Entwicklung von Messinstrumenten zum Kompetenzzuwachs anhand von Modellen der IRT

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, winkelmann@physik.uni-frankfurt.de, roger.erb@physik.uni-frankfurt.de, wenzel@psych.uni-frankfurt.de, m.ullrich@psych.uni-frankfurt.de, horz@psych.uni-frankfurt.de

Kurzfassung

Im BMBF-geförderten Projekt „Kompetenzmessung und Kompetenzförderung in leistungsheterogenen Lerngruppen im experimentierbasierten Physikunterricht (KoPhy)“ sollen ca. 2000 Schülerinnen und Schüler in einer längsschnittlich angelegten Interventionsstudie hinsichtlich ihres Kompetenzzuwachses durch drei Experimentiersituationen untersucht werden.
In diesem Beitrag sollen zuerst die Motivation für die vorgestellte Studie sowie die bisherigen Erkenntnisse dargestellt werden. Dem schließt sich eine Darstellung der Hauptstudie an, in welcher sowohl die grundlegende Fragestellung als auch die Messinstrumente vorgestellt werden. Die bisher abgeschlossenen Vorstudien zur Entwicklung eines IRT-skalierten Fachwissenstest als Messinstrument zum Einsatz in der Hauptstudie stellen den Abschluss dieses Beitrags dar.

1. Motivation

1.1 Experimente im Physikunterricht

Der Physikunterricht in der Schule wird als besonders unbeliebt und schwierig empfunden, was laut Merzyn ([1], [2]) an der starken Abstrahierung des Physikunterrichtes liegt. Bereits Wagenschein [3] wünscht sich einen weniger abstrakten und stärker experimentell ausgerichteten Physikunterricht. Folgerichtig schreiben Duit und Wodzinski [4] in einem Bericht über die Merkmale guten Physikunterrichts, dass Experimente einen großen Anteil der Unterrichtszeit einnehmen. Dabei legen Lehrkräfte etwas mehr Wert auf Schülerexperimente als Demonstrationsexperimente (11% der Unterrichtszeit im Gegensatz zu 7%, [4], S. 1). Duit und Wodzinski beklagen jedoch, dass „die Schülerinnen und Schüler in der Regel nur wenige Gelegenheiten haben, Experimente eigenständig zu planen, durchzuführen und auszuwerten“ ([4], S. 1). Auch Hofstein und Lunetta [6] berichten hauptsächlich von „Kochbuch“-Anleitungen („‘cook-book‘ lists of tasks“, S. 47, [6]).

Entsprechend fordern Hofstein und Lunetta bereits eine stärkere Konzentration auf fragengestützten Experimentierunterricht, der die Lernenden dann auch trotz verschiedener Fähigkeiten, Lernstile und kultureller Kontexte motiviert.

Ob jedoch eher Schülerexperimente oder Demonstrationsexperimente den Lernerfolg der Schülerinnen und Schüler unterstützen, ist unklar. Winkelmann ([5] S. 15) schreibt in einem Überblick, dass es Hinweise für Vorteile beider Experimentiersituationen gibt.

1.2 Experimentiersituationen

Um verschiedene Experimentiersituationen zu vergleichen, hat Winkelmann [5] verschiedene „Treatments“ zum Fachgebiet der geometrischen Optik entwickelt. Jedes der drei Treatments besteht aus vergleichbaren Experimenten zur Lichtbrechung und unterscheidet sich nur in der Darbietung. Die Experimente thematisieren folgende 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 Treatments unterscheiden sich darin, dass jedes Experiment in drei Varianten vorliegt:

Ein „Demonstrationsexperiment“, welches von dem Lehrenden durchgeführt wird, ein „Kochbuchexperiment“, welches nach genauen Instruktionen von den Schülerinnen und Schülern durchgeführt wird und ein „angeleitetes Experiment“, welches den Schülerinnen und Schülern grobe Vorgaben für das Experiment macht und eine anleitende Frage stellt.

