenten der Sekundarstufe I. Sie wurden mit Hilfe umfangreicher Lehrplan- und Schulbuchanalysen entwickelt [35]. Die Passung der Inhalte und Anforderungen der Aufgabenstellungen zur schulischen Praxis wurde durch eine Befragung von 53 erfahrenen Physiklehrkräften aus neun Bundesländern abgesichert sowie retrospektiv durch eine Schülerbefragung im Rahmen der Large-Scale-Erprobung des Tests überprüft (n = 1194; vgl. 4.). Die Tabelle im Anhang C gibt einen Überblick über die Testaufgaben. Aus dem Repertoire der Units lassen sich flexibel Testhefte als Erhebungsinstrumente zusammenstellen. Für die Bearbeitung einer Unit benötigten Schülerinnen und Schüler in den Erprobungen 10 bis 25 Minuten.

Zusätzlich zu den zwölf Units wurde eine weitere experimentelle Aufgabenstellung aus dem Bereich der Elektrizitätslehre zu einer Trainingsunit ausgearbeitet. Die Schülerinnen und Schüler können sich daran mit dem Aufbau des Tests und der Handhabung der interaktiven Komponenten vertraut machen.

Zu den entwickelten Materialien gehören weiterhin Testleitermanuale und detaillierte Kodiermanuale für die Bewertung der Itembearbeitungen.

3.3. Bewertung der Itembearbeitungen

Die Auswertung erfolgt mithilfe eines speziell für dieses Testverfahren entwickelten Software-Tools, das aus der Datenbank die Daten über die Itembearbeitungen ausliest und auf dem Bildschirm darstellt. Die Zustände der interaktiven Simulationen, Zeichnungen und Diagramme werden dabei aus den Da-

tenbankeinträgen auf dem Bildschirm rekonstruiert (Abb. 4). Dadurch wird die manuelle Auswertung effektiv unterstützt und eine auch im Large-Scale praktikable Auswertung erst ermöglicht. Die Bearbeitung des Items *Messung durchführen und dokumentieren* wird mit Hilfe eines Softwareskripts automatisch ausgewertet. Bewertet werden hierbei – für jede Unit angepasst – der Messbereich, die Anzahl der Messpunkte und die Passung zwischen dokumentierten und in der Simulation erzeugten Messwerten.

Abb. 4: Rekonstruktion der Schülerskizzen aus der Datenbank. Links: Drop-Down-Menü zur Auswahl von Schülercode (C27M), Unit (E1) und Zeitpunkt (4.7.2014, 12:27 Uhr). Rechts: rekonstruierter letzter Aufbau des Experiments („Kennlinie“) durch Schüler C27M.

Für die Bearbeitung von Items mit hohem Anteil an Interaktivität (*Versuchsplan entwerfen*, *Versuch aufbauen und testen*, *Messungen durchführen und dokumentieren*) sowie für Auswertungsitems, bei denen ein Diagramm erstellt werden muss, sind drei Bewertungsstufen vorgesehen: geeignet (Stufe 2), teilweise geeignet (Stufe 1) und ungeeignet (Stufe 0). Alle weiteren Items werden dichotom bewertet (geeignet oder ungeeignet). Im Kodiermanual sind für jedes Item inhaltliche Merkmale spezifiziert. Bei Items zum Aufbauen des Versuchs unterscheiden sich teilweise geeignete von geeigneten Lösungen dadurch, dass zwar der Grundaufbau korrekt ist, jedoch damit noch keine Messung möglich wäre, z.B. weil Stromstärke- oder Spannungsmessgeräte falsch in den Stromkreis eingebaut sind. Abbildung 5 zeigt drei Beispiellösungen. In analoger Weise werden die Items zur Planung und zur Durchführung einer Messung bewertet.

Die Objektivität der Kodierungen wurde durch Doppelkodierung von mindestens 10 % der Daten pro Item abgesichert. Für die Beurteilerübereinstimmung (Cohens Kappa [36]) ergeben sich zufriedenstellende bis sehr gute Werte ($.63 < \kappa < 1$, mittleres $\kappa = .84$). Der Zeitaufwand für die Kodierung einer Unit beträgt bei geschulten Kodierern für jeden Probanden ca. 3 Minuten.

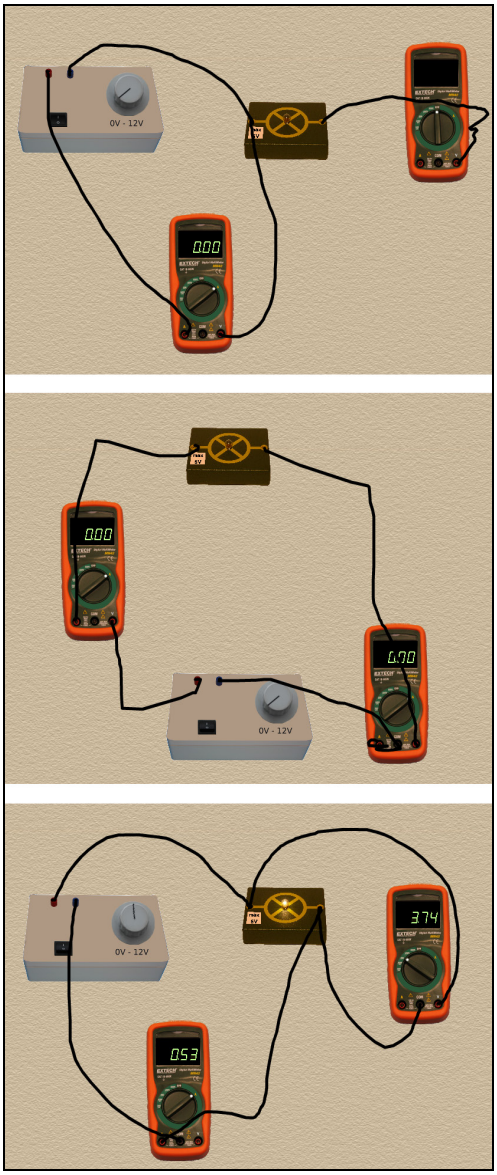


Abb. 5: Beispiele für die Bewertung des Aufbaus bei der Unit „Kennlinie“. Oben: kein funktionsfähiger Grundaufbau (0 Punkte), Mitte: Grundaufbau vorhanden, aber Messgeräte falsch angeschlossen (1 Punkt), unten: funktionsfähiger Aufbau (2 Punkte).

3.4. Validierungsstudien

Die Entwicklung des Tests wurde von umfangreichen Validierungsstudien begleitet. Eine Übersicht der Studien findet sich in [37] und [38]. In Anlehnung an die von der American Educational Research Association, der American Psychological Association und dem National Council on Measurement in Education [39] vorgeschlagenen Standards und orientiert am Validitätskonzept von Messick [21] wurden der inhaltliche, der kognitive, der externe und der strukturelle Aspekt untersucht. An dieser Stelle werden nur einige zentrale Ergebnisse der entsprechenden Studien zusammengefasst.

Die inhaltliche Validität der Aufgabenstellungen wurde mittels Lehrplananalysen ($n = 16$) und Lehr-

werkanalysen (n = 5) sowie Expertenbefragungen (n = 53) (vgl. 3.2) sichergestellt.

Zur Analyse der individuellen Vorgehensweisen der Schülerinnen und Schüler bei der Itembearbeitung (kognitiver Aspekt) wurden Studien mit Lautem Denken durchgeführt und die Überlegungen der Schülerinnen und Schüler zeitbasiert kodiert [40]. Es zeigt sich, dass experimentbezogene Überlegungen bei der Bearbeitung der on-screen Items klar im Vordergrund stehen und in vergleichbaren Anteilen auftreten, wie bei der Bearbeitung analoger Items mit Realexperimenten [40].

Hinsichtlich des externen Aspekts von Validität wurden die Testleistungen und die kognitive Belastung bei der Bearbeitung von on-screen Items und analog aufgebauten Items mit Realexperimenten verglichen. Hierbei zeigt sich eine hohe Konvergenz zwischen den Formaten [38]. Die Abgrenzung gegenüber Fachwissen und kognitiven Fähigkeiten (diskriminante Validität) und der strukturelle Aspekt der Validität (u. a. Abdeckung des Fähigkeitsspektrums) wurden im Rahmen einer Large-Scale-Erprobung überprüft [41].

4. Large-Scale-Erprobung des Tests

4.1. Stichprobe und Testdurchführung

Der Test wurde im Schuljahr 2014/2015 mit 1.194 Schülerinnen und Schülern aus 59 Klassen von 17 Schulen in vier Bundesländern am Ende der Sekundarstufe I durchgeführt.

Für die Erhebung wurden in einem Multi-Matrix-Design zwölf Testhefte mit je vier Units aus zwei verschiedenen Inhaltsbereichen zusammengestellt (Abb. 6). Jedes Testheft wurde von ca. 100 Schülerinnen und Schülern, jede Unit somit von ca. 400 Personen bearbeitet. Als Bearbeitungszeit wurden 25 Minuten pro Unit festgelegt (vgl. 3.2).

A	B	C	D	E	F	G	H	K	L	M	N
M1	M2	M4	M3	E1	E2	E3	E4	O1	O3	O2	O4
M4	M3	M1	M2	E3	E4	E1	E2	O2	O4	O1	O3
O1	O3	E1	E2	M2	M1	O2	O4	M3	M4	E4	E3
O2	O4	E3	E4	M3	M4	O1	O3	M2	M1	E2	E1
100	100	100	100	100	100	100	100	100	100	100	100

Abb. 6: Testheftdesign für die Large Scale Erprobung. EX: Elektrizitätslehre-Unit, MX: Mechanik-Unit, OX: Optik-Unit, Details siehe Anhang C

Vor der Bearbeitung der Testhefte führten die Schülerinnen und Schüler unter Anleitung durch Testleiter die Trainingsaufgabe durch (vgl. 3.2). Der Ablauf der Erhebung wurde durch detaillierte Testleitermanuale in möglichst hohem Maße standardisiert und – einschließlich etwaiger Besonderheiten – in Protokollbögen dokumentiert.

