Entwicklung eines Fachwissenstests zur geometrischen Optik

Jeremias Weber*, Jan Winkelmann*, Roger Erb*, S. Franziska C. Wenzel[+], Mark Ullrich[+], Holger Horz[+]

*Institut für Didaktik der Physik, Max-von-Laue-Straße 1, 60438 Frankfurt am Main, [+]Institut für Psychologie, Theodor-W.-Adorno-Platz 6, 60629 Frankfurt am Main
jeremias.weber@physik.uni-frankfurt.de

Kurzfassung

Im BMBF-geförderten Projekt 'Kompetenzmessung und Kompetenzförderung in leistungsheterogenen Lerngruppen im experimentierbasierten Physikunterricht' (KoPhy) werden etwa 2000 Schülerinnen und Schüler in einer längsschnittlich angelegten Interventionsstudie auf den Kompetenzzuwachs in drei Experimentiersituationen untersucht.
Anhand einer im Frühjahr 2016 durchgeführten Pilotstudie mit 310 Teilnehmerinnen und Teilnehmern konnten für die Hauptstudie IRT-skalierte Testhefte zur geometrischen Optik für den Kompetenzbereich 'Fachwissen' zusammengestellt werden. Zusammen mit vorhandenen Items zur „Erfassung prozessbezogener naturwissenschaftlicher Grundbildung“ (Glug, 2009) und dem aktuellen Interesse der Schülerinnen und Schüler an Physik (Schulz, 2011) werden die Teilnehmerinnen und Teilnehmer der Hauptstudie umfassend befragt, um Auswirkungen der Intervention auf diese drei Aspekte zu erfassen. Die teilnehmenden Lehrkräfte erhalten zusätzlich einen Fragebogen zu ihrer Einstellung zum Fach und zur Wissenschaft Physik.
Im vorliegenden Beitrag werden die Rahmenbedingungen der Hauptstudie sowie ihr Aufbau kurz skizziert. Danach folgt ein Überblick über die Erkenntnisse der Pilotstudie und deren Einfluss auf die Hauptstudie. Die Vorstellung erster deskriptiv statistischer Ergebnisse aus der Hauptstudie schließt sich daran an. Abschließend soll ein Ausblick auf die weiteren Forschungsschritte und die bisher gewonnenen Erkenntnisse gegeben werden.

1. Motivation

Bereits seit vielen Jahren werden, beispielsweise bei Wagenschein [1] und Merzyn ([2], [3]) Experimente im Physikunterricht gefordert. Dementsprechend sind auch seit vielen Jahren Experimente fest in schulischen Curricula zum Fach Physik verankert. Folgerichtig nehmen Experimente laut Tesch [4] einen großen Anteil der Unterrichtszeit ein. Dabei dominieren Schülerexperimente, wie auch Duit und Wodzinski [5] anmerken. Allerdings beklagen sie, dass Schülerinnen und Schüler normalerweise „nur wenige Gelegenheiten haben, Experimente eigenständig zu planen, durchzuführen und auszuwerten“ ([5], S.1). Auch laut Hofstein & Lunetta [6] dominieren „Kochbuch“-Experimente („‘cook-book‘ lists of tasks“, S. 47, [6]). Von ihnen wird daher ein stärkerer Fokus auf fragengestütztes Experimentieren gefordert, das ihrer Meinung nach Lernende verschiedener Fähigkeiten, Lernstile oder kultureller Kontexte motiviert. Hofstein und Lunetta berichten weiterhin von widersprüchlichen Ergebnissen zur Frage, ob der Lernerfolg der Lernenden eher von Schüler- oder Demonstrationsexperimenten unterstützt wird. Winkelmann [7] schreibt in einem Überblick, dass es Hinweise auf Vorteile beider Experimentiersituationen gibt (S. 15).