Abb. 1: Versuchsanordnung „Peilversuch“ ([5], S. 44)

Exemplarisch soll das an dem „Peilversuch“, einem Versuch zur Lichtbrechung an einer Wasseroberfläche kurz skizziert werden (Abb. 1):

a) Im Demonstrationsexperiment baut die Lehrkraft die abgebildete Versuchsanordnung auf und führt sie durch, mit eventueller Beteiligung der Schülerinnen und Schüler.
b) Im Kochbuchexperiment bauen die Schülerinnen und Schüler nach genauer Anleitung den Versuch auf und erhalten eine genaue Beobachtungsanweisung.
c) Im „angeleiteten Experiment“ wird ein genereller Arbeitsauftrag formuliert und eine Materialliste übergeben. Die Schülerinnen und Schüler werden aufgefordert, den Versuch selbstständig zu planen und durchzuführen, ohne dass ihnen erklärt wird, wie sie genau vorgehen sollen.

In Tabelle 1 werden die Treatments anhand der drei Phasen Planung, Durchführung und Auswertung unterschieden. Da die Auswertung bei allen Phasen im Plenum geschieht, werden Unterschiede nur in den Phasen Planung und Durchführung erzeugt.

	Experimentiersituation		
	Demoexperiment	Kochbuchexperiment	Angeleitetes Experiment
Planung	Lehrkraft	Lehrkraft	Schülerinnen und Schüler
Durchführung	Lehrkraft	Schülerinnen und Schüler	Schülerinnen und Schüler
Auswertung	Gemeinsam	gemeinsam	gemeinsam

Tab. 1: Darstellung der Treatmentunterschiede

Für die Studie des aktuellen Projektes „KoPhy“ werden diese Treatments übernommen.

1.3 Bisherige Erkenntnisse

Winkelmann [5] hat zu der Wirkung der Treatments auf den Fachwissenszuwachs festgestellt, dass die Schülerexperimente keinen signifikant positiven Einfluss haben, verglichen mit Demonstrationsexperimenten. Dabei hat er den Lernzuwachs zwischen zwei Messzeitpunkten (Pre-Posttest) untersucht. Unter Hinzunahme eines dritten Testzeitpunktes (ein Kurztest direkt nach Durchführung der Experimente) wurde jedoch ein signifikanter Effekt der Nachbereitung der Schülerexperimente (Auswertungsphase, vgl. Tab. 1) sichtbar.

Eine wichtige Erkenntnis ist weiterhin, dass der Effekt der Lehrkraft in Wechselwirkung mit der genutzten Experimentiersituation hochsignifikant ist, wenn auch klein. Daher halten wir es für „lohnenswert, zukünftig an diesem Punkt weitere Forschungsarbeit zu leisten“ (S. 134, [5]).

Bezogen auf die Heterogenität der Schülerinnen und Schüler stellt Winkelmann fest, dass leistungsschwache und –starke Schülerinnen und Schüler einen leichten Vorteil in stark angeleiteten Experimentiersituationen haben, Schülerinnen und Schüler mittlerer Fähigkeit jedoch eher vom „angeleiteten Experiment“ profitieren. Er konnte aber anhand der von ihm untersuchten Gruppe diese Tendenzen nicht statistisch belegen.

In der Studie wird nicht nur der Kompetenzbereich Fachwissen betrachtet, es wird unter anderem gefragt, ob „sich das (selbstständige) Experimentieren auf andere Facetten des Unterrichts gravierender ausübt“ (S. 134, [5]). Erkenntnisse, ob ein Zuwachs in anderen Kompetenzbereichen (bspw. Erkenntnisgewinnung) stattfindet, liegen also bisher nicht vor.

2. Design der KoPhy-Studie und Implementation der Messinstrumente

2.1 Forschungsfragen und Studiendesign

Aufbauend auf den oben dargestellten Erkenntnissen und aufgeworfenen Fragen wurden folgende Fragestellungen für die im Anschluss beschriebene Studie zur Kompetenzmessung und Kompetenzförderung, abgekürzt „KoPhy“ formuliert:

> 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 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?

Bereits in der in 1.3 dargestellten Studie wurde die Fragestellung 1.1 in Bezug auf den Kompetenzbereich Fachwissen formuliert. Im Rahmen der BMBF-geförderten „KoPhy“-Studie soll dies ausführlicher untersucht werden. Vor dem Hintergrund der aktuellen Bedeutung der Heterogenität im Unterricht (starke Prägung der Entwicklung eines Menschen durch verschiedene Hintergrundvariablen, vgl. [7], [8]), ist es auch von Interesse, inwiefern die Experimentiersituationen einen Einfluss auf unterschiedliche heterogene Lerngruppen haben. Entsprechende Hinweise wurden weiter oben bereits besprochen, daher soll diese Fragestellung jetzt weitergehend untersucht werden.