4.2. Skalierungsergebnisse

Die Bewertung der Itembearbeitungen erfolgte wie oben beschrieben anhand eines Kodiermanuals (vgl. 3.3). Die Daten wurden anschließend auf Grundlage des Partial Credit Rasch Modells [42] mit Hilfe der Statistiksoftware R und des Test Analysis Moduls (TAM, vgl. [43]) skaliert und anschließend auf die PISA-Skala transformiert (Mittelwert der Schülerfähigkeiten 500, Standardabweichung 100 Punkte, Lösungswahrscheinlichkeit .625).

Die Skalierung ergab eine sehr gute Passung des Partial Credit Rasch Modells auf die Daten. Die üblicherweise zur Beurteilung der Modellpassung herangezogenen Mean Square (MNSQ) Fit Statistiken weisen für 70 der insgesamt 72 Items (vgl. Anhang A) eine gute Modellpassung (0.80 < weighted MNSQ < 1.20) und für alle 72 Items eine befriedigende Modellpassung (0.70 < weighted MNSQ < 1.30) aus (vgl. [44] oder [45]). Die Analyse der Aufgabenschwierigkeiten (s. Abb. 7) belegt die Eignung der Aufgaben für die Erfassung der Schülerfähigkeiten in der Stichprobe. Der Test ist damit potentiell für die im Rahmen von Large-Scale-Assessments wie PISA untersuchte Population von 15-Jährigen geeignet. Die Differenz zwischen der mittleren Aufgabenschwierigkeit und der mittleren Personenfähigkeit mit 37 Punkten auf der PISA-Metrik ist vernachlässigbar. Die in Abbildung 6 dargestellte Gegenüberstellung der Aufgabenschwierigkeiten zu den Personenfähigkeiten zeigt eine gute Abdeckung des Fähigkeitsspektrums der Schülerinnen und Schüler durch die Aufgaben. Dies drückt sich auch in der sehr guten Reliabilität (α_{WLE} = .84) der Schätzer für die Personenfähigkeit (WLE) bei 24 bearbeiteten Teilaufgaben pro Person aus. Insgesamt ist der Test in seiner Schwierigkeit sehr gut auf die Population abgestimmt. Es lassen sich sowohl leistungsschwache als auch leistungsstarke Personen unterscheiden sowie insbesondere auch Personen im mittleren Leistungsbereich bezüglich ihrer Fähigkeiten klassifizieren.

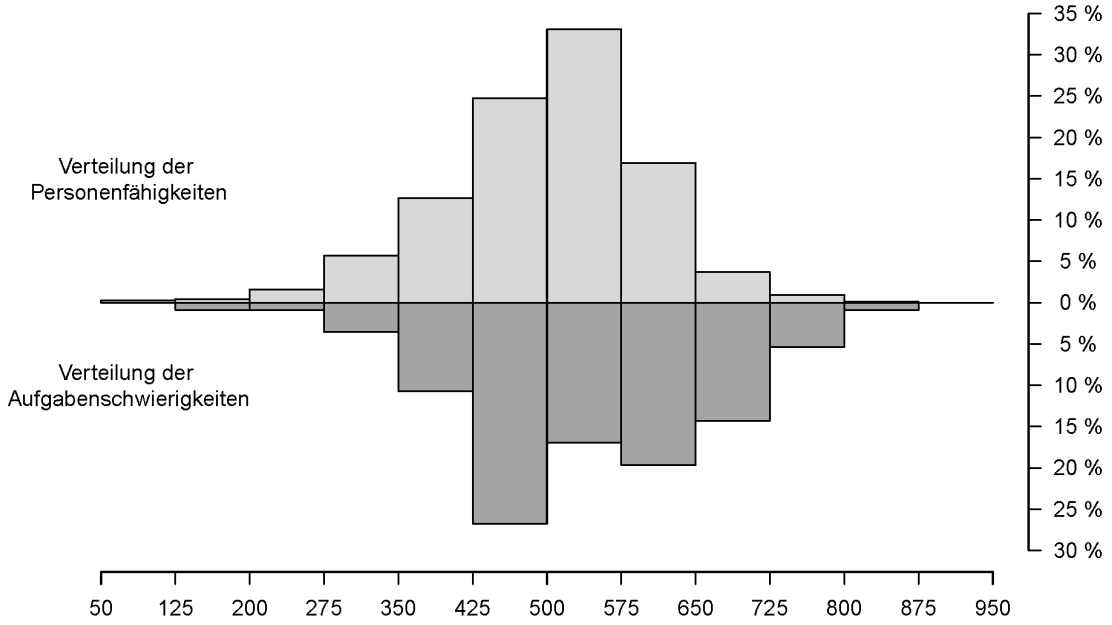


Abb. 7: Verteilung der Aufgabenschwierigkeiten gegenüber den Personenfähigkeiten (WLE) auf der PISA-Skala (siehe Text)

5. Inhaltliche Interpretation der Itemschwierigkeiten

Die Ergebnisse der Validierungsstudien (vgl. 3.4), die hohe interne Konsistenz und die gute Abdeckung des Fähigkeitsspektrums (vgl. 4.2) sind bereits Belege für die Aussagekraft des Tests. Weiter absichern lässt sich die Qualität, wenn sich die empirisch gefundenen Schwierigkeiten der Testaufgaben auch inhaltlich begründen lassen. Dazu werden in diesem Abschnitt Detailanalysen vorgestellt, die sich auf den zentralen Itemtyp zur Planung (*Versuchsplan entwerfen*) und die Items zur Durchführung mit den interaktiven Simulationen (*Versuch aufbauen und testen*; *Messung durchführen und dokumentieren*) beziehen. Im Folgenden werden diese Items kurz als Planungs-, Aufbau- und Messitems bezeichnet.

Abbildung 8, in der die Items nach Schwierigkeit geordnet aufgetragen sind (auf der PISA-Skala, vgl. 4.2), zeigt, dass bereits die Planungs-, Aufbau- und Messitems den Schwierigkeitsbereich vollständig abdecken. Da diese Items zweistufig bewertet wurden (vgl. 3.3), sind sie in der Abbildung jeweils zweimal vertreten („gelöst auf Stufe 1“ bzw. „gelöst auf Stufe 2“). Es fällt auf, dass bis auf wenige Ausnahmen die Items auf Stufe 1 unterhalb der mittleren Schwierigkeit von 500 Punkten liegen, die Items auf Stufe 2 oberhalb. Die mit dem Bewertungsmaßstab intendierte Differenzierung von Schwierigkeiten gelingt also nicht nur innerhalb eines Items (das ist durch die zweistufige Kodierung gegeben), sondern über alle Itemtypen (Aufbau-, Mess- und Planungsitems) und Inhaltsbereiche hinweg.

Im Folgenden werden die einzelnen Itemtypen, insbesondere Aufbau- und Messitems, bezüglich der inhaltlichen Anforderungen und der beobachteten Schwierigkeiten diskutiert.