2. Bisherige Vorarbeiten

Winkelmann [7] hat in Vorarbeiten verschiedene Experimentiersituationen formuliert und auf ihren Lernerfolg hin untersucht. Die Experimentiersituationen (Demonstrationsexperiment, Schülerexperiment „Kochbuch“, Schülerexperiment „Guided“) unterschieden sich dabei im Grad der Anleitung (während der Planung und Durchführung) sowie in der oder den handelnden bzw. ausführenden Person oder Personen. Alle Experimentiersituationen bestanden aus Experimenten zur Lichtbrechung, insbesondere der folgenden Phänomene:

a) Lichtbrechung an der Wasseroberfläche
b) Lichtbrechung an Glasoberflächen
c) Totalreflexion an Glasoberflächen
d) Lichtbündelung durch Sammellinsen
e) Bildentstehung an Sammellinsen und Abbildungsgesetz.

Die Auswertung der Experimente wurde in allen Experimentiersituationen im Plenum durchgeführt.

	Experimentiersituation		
	„Demo“	„Kochbuch“	„Guided“
Planung	Lehrkraft	Lehrkraft	Schülerinnen und Schüler
Durchführung	Lehrkraft	Schülerinnen und Schüler	Schülerinnen und Schüler
Auswertung	Plenum	Plenum	Plenum

Tab. 1: Unterschied der Experimentiersituationen

Dabei hat Winkelmann festgestellt, dass die Schülerexperimente, verglichen mit den Demonstrationsexperimenten, keinen signifikant positiven Einfluss auf

den Lernzuwachs über die Unterrichtsreihe hinweg haben.
Winkelmann fand dabei heraus, dass es eine kleine Wechselwirkung zwischen Experimentiersituation und Lehrkraft gab, weswegen er es für „lohnenswert [hält], zukünftig an diesem Punkt weitere Forschungsarbeiten zu leisten“ (S. 134, [7]).
Weiter wurde ein leichter Vorteil leistungsstarker und –schwacher Schülerinnen und Schüler in stark angeleiteten Experimentiersituationen, verglichen mit Schülerinnen und Schüler mittlerer Fähigkeit, festgestellt. Diese profitierten eher von der „Guided“-Experimentiersituation.

3. Studie zur Kompetenzmessung und Kompetenzförderung in leistungsheterogenen Lerngruppen im experimentierbasierten Physikunterricht (KoPhy-Studie)

3.1 Forschungsfragen und Studiendesign

Anhand der oben dargestellten Vorarbeiten konnten für die hier vorgestellte KoPhy-Studie folgende Forschungsfragen formuliert werden [8]:

> 1.1 Wie wirken sich die unterschiedlichen Experimentiersituationen im Physikunterricht auf die Entwicklung im Kompetenzbereich „Fachwissen“, im Kompetenzbereich „Erkenntnisgewinnung“ und auf das aktuelle Interesse der Schülerinnen und Schüler auf Physik aus?
> 1.2 Welche Unterschiede zeigen sich in heterogenen Leistungsgruppen aufgrund der unterschiedlichen Experimentiersituationen im Physikunterricht in Bezug auf die Entwicklung im Kompetenzbereich „Fachwissen“ und im Kompetenzbereich „Erkenntnisgewinnung“?
> 2. Welche Auswirkungen hat die Interaktion von Lehrercharakteristika und Experimentiersituation auf die Kompetenzentwicklung von Schülerinnen und Schülern im Fach Physik?

Die Studie ist als längsschnittliche Interventionsstudie angelegt. Um die Forschungsfragen zu beantworten werden die Unterrichtsreihen von Winkelmann [7] wiederverwendet bzw. leicht adaptiert, da diese sich bereits im praktischen Einsatz bewährt haben.
Dabei wird die Intervention bzw. Experimentiersituation zufällig teilnehmenden Klassen zugeteilt. Vor sowie nach der Intervention, also der jeweiligen Unterrichtsreihe, erhalten die Schülerinnen und Schüler einen Fragebogen mit Aufgaben und Fragen zu verschiedenen Bereichen. An drei weiteren Messzeitpunkten, im Abstand von jeweils vier Wochen, werden ebenfalls Fragebögen verteilt und Follow-Up-Tests durchgeführt. Um dabei die Testbelastung für die teilnehmenden Schulen möglichst gering zu halten, wird hierbei ein Planned-Missing-Design [9] genutzt. Die Teilnehmerinnen und Teilnehmer werden dabei nur an einem der drei nachfolgenden Follow-Up-Termine befragt und somit nur an drei von insgesamt fünf Messzeitpunkten. Der Termin zur Follow-Up-Befragung wurde den Klassen randomisiert zugewiesen.
Der zeitliche Ablauf der Studie wird in der folgenden Tabelle 2 noch einmal graphisch dargestellt:

2 Unterrichtsstunden	6 Unterrichtsstunden	2 Unterrichtsstunden	2 Unterrichtsstunden	2 Unterrichtsstunden	2 Unterrichtsstunden
	Intervention, Experiment				
Demoexperimentiergruppe: „Demo“					
Schülerexperimentiergruppe 1: „Kochbuch“					
Schülerexperimentiergruppe 2: „Guided“					
Pretest		Posttest	Follow-Up-Test	Follow-Up-Test	Follow-Up-Test
Sept. 2016	Oktober - Dezember 2016		Januar 2017	Februar 2017	März 2017

Tab. 2: Design der Studie

3.2 Messinstrumente

Zur Beantwortung der Forschungsfragen wurden verschiedene Messinstrumente genutzt, die an den jeweiligen Messzeitpunkten eingesetzt werden.
Im Pretest wird neben personenbezogenen Daten die kognitive Leistungsfähigkeit (KFT-V3, -N2) nach Heller & Perleth [10] als Kontrollvariable erhoben. Diese wird nur zu diesem Messzeitpunkt erhoben, da sie als stabiles Merkmal angenommen werden kann.
Ebenfalls werden während des Pretests die Lehrkräfte anhand eines Auszugs des Fragebogens von Lamprecht [9] zu ihren Überzeugungen zum Physikunterricht und zur Physikwissenschaft befragt. Die Erkenntnisse dieser Befragung sollen in Kombination mit den übrigen Messinstrumenten wertvolle Hinweise zur Beantwortung der zweiten Forschungsfrage liefern. Da diese Überzeugungen über den Untersuchungszeitraum hinweg stabil sind, wird dieses Messinstrument nur während des Pretests genutzt.
Das aktuelle Interesse der Schülerinnen und Schüler an Physik wird im Posttest anhand einer an den Physikunterricht angepassten Skala von Schulz [11] erhoben.
An allen Messzeitpunkten wird die Entwicklung in den Kompetenzbereichen „Erkenntnisgewinnung“ und „Fachwissen“ gemessen. Durch die Befragung zu fünf unterschiedlichen Zeitpunkten kann neben einer linearen auch eine nichtlineare Veränderung der Kompetenz der Schülerinnen und Schüler untersucht werden. Für den Kompetenzbereich „Erkenntnisgewinnung“ werden dabei Itemsets aus dem Test zur prozessbezogenen naturwissenschaftlichen Grundbildung von Glug [12] eingesetzt, die bereits nach der Item-Response-Theorie (IRT, [13]) kalibriert sind. Für die Messung der Entwicklung im Bereich „Fachwissen“ gibt es für das Themengebiet der „Geometrischen Optik“ keine nach den Methoden der Item-Response-Theorie kalibrierten Testinstrumente. Daher wurde im Vorfeld der Studie ein solcher Test entwickelt.

3.3 Konzeption eines Fachwissenstests

Mit dem Ziel, einen IRT-skalierten Fachwissenstest im Bereich der „Geometrischen Optik“ für Schülerinnen und Schüler zu entwickeln, welcher eine reliable Messung der Veränderung über mehrere Messzeitpunkte hinweg erlaubt, wurden die folgenden Schritte unternommen.