Winkelmann [5] hat tieferliegende Einflüsse durch die Wahl der Lehrkraft auf die Wirkung der Experimentiersituation beobachtet (s.o.). Daher stellt die Untersuchung der Wechselwirkung zwischen den Überzeugungen der Lehrperson zu Physik und Physik im Unterricht und der jeweiligen Experimentiersituation die zweite wichtige Fragestellung der „KoPhy“-Studie dar.

Die Hauptstudie ist als eine längsschnittliche Interventionsstudie angelegt. Die Schülerinnen und Schüler erhalten vor der Intervention als Pretest einen Fragebogen, dann folgt die Intervention, in diesem Fall die Durchführung einer Unterrichtsreihe mit den oben vorgestellten Unterschieden in der Experimentiersituation. Während der Intervention soll der Schulunterricht punktuell mitgeschnitten werden, um die Umsetzung der verschiedenen Experimentiersituationen in ihrer intendierten Spezifität zu dokumentieren. Nach der Intervention wird ein Posttest durchgeführt. Im Abstand von bis zu drei Monaten werden dann drei Follow-Up-Tests durchgeführt. Dabei wird hier ein Planned-Missing-Design genutzt, um die Menge an Follow-Up-Tests pro Schule gering zu halten: Jeder Follow-Up-Testzeitpunkt findet nur an einem zufällig (und proportional zum Anteil der Experimentiersituationen) ausgewählten Drittel aller teilnehmenden Schülerinnen und Schüler statt. Während der Studie werden die Schülerinnen und Schüler im Klassenverband belassen, es handelt sich damit um eine quasiexperimentelle Studie.

In Tabelle 2 wird dieses Design noch einmal graphisch vorgestellt.

2 Unterrichtsstunden	6 Unterrichtsstunden	2 Unterrichtsstunden	2 Unterrichtsstunden	2 Unterrichtsstunden	2 Unterrichtsstunden
	Intervention, Experiment				
Demoexperimentiergruppe: „Demo“					
Schülerexperimentiergruppe 1: „Kochbuch“					
Schülerexperimentiergruppe 2: „Guided“					
Pretest	Videographie	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 Hauptstudie

Da die Überzeugungen der Lehrkräfte als stabil angenommen werden, werden diese nur einmal, während des Pretests, erhoben.

2.2 Messinstrumente

Um die oben formulierten Forschungsfragen zu beantworten, sollen an den Messzeitpunkten verschiedene Messinstrumente benutzt werden.

Nur im Pretest wird neben den personenbezogenen Daten als Kontrollvariable die kognitive Leistungsfähigkeit (KFT) nach Heller & Perleth [9] erhoben. Diese ist auch über mehrere Testzeitpunkte hinweg stabil und muss daher nicht erneut gemessen werden.

Nur im Post-Test und im Follow-Up-Test wird das aktuelle Interesse der Schülerinnen und Schüler erhoben. Dabei werden die Testfragen von Schulz [10] übernommen, wie dies bereits in der oben dargestellten Vorstudie geschah.

Die Entwicklung in den Kompetenzbereichen „Erkenntnisgewinnung“ und „Fachwissen“ wird an allen fünf Testzeitpunkten gemessen. Anhand dieser verschiedenen Messzeitpunkte kann neben einer linearen auch eine nichtlineare Veränderung der Kompetenz der Schülerinnen und Schüler untersuchtwerden. Für die Messung der Entwicklung im Bereich „Erkenntnisgewinnung“ soll der Test zur prozessbezogenen naturwissenschaftlichen Grundbildung von Glug [11] eingesetzt werden. Für die Messung der Entwicklung im Bereich „Fachwissen“ gibt es für das Themengebiet der „Geometrischen Optik“ keine nach den Methoden der Item-Response-Theorie modellierten Tests. Der Vorteil solcher Tests ist unter anderem, dass nicht identische Tests miteinander verknüpft werden können und so Erinnerungseffekte vermieden

werden. Aus diesem Grund wird dieser Test im Verlauf der Studie selbst entwickelt, was im Folgenden dargestellt werden soll.