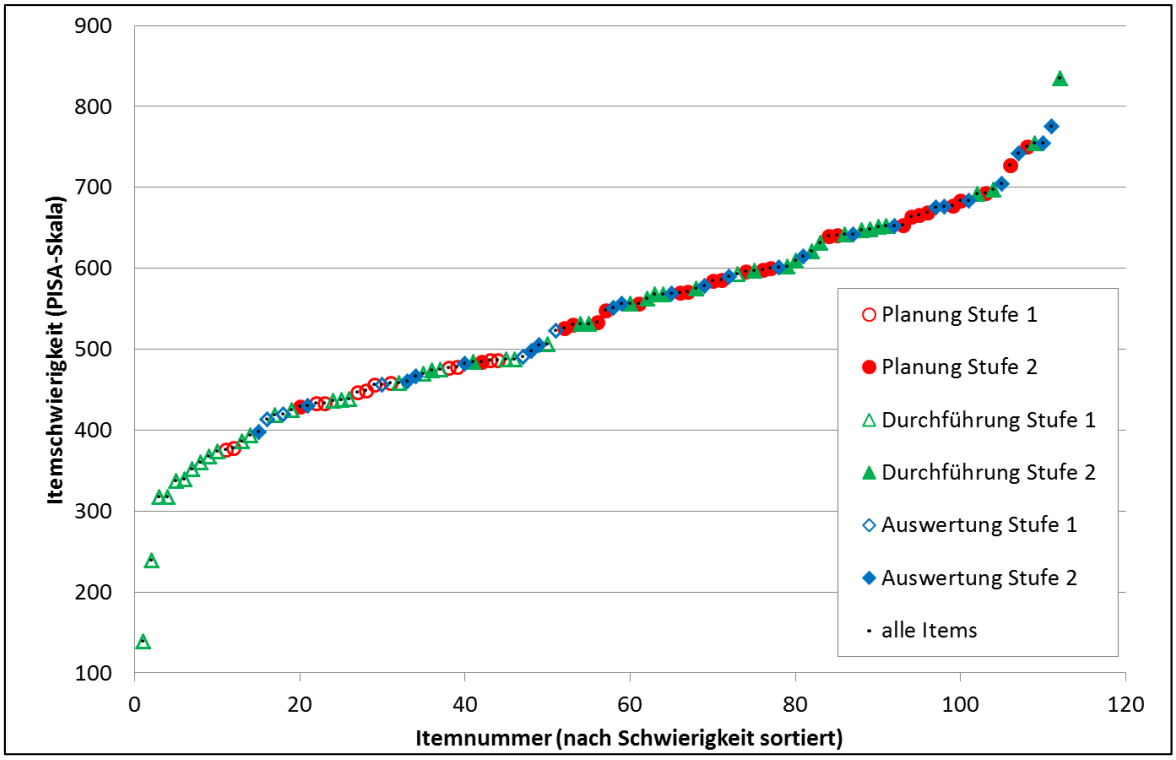


Abb. 8: Itemschwierigkeiten; Planung-, Aufbau- und Messitems hervorgehoben

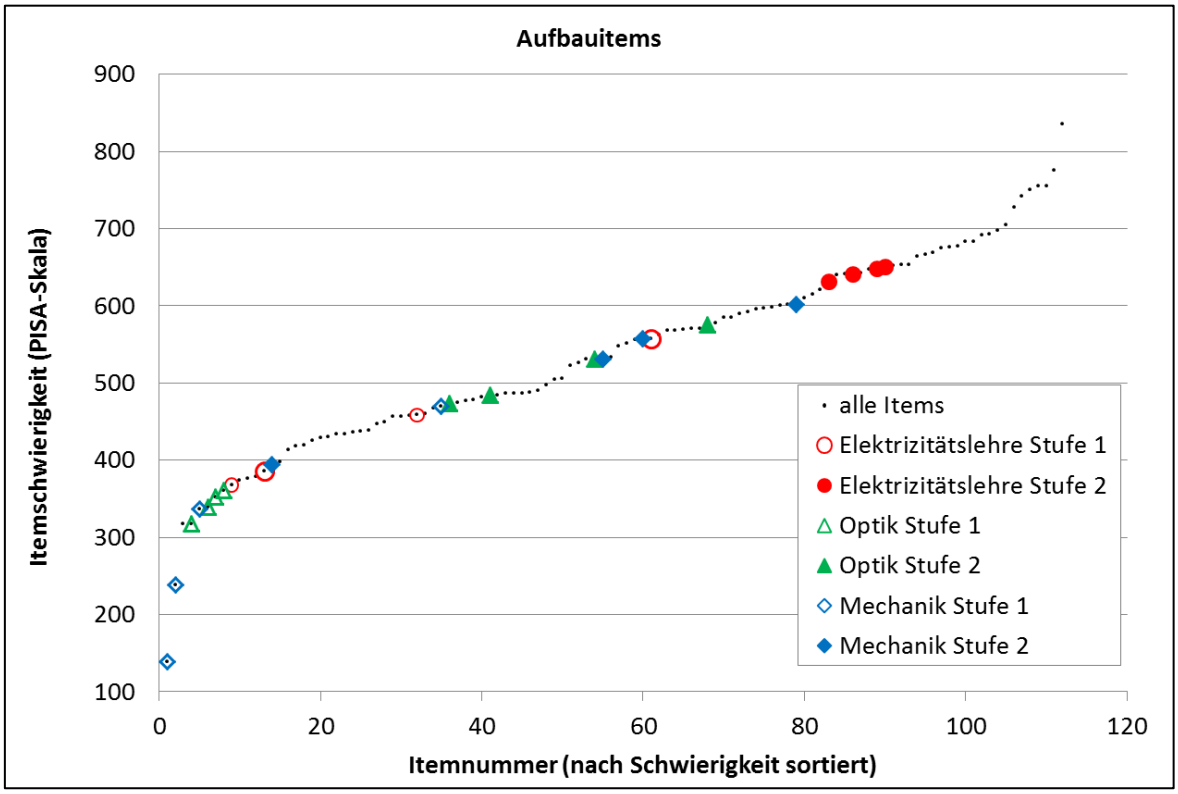


Abb. 9: Schwierigkeiten der Aufbauitems, nach Inhaltsbereichen und Bewertungsstufen getrennt

5.1. Anforderungen und Schwierigkeiten der Aufbauitems

In Abbildung 9 sind die Itemschwierigkeiten für die Aufbauitems, nach Inhaltsbereichen und Bewertungsstufen getrennt, farblich hervorgehoben.

Bei den Aufbauitems stehen eine Skizze des Versuchsaufbaus, das notwendige Material (on-screen) und eine kurze Beschreibung der geplanten Vorgehensweise zur Verfügung. Mit den auf Stufe 1 bewerteten Aufbauitems lassen sich die Schülerfähigkeiten am unteren Ende des Fähigkeitsspektrums auflösen. Um als „teilweise geeignet“ (Stufe 1) kodiert zu werden, muss der in der Skizze vorgegebene Aufbau mit dem vorgegebenen Material in Grundzügen „nachgebaut“ werden, was für die Schülerinnen und Schüler offenbar vergleichsweise leicht ist. Dies gilt insbesondere in der Optik und Mechanik (Abb. 9), wo gegenständliche Skizzen vorliegen. Die Skizzen in der Elektrizitätslehre sind hingegen Schaltskizzen, die zunächst auf einer höheren Abstraktionsebene entschlüsselt werden müssen (eine Erklärung der verwendeten Schaltsymbole wird jeweils beim Planungsitem gegeben). Abbildung 10 zeigt die Musterlösung, die bei einem der einfachsten Aufbauitems aus der Optik (318 Punkte auf der PISA-Skala) zur Verfügung gestellt wird. Ein Nachbau dieser Anordnung in der Simulation kann ohne spezifische Kenntnisse optischer Geräte erfolgreich sein.

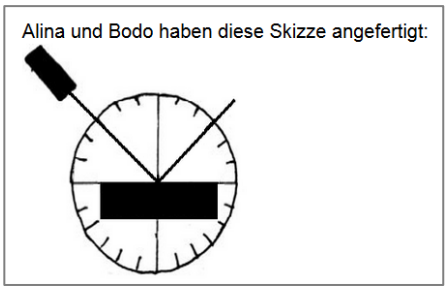


Abb. 10: Skizze, die beim Aufbauitem der Unit O2 vorgegeben ist

Wird bei einem Aufbauitem in der Mechanik oder Optik nur Stufe 1 erreicht, so liegt das in der Regel an einer unvollständigen Übertragung der Skizze. Um Stufe 2 zu erreichen, muss der Aufbau geeignet sein, um die in der Vorgehensweise angegebene Messung durchzuführen. In der Elektrizitätslehre kommt hier als Anforderung hinzu, dass mindestens ein Messgerät korrekt angeschlossen und eingestellt sein muss. Letzteres kann man der vorgegebenen Skizze nicht entnehmen. Das erklärt vermutlich, warum die Aufbauitems der Elektrizitätslehre innerhalb der Gruppe der Aufbauitems auch auf Stufe 2 die größte Schwierigkeit aufweisen (Abb. 9).

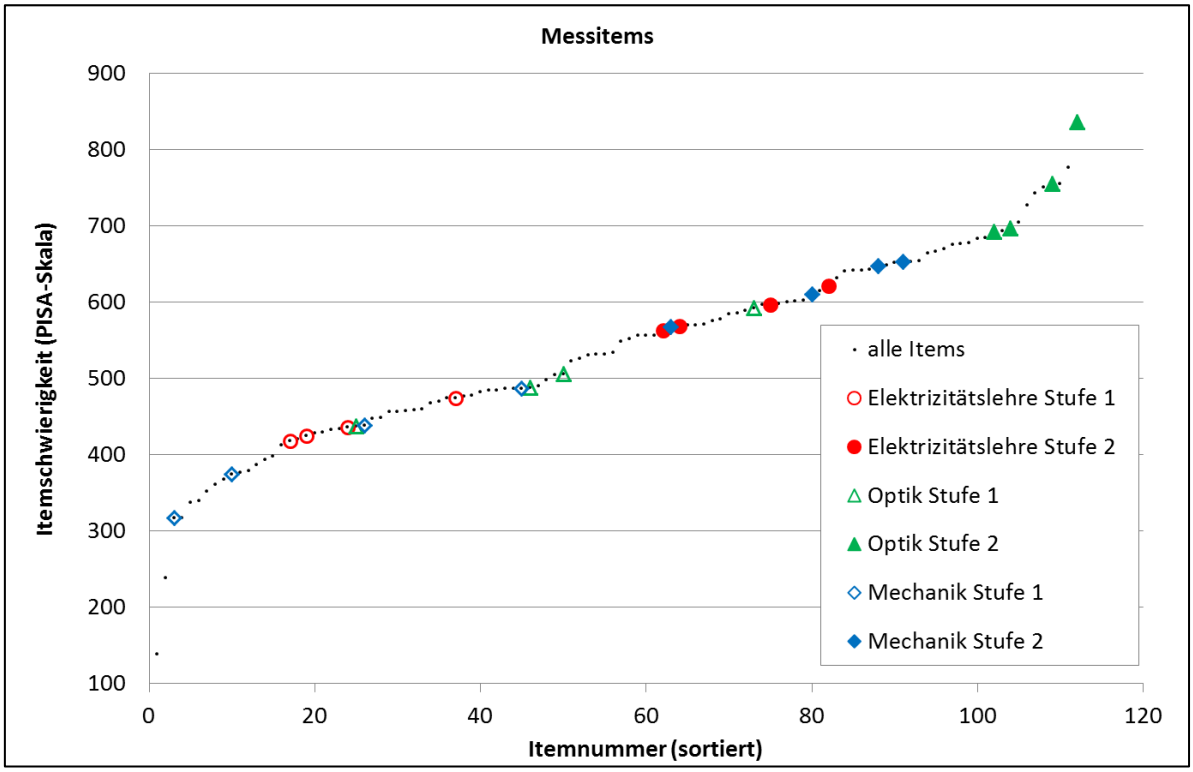


Abb. 11: Schwierigkeiten der Messitems, nach Inhaltsbereichen und Bewertungsstufen getrennt

5.2. Anforderungen und Schwierigkeiten der Messitems

Abbildung 11 zeigt die Schwierigkeiten der Messitems. Sie liegen tendenziell im oberen Bereich des Schwierigkeitsspektrums. Insbesondere die Messitems der Optik auf Stufe 2 sind geeignet, das obere Ende des Fähigkeitsspektrums aufzulösen.