3.3.1 Reanalyse existierender Items

Zunächst wurden die von Winkelmann [7] eingesetzten Items hinsichtlich zentraler psychometrischer Aspekte und unter Nutzung von Modellen der Item-Response-Theorie analysiert (Schritt 1) und problematische Items überarbeitet, sofern möglich. Insgesamt hat Winkelmann 54 Items eingesetzt, bei einer Stichprobe von insgesamt 951 Schülerinnen und Schülern. Aus diesen konnten 32 Items, die einem höheren Schwierigkeitsspektrum zuzuordnen waren, selektiert werden. Zusätzlich wurden in einer Expertengruppe 30 weitere Items entwickelt, die dann mit den vorhandenen Items verknüpft wurden, um sie in einer Vorstudie zu kalibrieren.

3.3.2 Vorstudie

Überarbeitete und neu entwickelte Items wurden im Rahmen einer Pilotstudie IRT-skaliert und entsprechend ihrer psychometrischen Güte für den Fachwissenstest selektiert oder eliminiert (Schritt 2). Die 51 Items wurden dabei auf 13 Testhefte aufgeteilt, wobei ein balanciertes unvollständiges Testheftdesign [14] genutzt wurde. Die Testhefte enthielten nicht mehr als 15 Items und wurden von insgesamt 310 Personen bearbeitet. Dazu zählten Schülerinnen und Schüler und Lehramtsstudierende der Universitäten Köln und Frankfurt. Genauso wie im ersten Schritt wurden die Items nach der Durchführung der Vorstudie skaliert und selektiert.

3.3.3 Erstellung des Fachwissenstests

Zuletzt wurden aus dem so gewonnenen Itempool Testhefte zusammengestellt, die eine Erfassung des Fachwissens im Bereich „Geometrische Optik“ über verschiedene Messzeitpunkte hinweg, und damit die Entwicklung des Fachwissens über die Zeit, erlauben (Schritt 3). Der zur Testhefterstellung zur Verfügung stehende Itempool umfasste schließlich 60 Items. 48 dieser Items wurden zu verschiedenen Testheften für die unterschiedlichen Messzeitpunkte zusammengestellt. Dabei wiederholt sich jeweils nur ein Teil der Items über mehr als einen Messzeitpunkt, diese werden Ankeritems genannt. Die übrigen Items sind messzeitpunktspezifisch ausgewählt. Insbesondere wurde dabei die Schwierigkeit der einzelnen Items an die zu erwartende Personenfähigkeit angepasst: Im Pretest wurden also besonders einfache Items genutzt, im Posttest dafür schwerere Items.
Um Reihenfolgeeffekte soweit wie möglich zu minimieren, wurden für jeden Messzeitpunkt fünf Testhefte anhand eines balancierten unvollständigen Testheftdesigns [15] erstellt. Insgesamt sollte der so entwickelte Fachwissenstest eine Bearbeitungsdauer von ca. 20 Minuten aufweisen, aber trotzdem eine änderungssensitive und reliable Erfassung des Fachwissens erlauben.

4. Studiendurchführung und erste Ergebnisse

4.1 Bisherige Durchführung

Für die Hauptstudie wurden bis zum zweiten Halbjahr des Schuljahres 2016/17 insgesamt 63 Klassen rekrutiert. Die beteiligten Lehrkräfte meldeten zurück, dass die vorgegebenen Unterrichtsreihen gut durchführbar waren. Um verschiedene Stundenraster zu berücksichtigen, wurden sowohl die Testhefte als auch die vorgegebenen Stunden in Versionen für Einzel- und Doppelstunden bereitgestellt. Dadurch und durch die Vorbereitung der Arbeitsmaterialien durch das Forscherteam wurde die Akzeptanz der Studie bei den Lehrkräften deutlich erhöht. In den insgesamt acht Fällen, bei denen die Studie in teilnehmenden Klassen abgebrochen wurde, lagen dem Abbruch organisatorische Aspekte (u.a. Schulwechsel der Lehrkraft, Elternzeit) zugrunde. Bei einer zu erwartenden durchschnittlichen Klassengröße von 26 Schülerinnen und Schülern [16] kann im Moment von einer Stichprobengröße von 1430 Schülerinnen und Schülern ausgegangen werden. Bis zum Ende des Schuljahres 2016/17 soll eine Nachrekrutierung unter identischen Rahmenbedingungen durchgeführt werden, um die Stichprobe weiter zu vergrößern.