Bei allen Messitems stehen ein funktionsfähiger Aufbau (on-screen) und ein vorbereitetes Messprotokoll zur Verfügung. Darüber hinaus wird die Vorgehensweise (welche Größen zu variieren und zu messen sind) in Stichworten vorgegeben.

Um Stufe 1 zu erreichen, ist die Aufnahme eines korrekten Messwertes erforderlich. Korrekt bedeutet dabei, dass der in der Simulation eingestellte Wert mit dem im Messprotokoll eingetragenen Wert übereinstimmt. Dies fällt den Schülerinnen und Schülern offenbar relativ leicht, wenngleich es in der Optik tendenziell schwieriger ist als in der Mechanik und Elektrizitätslehre. Dies kann darauf zurückzuführen sein, dass in der Elektrizitätslehre und der Mechanik Werte von digitalen Displays oder alltagstypischen Messgeräten, wie Maßstab und Stoppuhr, direkt ablesbar sind. In der Optik sind hingegen in der Regel Einfalls-, Brechungs- oder Reflexionswinkel auf einer unbeschrifteten Winkelscheibe abzulesen. Hier können Fehler z. B durch eine falsche Interpretation der Skalierung der Winkelscheibe oder die falsche Wahl der Schenkel für die Winkelmessung auftreten.

Um bei den Messitems Stufe 2 zu erreichen, muss eine vollständige Messreihe weitgehend korrekt und in einem sinnvoll selbst gewählten Bereich von Messwerten aufgenommen werden. Auch hier erweisen sich die Optikitems wieder als tendenziell schwieriger und sind geeignet, das obere Ende des Fähigkeitsspektrums der Schülerinnen und Schüler aufzulösen.

5.3. Vergleich der Itemtypen hinsichtlich der Schwierigkeit

Im Vergleich der Itemtypen sind die Aufbauitems sowohl innerhalb von Stufe 1 als auch innerhalb von Stufe 2 tendenziell einfacher als die Planungs- und Messitems. Das passt zu Befunden aus Videostudien, nach denen im Unterricht viel Zeit auf die Durchführung und deutlich weniger Zeit auf die Planung von Experimenten verwendet wird und die Schülerinnen und Schüler in der Planung fast nie Gelegenheit haben, eigene Ideen einzubringen [46]. Die Schülerinnen und Schüler sind daher vermutlich geübt darin, Experimente nach Anleitung aufzubauen und nutzen die Musterlösung aus dem Planungsitem im Aufbauitem als detaillierte Vorlage.

Die hohe Schwierigkeit der Messitems lässt sich vermutlich darauf zurückführen, dass Schülerinnen und Schüler im Unterricht zwar häufig Messungen durchführen, jedoch dabei keine eigenen Überlegungen zur Anzahl der Messpunkte und zum Wertebereich anstellen müssen.

6. Zusammenfassung und Ausblick

Mit dem in diesem Beitrag vorgestellten Test können experimentelle Fähigkeiten von Schülerinnen und Schülern am Ende der Sekundarstufe I diagnostiziert werden. Als Vorteil gegenüber anderen Instrumenten wird durch den Einsatz interaktiver Simulationen insbesondere der Bereich der Durchführung erfasst.

Die Large-Scale-Erprobung hat eine hohe Reliabilität der Schätzer für die Personenfähigkeit (vgl. 4.2) und eine sehr gute Passung der Itemschwierigkeiten zu den Schülerfähigkeiten ergeben. Die empirisch gefundenen Itemschwierigkeiten können inhaltlich plausibel erklärt werden. Dies ist neben den Ergebnissen der Studien zur inhaltlichen, kognitiven und konvergenten Validität ein weiterer Hinweis darauf, dass die Testergebnisse valide Aussagen über die experimentellen Fähigkeiten der Schülerinnen und Schüler zulassen.

Die Erprobung mit mehr als 1100 Schülerinnen und Schülern hat gezeigt, dass das Testverfahren sowohl bei der Durchführung der Erhebung als auch bei der Auswertung der anfallenden Datenmengen bei großen Stichproben praktikabel und effizient ist. Somit sind die wesentlichen Voraussetzungen für einen Einsatz des Testverfahrens im Rahmen des nationalen Bildungsmonitorings gegeben. Die notwendigen Adaptionen liegen im Wesentlichen darin, den Test an die jeweils aktuellen technischen Rahmenbedingungen für die online-Erhebung anzupassen. Der zu erwartende Ertrag aus einer Erhebung mit einer repräsentativen Stichprobe besteht in Erkenntnissen über die Ausprägung experimenteller Fähigkeiten deutscher Schülerinnen und Schüler am Ende der Sekundarstufe I. Daraus ließen sich spezifische Förderbedarfe in diesem Bereich fachmethodischer Kompetenzen ableiten.

Neben dem Einsatz im Rahmen des Bildungsmonitorings eröffnen die Ergebnisse des Projektes auch die Möglichkeit, die Wirksamkeit von Interventionen zur Förderung experimenteller Fähigkeiten effizient zu untersuchen und diese Interventionen auf Basis der Evaluationsergebnisse gezielt zu optimieren. Dafür ist jedoch noch der Nachweis zu erbringen, dass das Testverfahren ausreichend empfindlich misst, d.h. in der Lage ist, die Entwicklung experimenteller Fähigkeiten über vergleichsweise kurze Interventionszeiträume von wenigen Wochen oder Monaten hinweg aufzulösen.

7. Literatur

[1] Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (2005): Bildungsstandards im Fach Physik für den Mittleren Schulabschluss. Beschluss vom 16.12.2004. München: Luchterhand.

[2] NGSS Lead States (2013). Next Generation Science Standards: For States, By States.

Washington, DC: The National Academies Press.
[3] Klos, S. (2009): Kompetenzförderung im naturwissenschaftlichen Anfangsunterricht - Der Einfluss eines integrierten Unterrichtskonzepts, Berlin: Logos.
[4] Pant. H. A.; Stanat, P.; Schroeders, U.; Roppelt, A.; Siegle, T. & Pöhlmann, C. (Hrsg) (2013). IQB-Ländervergleich 2012. Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I. Münster: Waxmann.
[5] Straube, P. & Nordmeier. V. (2012). KoWADiS - Wohin geht es? In V. Nordmeier & H. Grötzebauch (Hrsg.), PhyDid B, Didaktik der Physik, Beiträge zur DPG-Frühjahrstagung, Mainz 2012). http://phydid.physik.fu-berlin.de/index.php/phydid-b/article/view/409/548 [24.4.2016].
[6] Stecher, B. M. & Klein, S. P. (1997): The Cost of Performance Assessments in Large-Scale Testing Programs. In: Educational Evaluation and Policy Analysis, 19(1), S. 1–14.
[7] Kircher, E.; Girwidz, R. & Häußler, P. (2010): Physikdidaktik: Theorie und Praxis, Berlin, Heidelberg: Springer.
[8] Schulz, A., Wirtz, M. & Starauschek, E. (2012). Das Experiment in den Naturwissenschaften. In W. Rieß, M. Wirtz, B. Barzel. & A. Schulz (Hrsg.), Experimentieren im mathematisch-naturwissenschaftlichen Unterricht. Schüler lernen wissenschaftlich denken und arbeiten (S. 15-38). Münster: Waxmann.
[9] Adamia, M.; Labudde, P.; Gingins, F.; Nidegger, C.; Bazzigher, L.; Bringold, B. et al. (2008): HarmoS Naturwissenschaften+: Kompetenzmodell und Vorschläge für Bildungsstandards. Wissenschaftlicher Schlussbericht, Bern: HarmoS Konsortium Naturwissenschaften+.
[10] Mayer, J.; Grube, C. & Möller, A. (2008): Kompetenzmodell naturwissenschaftlicher Erkenntnisgewinnung. In Harms, U.; Sandmann, A. (Hrsg.): Lehr- und Lernforschung in der Biologiedidaktik Bd.3, S. 63-79. Innsbruck: Studienverlag.
[11] Schreiber, N.; Theyßen, H. & Schecker, H. (2009): Experimentelle Kompetenz messen?! In: Physik und Didaktik in Schule und Hochschule, 8(3), S. 92-101. URL: http://www.phydid.de
[12] Wirth, J.; Thillmann, H.; Künsting, J.; Fischer, H. E. & Leutner, D. (2008): Das Schülerexperiment im naturwissenschaftlichen Unterricht - Bedingungen der Lernförderlichkeit dieser Lehrmethode. In: Zeitschrift für Pädagogik, 54(3), S. 361-375.
[13] Emden, M. & Sumfleth, E. (2012): Prozessorientierte Leistungsbewertung des experimentellen Arbeitens. Zur Eignung einer Protokollmethode zur Bewertung von Experimentierprozessen. In: Der mathematische und naturwissenschaftliche Unterricht, 65(2), S. 68-75.
[14] Nawrath, D.; Maiseyenka, V. & Schecker, H. (2011): Experimentelle Kompetenz - Ein Modell für die Unterrichtspraxis. In: Praxis der Naturwissenschaften - Physik in der Schule, 60(6), S. 42-48.
[15] Henke, C. (2007): Experimentell-naturwissenschaftliche Arbeitsweisen in der Oberstufe: Untersuchung am Beispiel des HIGHSEA-Projekts in Bremerhaven, Berlin: Logos.
[16] Rumann, S. (2005): Kooperatives Arbeiten im Chemieunterricht: Entwicklung und Evaluation einer Interventionsstudie zur Säure-Base-Thematik, Berlin: Logos.
[17] Walpuski, M. (2006): Optimierung von experimenteller Kleingruppenarbeit durch Strukturierungshilfen und Feedback: Eine empirische Studie, Berlin: Logos.
[18] Baxter, G. P.; Shavelson, R. J.; Goldman, S. R. & Pine, J. (1992): Evaluation of Procedure-Based Scoring for Hands-On Science Assessment. In: Journal of Educational Measurement, 29 (1), S. 1-17.
[19] Stecher, B. M.; Klein, S. P.; Solano-Flores, G.; McCaffrey, D.; Robyn, A.; Shavelson, R. J. et al. (2000): The Effects of Content, Format and Inquiry Level on Science Performance Assessment Scores. In: Applied Measurement in Education, 13(2), S. 139-160.
[20] Schreiber, N.; Theyßen, H. & Schecker, H. (2014): Diagnostik experimenteller Kompetenz: Kann man Realexperimente durch Simulationen ersetzen? In: Zeitschrift für Didaktik der Naturwissenschaften, 20(1), S. 161-173.
[21] 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.
[22] Ayala, C. C.; Shavelson, R. J. & Yin, Y. (2002): Reasoning Dimensions Underlying Science Achievement: The Case of Performance Assessment 1610. In: Educational Assessment, 8(2), S. 101-121.
[23] Baxter, G. P.; Elder, A. D. & Glaser, R. (1996): Knowledge-Based Cognition and Performance Assessment in the Science Classroom. In: Educational Psychologist, 31(2), S. 133-140.
[24] Baxter, G. P.; Shavelson, R. J.; Goldman, S. R. & Pine, J. (1992): Evaluation of Procedure-Based Scoring for Hands-On Science Assessment. In: Journal of Educational Measurement, 29(1), S. 1-17.
[25] Shavelson, R.; Baxter, G. & Pine, J. (1991): Performance Assessment in Science. In: Applied Measurement in Education, 4(4), S. 347-362.

[26] Stebler, R.; Reusser, K. & Ramseier, E. (1997): Spitzenleistungen der Schweizer Siebtklässler im TIMSS Experimentiertest 1721. In: SLZ (10), S. 18-21.
[27] Gut-Glanzmann (2012): Modellierung und Messung experimenteller Kompetenz. Analyse eines large-scale Experimentiertests, Berlin: Logos Verlag.
[28] Koretz, D.; Stecher, B. M.; Klein, S. & McCaffrey, D. (1994): The Vermont portfolio assessment program: Findings and implications. In: Educational Measurement: Issues and Practices, 13(3), S. 5-16.
[29] Shavelson, R. J.; Baxter, G. P. & Gao, X. (1993): Sampling Variability of Performance Assessments. In: Journal of Educational Measurement, 30(3), S. 215-232.
[30] Shavelson, R. J.; Ruiz-Primo, M. A. & Wiley, E. W. (1999): Notes on Sources of Sampling Variability in Science Performance Assessments. In Journal of Educational Measurement, 36(1), S. 61-71.
[31] Schreiber, N. (2012): Diagnostik experimenteller Kompetenz – Validierung technologiegestützter Testverfahren im Rahmen eines Kompetenzstrukturmodells, Berlin: Logos.
[32] OECD (1999): Measuring Student Knowledge and Skills. A New Framework for Assessment, Paris: OECD.
[33] http://zib.education/fileadmin/user_upload/PDFs/PISA/21_Pisa-Bro_2015_Lay01_141201_Einzelseiten.pdf [20.5.2016].
[34] 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.
[35] Dickmann, M. & Theyßen, H. (2013): Curriculare Validität von Units zur Messung experimenteller Kompetenz. In: Bernholt, S. (Hrsg.): Inquiry-based Learning - Forschendes Lernen: Gesellschaft für Didaktik der Chemie und Physik, Jahrestagung in Hannover 2012, S. 587-589. Kiel: IPN.
[36] Hammann, M.; Jördens, J. & Schecker, H. (2014): Übereinstimmung zwischen Beurteilern: Cohens Kappa (κ). In Krüger, D.; Parchmann, I.; Schecker, H. (Hrsg.): Methoden in der naturwissenschaftsdidaktischen Forschung. Berlin: Springer.
[37] Theyßen, H.; Schecker, H.; Neumann, K.; Dickmann, M. & Eickhorst, B. (2013): Messung experimenteller Kompetenz in Large-Scale Assessments. In: Bernholt, S. (Hrsg.): Inquiry-based Learning - Forschendes Lernen: Gesellschaft für Didaktik der Chemie und Physik, Jahrestagung in Hannover 2012, S. 596-598. Kiel: IPN.
[38] Dickmann, M. (2016). Messung von Experimentierfähigkeiten. Validierungsstudien zur Qualität eines computerbasierten Testverfahrens. Dissertation Universität Duisburg-Essen.
[39] AERA, APA & NCME (2014). Standards for educational and psychological testing. Washington D.C.: American Educational Research Association.
[40] Theyßen, H., Dickmann, M., Neumann, K., Schecker, H. & Eickhorst, B. (2016). Measuring Experimental Skills in Large Scale Assessments: A Simulation-Based Test Instrument. In J. Lavonen, K. Juuti, J. Lampiselkä, A.; Uitto & K. Hahl (Eds.), Electronic Proceedings of the ESERA 2015 Conference. Science education research: Engaging learners for a sustainable future, Part 11 (co-ed. J. Dolin & P. Kind), (pp. 1598 - 1606). Helsinki, Finland: University of Helsinki. ISBN 978-951-51-1541-6.
[41] Theyßen, H.; Schecker, H.; Dickmann, M.; Eickhorst, B. & Neumann, K. (2016). Messung experimenteller Kompetenz in Large-Scale-Assessments. In BMBF (Hrsg.): Forschungsvorhaben in Ankopplung an Large-Scale-Assessments. Bildungsforschung, Band 44 (S. 83 - 96). Bonn, Berlin: BMBF.
[42] Masters, G. N. (1982): A Rasch model for partial credit scoring. In: Psychometrika, 47(2), S. 149–174.
[43] Kiefer, T.; Robitzsch, A. & Wu, M. (2015): Tam: Test analysis modules. Computer software manual. https://cran.r-project.org/web/packages/TAM/TAM.pdf [11.7.2015] (R package version 1.3).
[44] Neumann, I.; Neumann, K.; & Nehm, R. (2011): Evaluating Instrument Quality in Science Education: Rasch-based analyses of a Nature of Science test. In: International Journal of Science Education, 10(33), S. 1–33.
[45] Bond, T. G. & Fox, C. M. (2007): Applying the Rasch model: Fundamental measurement in the human sciences (2nd ed.), Mahwah; N.J: Lawrence Erlbaum Associates Publishers.
[46] Tesch, M. & Duit, R. (2004): Experimentieren im Physikunterricht – Ergebnisse einer Videostudie. In: ZfDN, 10, S. 51-69.

Anhang A: Zusammenstellung der Units

In der folgenden Tabelle ist für alle Units dargestellt, welche Modellkomponenten (Abb. 1) darin als Items repräsentiert sind.

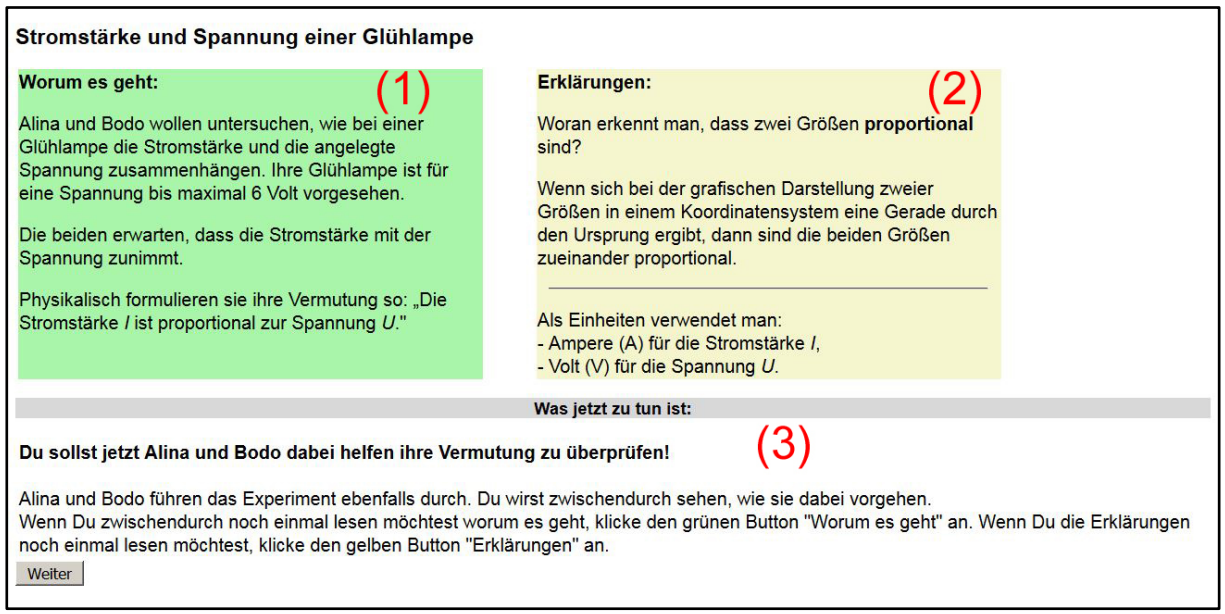
Modellkomponente bzw. Item → **in Unit ↓**	**Grundidee skizzieren**	**Versuchsplan entwerfen**	**Messprotokoll vorbereiten**	**Versuch aufbauen und testen**	**Messung durchführen und testen**	**Vorgehen bei Datenauswertung planen**	**Datenauswertung durchführen**	**Schlüsse ziehen**
U-I-Kennlinie einer Glühlampe	X	X		X	X		X	X
Leistung von Glühlampen		X	X	X	X	X	X	
Parallelschaltung von Glühlampen	X	X		X	X	X	X	
Reihenschaltung von Glühlampen	X	X		X	X	X		X
Lichtbrechung am Halbkreisblock	X	X		X	X		X	X
Reflexion am Halbkreisblock		X	X	X	X	X		X
Totalreflexion		X	X	X	X	X		X
Brennweitenbestimmung einer Linse		X	X	X	X	X		X
Dichtebestimmung		X	X	X	X	X	X	
Ausdehnung eines Gummibandes	X	X		X	X		X	X
Auftriebskraft in Wasser		X	X	X	X	X	X	
Fahrzeit auf der schiefen Ebene	X	X		X	X		X	X
Gesamt	**6**	**12**	**6**	**12**	**12**	**8**	**8**	**8**

Anhang B: Unit „Kennlinie“

Die folgenden Seiten zeigen Bildschirmkopien der Units zur Stromstärke-Spannungskennlinie einer Glühlampe.