4.2 Erste deskriptive Ergebnisse

Von den zum aktuellen Zeitpunkt 50 zurückgesandten Klassensätzen an Pre- und Posttest-Fragebögen sind zehn Sätze teilweise digitalisiert und aufbereitet worden, mit insgesamt 167 teilnehmenden Schülerinnen und Schülern. Mit einer so kleinen Stichprobe können die in der Studie geplanten Mehrebenenanalysen noch nicht durchgeführt werden. Im Folgenden sind daher nur Ausblicke, basierend auf dieser Teilstichprobe, zu sehen. In der folgenden Tabelle 3 sind die bisher aufbereiteten Fragebogensätze im Vergleich zu den bisher zurückgesandten Sätzen und rekrutierten Klassen, jeweils bezogen auf die Experimentiersituation, zusammengefasst.

	Demoexp.	Kochbuch	Guided
Rekrutierung	20	22	21
Bisheriger Rücklauf	18	16	16
Aufbereitete Sätze	4	3	3

Tab. 3: Rücklaufquote und Stand der Datenaufbereitung (als Anzahl von Klassen)

Die hier näher betrachtete Teilstichprobe setzt sich aus deutlich mehr Mädchen (54%) als Jungen (29%) zusammen, wobei ein recht großer Teil der Befragten (17%) keine Angabe zu seinem Geschlecht machte.
In dieser Stichprobe waren die Schülerinnen und Schüler wie in folgender Tabelle dargestellt auf Klassenstufen und einzelne Klassen verteilt (Tab. 4). Es sind nur die Schülerinnen und Schüler berücksichtigt, die ihre Klassenstufe angegeben haben.

Klasse	Klassenstufe	Eingescannte Bögen
1	8	8
2	7	19
3	7	21
4	7	9
5	7	6
6	7	13
7	7	13
8	8	24
9	8	15
10	8	13

Tab. 4: Verteilung der Teilnehmerinnen und Teilnehmer

In der rekrutierten Gesamtstichprobe ist die Klassenstufe 7 stärker repräsentiert (in einem Verhältnis 3:1 zur Klassenstufe 8). Dagegen ist hier eine Verteilung von 3:2 zu finden.
Die Vornoten in den Fächern Deutsch, Mathematik und Physik korrelieren untereinander und waren im Mittel jeweils „gut“, dabei war das gesamte Notenspektrum vorhanden.

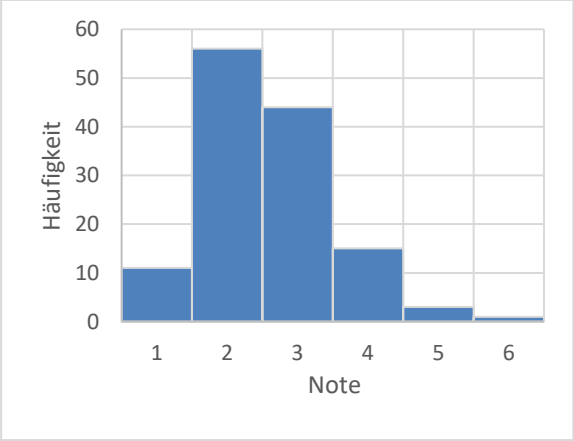

Abb. 1: Verteilung der Schulnoten in der Teilstichprobe

Weder anhand der Schulnoten noch der Anzahl der gelösten Items kann eine endgültige Aussage über die Heterogenität der verschiedenen Klassen getroffen werden. Anhand der Skalen zur kognitiven Leistungsfähigkeit wird in der Auswertung der Gesamtstichprobe diese Frage genauer untersucht werden.
In der folgenden Tabelle 6 ist ein Überblick über die mittlere Anzahl gelöster Items im Pretest pro Klasse zu finden.

Klasse	Mittlere Anzahl	Standardabweichung
1	7,1	2,5
2	7,4	2,7
3	6,2	2,3
4	7,2	2,2
5	7,5	2,1
6	6,7	1,9
7	6,9	2,0
8	6,8	1,9
9	5,4	1,7
10	4,9	2,2

Tab. 5: Anzahl richtig gelöster Items im Pretest

Im Mittel nahm die Anzahl der richtig gelösten Physikitems vom Pretest zum Posttest leicht ab. Um das korrekt interpretieren zu können, müssen die bereits in 3.3 beschriebenen IRT-basierten Itemanalysen realisiert werden, was erst bei Vorlage der Gesamtstichprobe möglich ist. Zur Vermeidung von Gedächtniseffekten haben die Testhefte verschiedene Items zu verschiedenen Messzeitpunkten, bis auf die identischen Ankeritems (vgl. 3.3.3). Bei der hier vorliegenden Teilstichprobe sind das insgesamt vier Items. Es lässt sich dabei feststellen, dass zum Pretest im Mittel eines dieser vier Items, beim Posttest 1,4 von vier Items gelöst werden kann. Aufgrund der geringen Itemzahl ist das aber nur als Hinweis auf eine positive Tendenz zu verstehen. Weitere Aussagen, insbesondere über Unterschiede zwischen den Treatments, können an dieser Stelle noch nicht getroffen werden.
Über die interessensbezogenen Fragen nach Schulz [11] kann bereits eine hohe Reliabilität der Skala festgehalten werden, das Cronbach-Alpha als Maß für die interne Konsistenz beträgt 0.85. Außerdem gibt es einen ersten Hinweis darauf, dass die Schülerinnen und Schüler ein höheres Interesse am Fach Physik im Nachgang der Unterrichtsreihe haben, wenn sie Erfahrungen mit schülerorientierten Experimentiersituation sammeln konnten. Wie in der folgenden Tabelle verdeutlicht, ist dies aber nur als Tendenz zu verstehen.

Experimentiersituation	Mittleres Interesse	Standardabweichung
Demo	2,7	0,8
Kochbuch	3,2	0,8
Guided	3,3	0,8

Tab. 6: Mittleres Interesse in den Experimentiersituationen

5. **Fazit und Ausblick**

Die Studie in der vorgestellten Form stieß auf hohe Akzeptanz bei den teilnehmenden Schulen. Insbesondere die empirische Erprobung eines Fachwissenstests wurde von den teilnehmenden Lehrkräften sehr

positiv aufgenommen, in verschiedenen Rückmeldungen wurde die Evaluation des Fachwissenstests als wichtiges Kriterium für die Teilnahme an der Studie angegeben.
Auch die flexiblen Durchführungsmöglichkeiten und die umfangreiche Unterstützung durch Arbeitsmaterialien wurden von den Lehrkräften positiv erwähnt.

Der hier vorgestellte Fachwissenstest zur geometrischen Optik wurde nach Modellen der Item-Response Theorie konzeptioniert. Der Einsatz in der KoPhy-Studie erlaubt es, ihn im Folgenden nochmals zu reskalieren. Der so entwickelte Test kann in anderen Studien im Bereich der Unterrichtsforschung als eigenständiger Fachwissenstest oder auch im Kontext von Large-Scale-Assessments (großangelegte Vergleichsstudien, wie beispielsweise PISA) genutzt werden.

Die ersten, anhand einer kleinen Teilstichprobe erzielten deskriptiven Ergebnisse in Bezug auf das Fachwissen können natürlich noch keine Untersuchung des Lernzuwachs ermöglichen. Um Gedächtniseffekte zu vermeiden, wurden bei der hier dargestellten Studie unterschiedliche Fragebögen mit einer an den Messzeitpunkt angepassten Schwierigkeit verwendet. Eine abschließende Aussage zur Entwicklung des Fachwissens kann daher erst erfolgen, wenn der vollständige Datensatz digitalisiert und aufbereitet wurde.
Im weiteren Verlauf der KoPhy-Studie werden die Daten aus den verschiedenen Tests dann mehrebenen-analytisch betrachtet, um die in Frage stehenden Auswirkungen der Experimentiersituationen und insbesondere auch die Wechselwirkungen zwischen Experimentiersituation und den Überzeugungen der Lehrkräfte zu untersuchen. So sollen Antworten auf die eingangs formulierten Forschungsfragen gefunden werden.