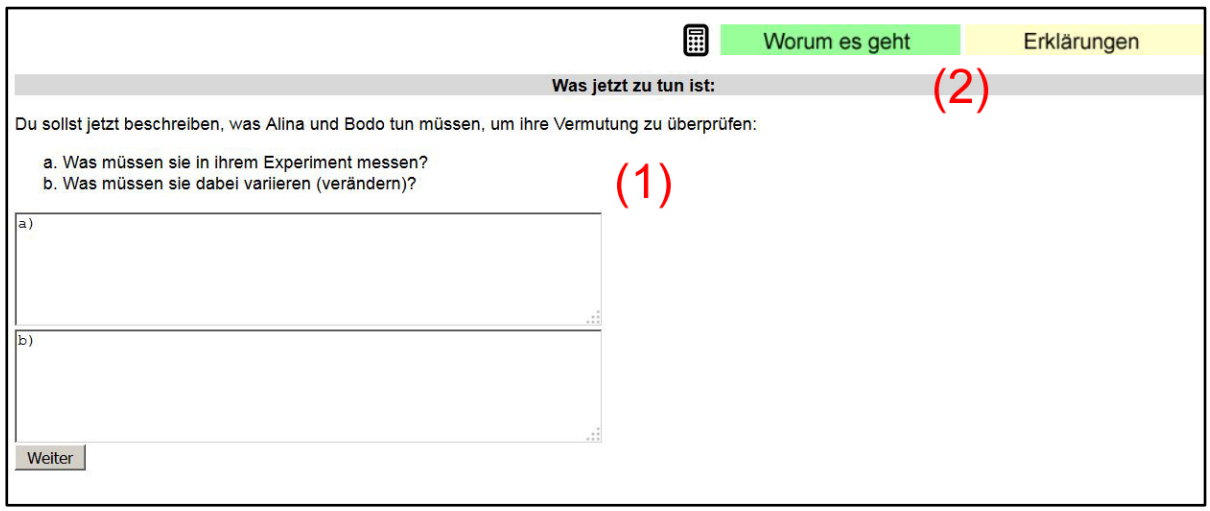
Seite 1: Aufgabenstamm

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



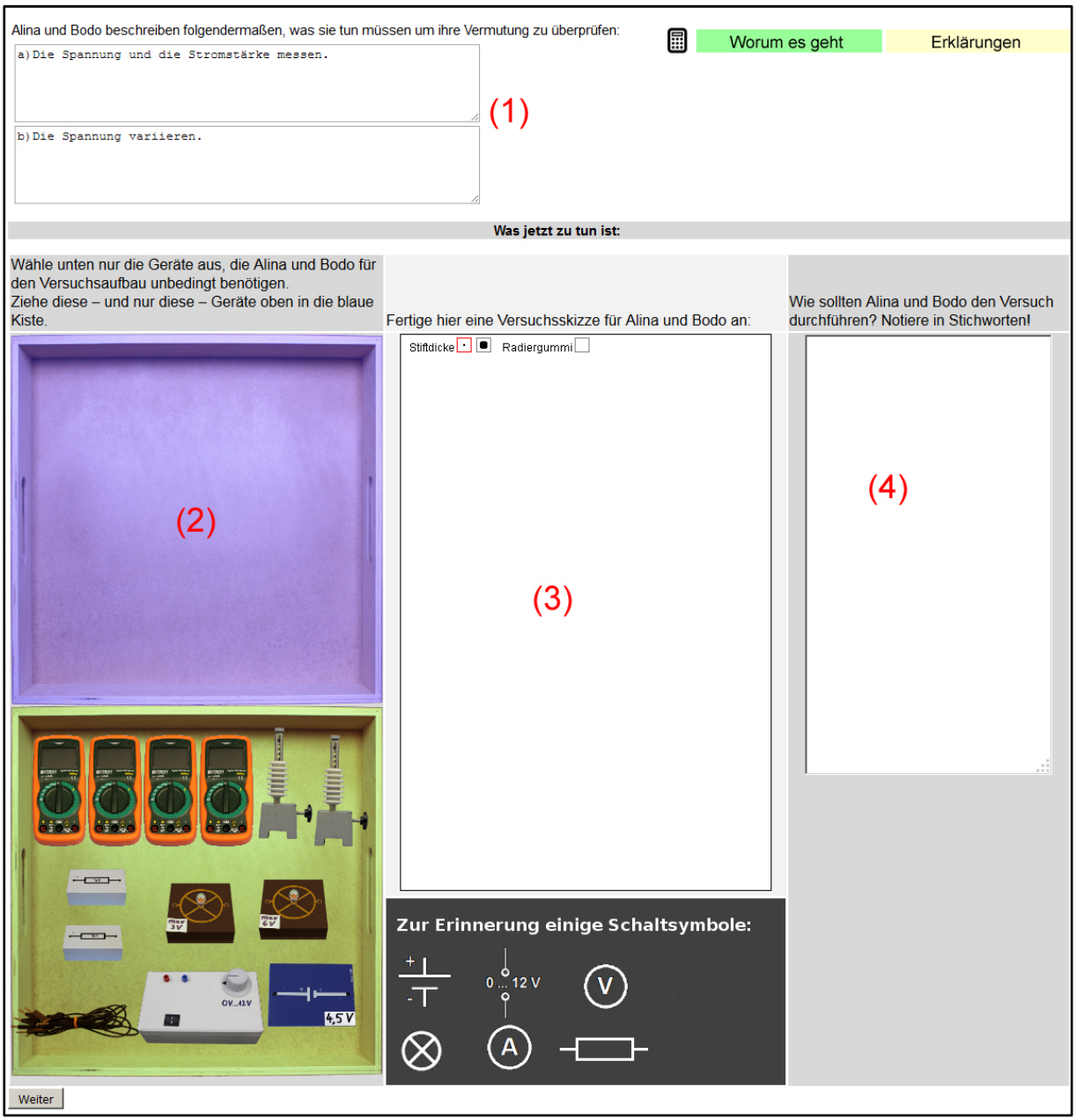
Seite 2: Teilaufgabe „Grundidee skizzieren“

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).



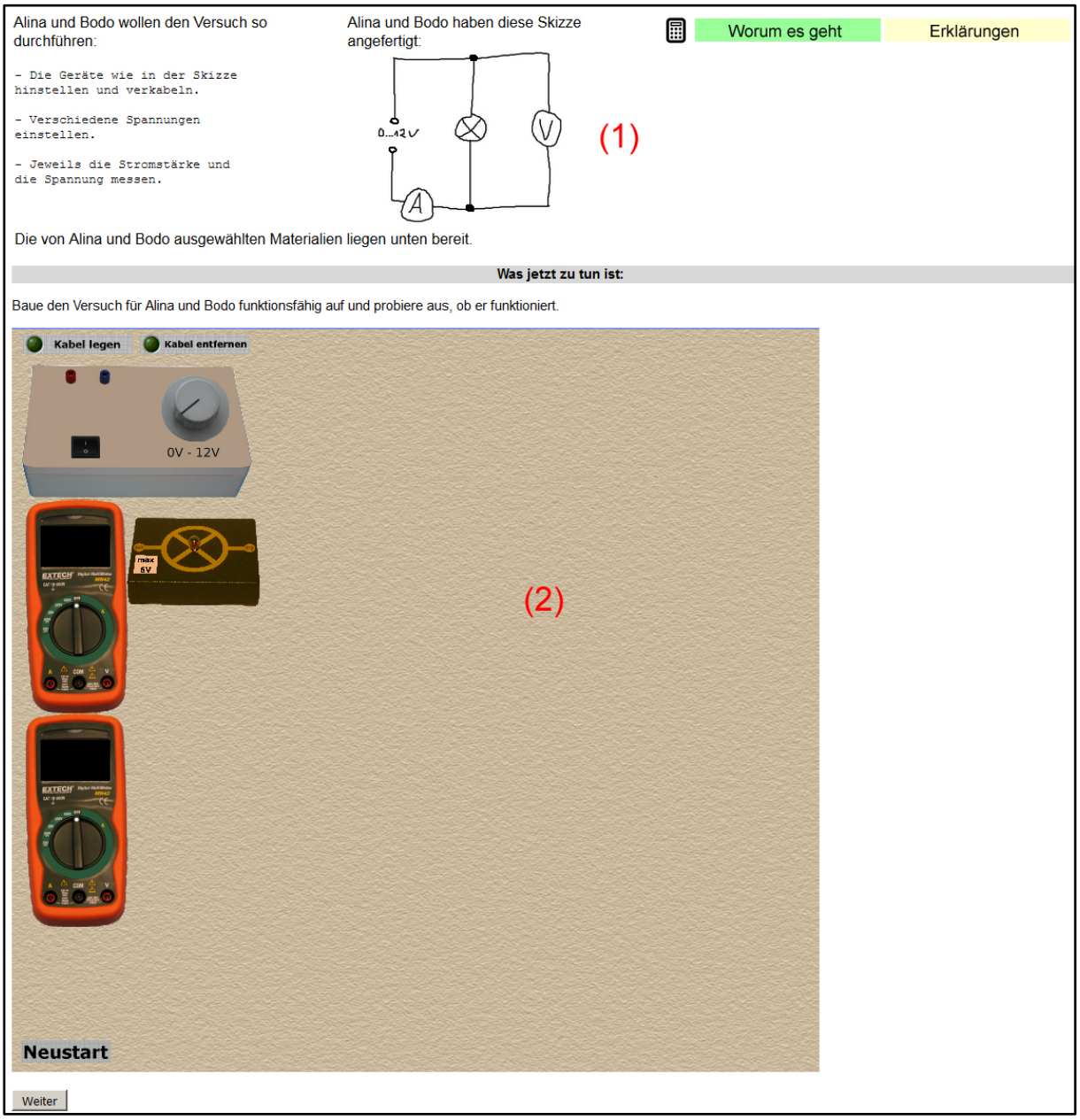
Seite 3: Teilaufgabe „Versuchsplan entwerfen“

Vorgegeben ist die Grundidee (1). Die Schülerinnen und Schüler sollen das benötigte Material auswählen (2), den Aufbau skizzieren (3) und die Vorgehensweise in Stichworten beschreiben (4). Bei der Geräteauswahl können sie die Geräte mit der Maus „greifen“, rotieren und in die blaue Kiste verschieben.