6. **Literatur**

[1] Wagenschein, M. (1976). „Rettet die Phänomene (Der Vorrang des Unmittelbaren)“. In: Der Mathematische und Naturwissenschaftliche Unterricht, 1977, S. 129–137.

[2] Merzyn, G., (2010). „Physik – ein schwieriges Fach“. In: Praxis der Naturwissenschaften, 5/59, 9-12.

[3] Merzyn, G., (2008). „Naturwissenschaften, Mathematik, Technik – immer unbeliebter?“, Baltmannsweiler.

[4] Tesch, M. (2005). Das Experiment im Physikunterricht: Didaktische Konzepte und Ergebnisse einer Videostudie. In H. Niedderer, H. Fischler & E. Sumfleth (Eds.). Studien zum Physik und Chemielernen (Vol.42). Berlin: Logos Verlag.

[5] Duit, R. & Wodzinski, C.T. (2010). Merkmale guten Physikunterrichts. In: Duit, R. (Hrsg.). Piko-Briefe. Der fachdidaktische Forschungsstand kurzgefasst. IPN Kiel. Abgerufen von: http://www.ipn.uni-kiel.de/de/das-ipn/abteilungen/didaktik-der-physik/piko/piko-briefe032010.pdf

[6] Hofstein, A., & Lunetta, V. N. (2004). The Laboratory in Science Education: Foundations for the Twenty-First Century. Science Education, 88, 28-54.

[7] Winkelmann, J. (2015). Auswirkungen auf den Fachwissenszuwachs und auf affektive Schülermerkmale durch Schüler- und Demonstrationsexperimente im Physikunterricht. In H. Niedderer, H. Fischler, E. Sumfleth (Hrsg.). Studien zum Physik- und Chemielernen. Band 179. Berlin: Logos Verlag.

[8] Weber, J., Winkelmann, J., Erb, R., Wenzel, S. F. C., Ullrich, M., & Horz, H. (2016). Entwicklung von Messinstrumenten zum Kompetenzzuwachs anhand von Modellen der IRT. In: PhyDid B – Beiträge zur DPG-Frühjahrstagung des Fachverbandes Didaktik der Physik in Hannover.

[9] Little, T. D., & Rhemtulla, M. (2013). Planned Missing Data Designs for Developmental Researchers. Child Development Perspectives, 7, 199–204. doi:10.1111/cdep.12043.

[10] Heller, K. A., & Perleth, C. (2000). Kognitiver Fähigkeitstest für 4. bis 12. Klassen, Revision (KFT 4-12+R). Göttingen: Beltz Test GmbH.

[11] Schulz, A. (2011). Experimentierspezifische Qualitätsmerkmale im Chemieunterricht: Eine Videostudie. In: H. Niedderer, H. Fischler, & E. Sumfleth (Hrsg.). Studien zum Physik und Chemielernen (Vol. 113). Berlin: Logos Verlag.

[12] Glug, I. (2009). Entwicklung und Validierung eines Multiple-Choice-Tests zur Erfassung prozessbezogener naturwissenschaftlicher Grundbildung. Kiel: IPN.

[13] van der Linden, W. J. & Hambleton, R. K. (Eds.). (2013). Handbook of modern item response theory. Springer Science & Business Media.

[14] Frey, A., Hartig, J. & Rupp, A. (2009). Booklet designs in large-scale assessments of student achievement: Theory and practice. *Educational Measurement: Issues and Practice, 28*, 39-53.

[15] Osterlind, S. J. & Everson, H. T. (2009). Differential item functioning (Vol. 161). Sage Publications.

[16] Statistisches Bundesamt (2016). *Schulen auf einen Blick*. Wiesbaden: Statistisches Bundesamt. Abgerufen von: https://www.destatis.de/DE/Publikationen/Thematisch/BildungForschungKultur/Schulen/BroschuereSchulen-Blick0110018169004.pdf?__blob=publicationFile