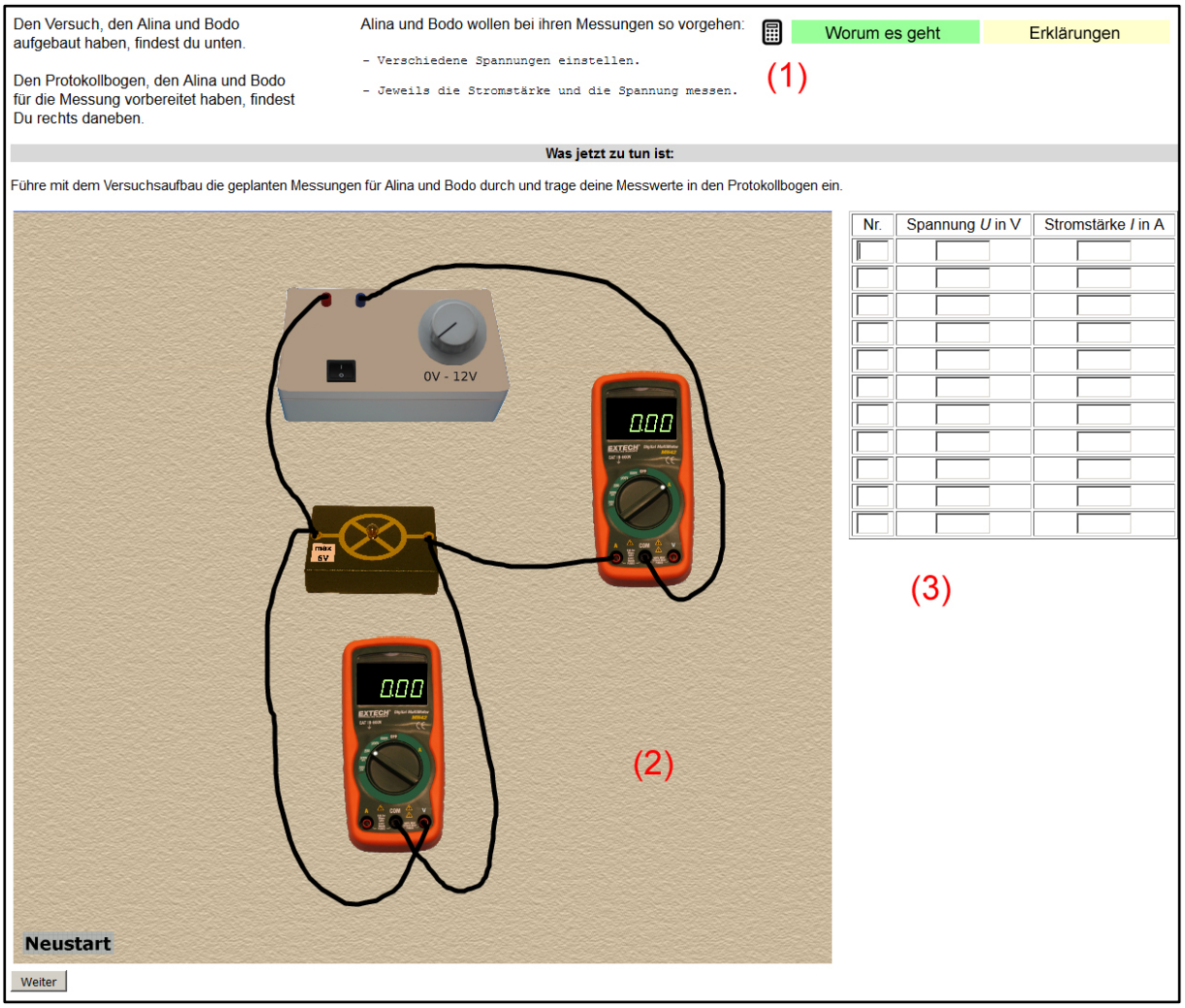
Seite 4: Teilaufgabe „Versuch aufbauen und testen“

Die Geräte, eine Skizze und die Beschreibung der Vorgehensweise sind vorgegeben (1). Die Schülerinnen und Schüler sollen nun den Versuch aufbauen und testen (2). In der Simulation können sie die Geräte beliebig anordnen und mit Kabeln verbinden. Sie können die Wählräder an den Multimetern bedienen, die Spannungsquelle einschalten und die Spannung einstellen. Wurde die Glühlampe oder ein Multimeter beim Test der Funktionsfähigkeit zerstört, kann die Ausgangskonfiguration mit intakter Glühlampe wiederhergestellt werden („Neustart“).



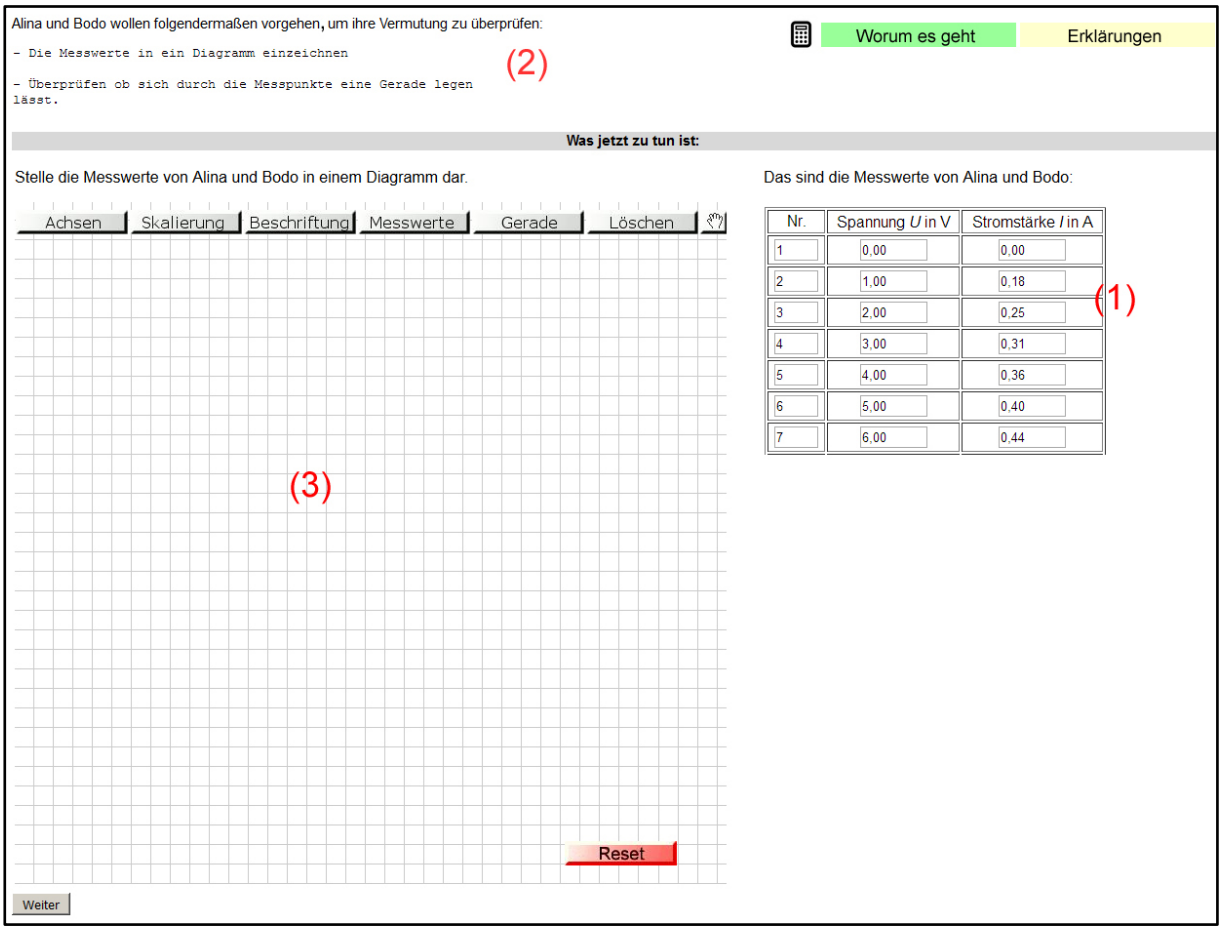
Seite 5: Teilaufgabe „Messung durchführen und dokumentieren“

Eine Beschreibung der Vorgehensweise (1), ein funktionsfähiger Aufbau (2) und eine Tabelle (3) sind für die Messung vorgegeben. Die Schülerinnen und Schüler sollen die Messung mit dem Aufbau durchführen und die Messwerte in der Tabelle dokumentieren. In der Simulation können sie nun nur noch die Spannung verändern und Messwerte ablesen. Wurde die Glühlampe durch eine zu hohe Spannung zerstört, kann das Experiment mit intakter Glühlampe neu gestartet werden („Neustart“).



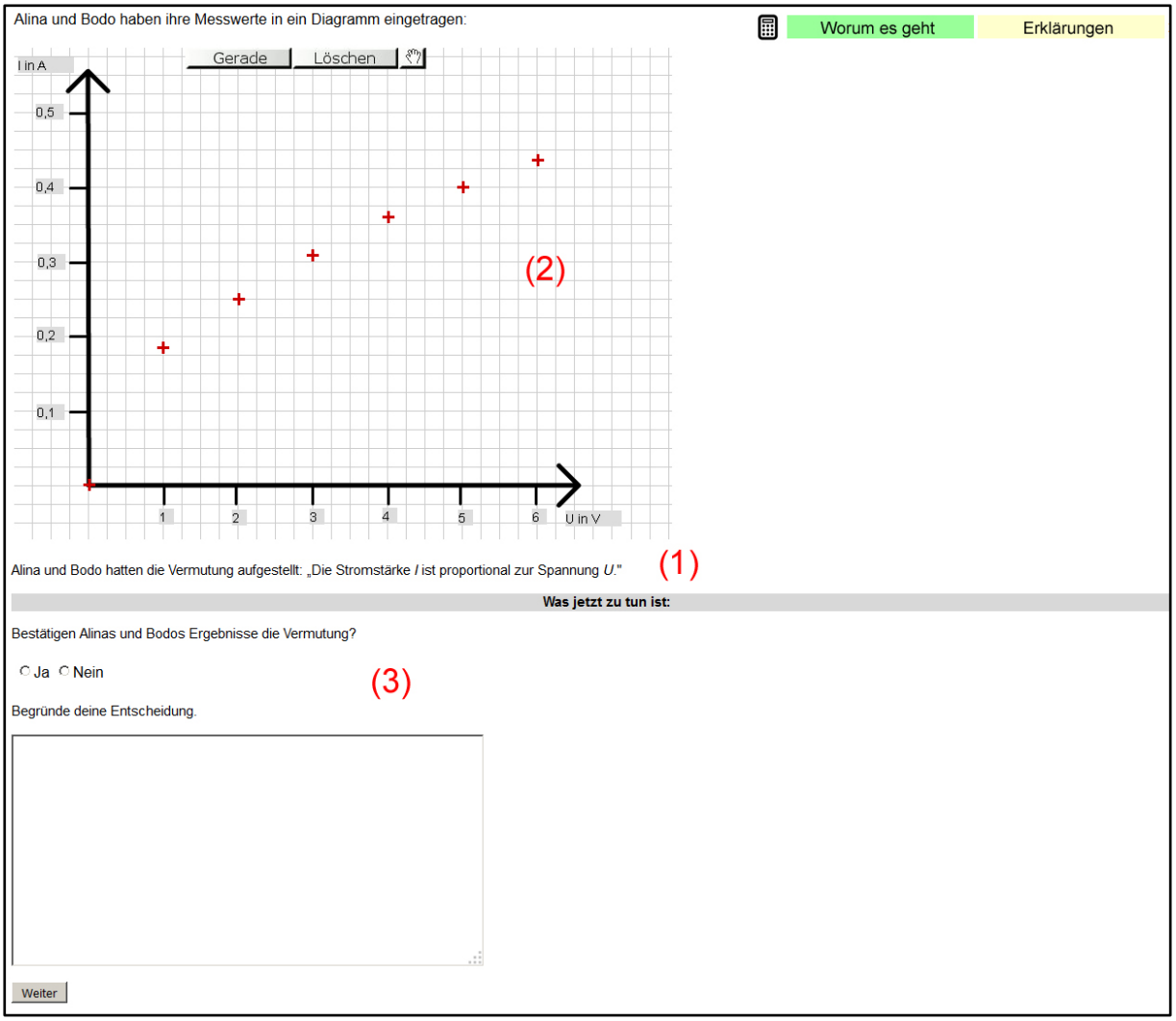
Seite 6: Teilaufgabe „Datenauswertung durchführen“

Die Schülerinnen und Schüler sollen die vorgegebenen Messwerte (1) auf vorgegebene Weise (2) auswerten, indem sie ein Diagramm erstellen (3). Mit dem Tool können Sie Achsen erstellen, skalieren und beschriften, Messwerte eintragen und Geraden einzeichnen. Einzelne Elemente oder ganze Bereiche können gelöscht oder verschoben werden.

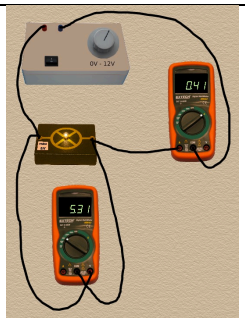
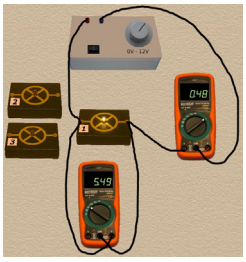
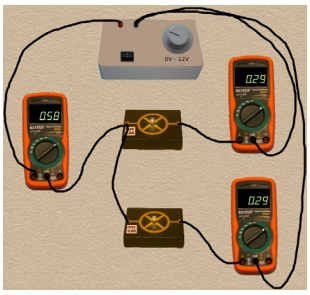
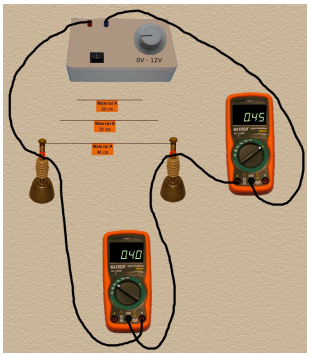


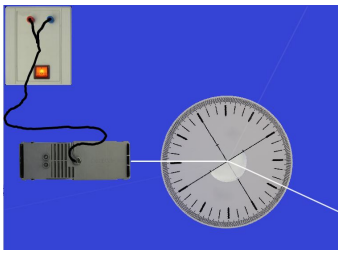
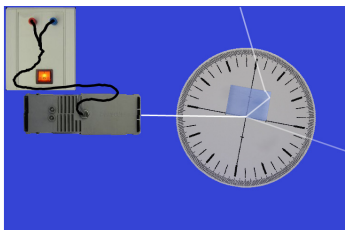
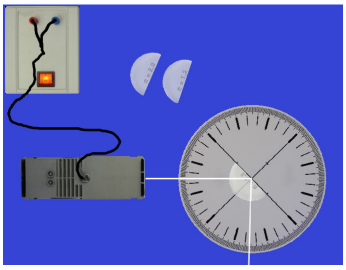
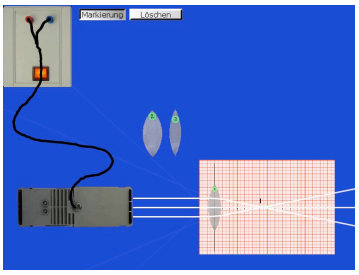
Seite 7: Teilaufgabe „Schlüsse ziehen“

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). Dazu können sie in dem vorgegebenen Diagramm eine Gerade einzeichnen, aber das Diagramm ansonsten nicht verändern.



Anhang C: Übersicht über die Units

Themenbereich, Thema	**Aufgabenstellung bzw. zu überprüfende Vermutung**	**Versuchsanordnung**
Elektrizitätslehre 1 Spannungs-Strom-Kennlinie einer Glühlampe	„Die Stromstärke in der Glühlampe ist proportional zur Spannung an der Glühlampe."	
Elektrizitätslehre 2 Leistung von Glühlampen	"Welche maximalen Leistungen haben die drei Glühlampen."	
Elektrizitätslehre 3 Parallelschaltung von Glühlampen	„Die Summe der Spannungen an den Glühlampen ist gleich der Gesamtspannung an der Spannungsquelle."	
Elektrizitätslehre 4 Reihenschaltung von Glühlampen	"Die Gesamtstromstärke ist gleich der Summe der Stromstärken in den Glühlampen."	
Elektrizitätslehre 5 (Trainingsunit) Widerstand eines Drahtstücks	„Der Widerstand eines Drahtes ist proportional zu seiner Länge."	

Optik 1 Brechung am Halbkreisblock	„Der Brechungswinkel ist proportional zum Einfallswinkel."	
Optik 2 Reflexion an einem Plexiglasblock	„Auch wenn ein Lichtstrahl auf eine nichtverspiegelte Oberfläche trifft, wird er reflektiert und es gilt das Reflexionsgesetz: Einfallswinkel gleich Ausfallswinkel."	
Optik 3 Totalreflexion	„Je größer der Brechungsindex des Materials, desto größer ist auch der Grenzwinkel der Totalreflexion."	
Optik 4 Brennweitenbestimmung einer Linse	„Je größer die Linsendicke ist, desto kleiner ist die Brennweite."	
Mechanik 1 Dichtebestimmung	Welche Dichten haben die drei Zylinder?	
Mechanik 2 Kraft-Dehnungs-Verhalten eines Gummibandes	"Die Ausdehnung des Gummibands ist proportional zur Masse der angehängten Gewichtsstücke."	

Mechanik 3 Auftriebskraft in Wasser	„Je größer - bei gleichem Volumen - der Durchmesser des Zylinders ist, desto größer ist die Auftriebskraft auf den Zylinder."	
Mechanik 4 Bewegung auf der Schiefen Ebene	"Die Fahrzeit des Autos ist proportional zum Kehrwert des Neigungswinkels der Rampe."