3. **Konzeption eines Fachwissenstests**

3.1 Analyse existierender Items

In der Studie von Winkelmann [5] wurden zwei Itempools mit 30 bzw. 24 Items entwickelt. Diese Items wurden in einer Befragung mit insgesamt 1032 Schülerinnen und Schülern eingesetzt. Anhand der Ergebnisse seiner Studie konnten diese Items nun anhand des Rasch-Modells skaliert werden, um damit die Itemschwierigkeit sowie die Personenfähigkeit der Zielpopulation zu schätzen.

Im nächsten Schritt wurden dann Items nach ihrer psychometrischen Güte ausgewählt. Dabei wurden als Kriterien sowohl die Passung der einzelnen Items mit dem Rasch-Modell (Itemfit), die Trennschärfe (als Korrelation des Itemscores mit dem gesamten Testscore) als auch Effekte des Geschlechts, der Mathematik- und der Deutschnote auf die Lösungswahrscheinlichkeit eines Items (differentielles Itemfunktionieren, DIF, [12]) herangezogen. Items, die durch die Analysen als problematisch identifiziert wurden (bspw. eine niedrige Trennschärfe oder Items sind für Mädchen leichter zu lösen als für Jungen), wurden von einer Expertengruppe untersucht und entweder verbessert oder eliminiert.

Insgesamt wurden aus den Itempools 33 Items selektiert, von denen 18 besonders gut funktionierende als Ankeritems für die Vorstudie ausgewählt wurden. Drei Items wurden überarbeitet und auch in die Vorstudie eingebracht. Zehn Items mussten wegen nicht passendem Inhaltsbereich (Fragen zur Reflexion) eliminiert werden. Da durch die ausgewählten 33 Items jedoch vor allem im hohen Fähigkeitsbereich gut zwischen Personen unterschieden werden kann, im niedrigeren Leistungsspektrum dies aber nicht so gut gelingt, und für den Einsatz über mehrere Messzeitpunkte in der Hauptstudie ein größerer Itempool benötigt wird, wurden weitere Items entwickelt. Insgesamt wurden von einer Expertengruppe weitere 30 Items erstellt. Dabei wurde sich in Form und Inhalt an den selektierten Items orientiert, mit einem starken Fokus auf relativ einfache Aufgaben. Zur detaillierten und fortlaufenden Dokumentation der Items, auch über Bearbeitungsschleifen hinweg, wurden Stammdatenblätter genutzt, die als roter Faden bei der Entwicklung neuer Items hilfreich waren.

Die so selektierten und neu geschaffenen insgesamt 51 Items wurden dann in einer Vorstudie (erneut) erprobt, um den oben erwähnten Fachwissenstest zu konstruieren.

3.2 Durchführung der Vorstudie

Um im Zuge der Vorstudie alle 51 Items erproben zu können, die einzelne Versuchsperson aber nicht mit mehr als 15 Minuten für die Testbearbeitung zu beanspruchen, wurde ein balanciertes unvollständiges Testheftdesign (z. B. [13]) verwendet. Dies bedeutet, dass einzelne Personen jeweils nur einen Teil der Items zur Bearbeitung vorgelegt bekamen. Die Items wurden auf insgesamt 13 Testhefte mit jeweils ca. 15 Items verteilt. Dabei enthielt jedes Testheft ungefähr gleich viele alte (selektierte) wie neue (von der Expertengruppe formulierte) Items. So sollte gewährleistet sein, dass die Personenfähigkeit der Probanden anhand der alten Items geschätzt werden kann und anhand dieser Werte wieder die Eignung der neuen Items abgeschätzt werden kann.

An der Vorstudie nahmen 301 Versuchspersonen teil, darunter sowohl Schülerinnen und Schüler als auch Lehramtsstudierende der Universitäten Frankfurt und Köln. Dadurch konnte eine breite Verteilung der Personenfähigkeit gewährleistet werden.

Die neuen Items wurden auf Basis der bereits geschätzten Schwierigkeiten der zuvor selektierten Items skaliert und nach den gleichen Kriterien, wie oben beschrieben, analysiert.

3.3 Ergebnisse der Vorstudie

Nach der Analyse der Vorstudie wurden insgesamt 41 Items selektiert. Die meisten Items sind weiterhin als schwer anzusehen (mit einer mittleren Schwierigkeit von 1,96). Es gibt mehr leichte Items als in den bisherigen Itempools sowie eine deutlich kontinuierlichere Verteilung der Itemschwierigkeiten (s. Abb. 2).

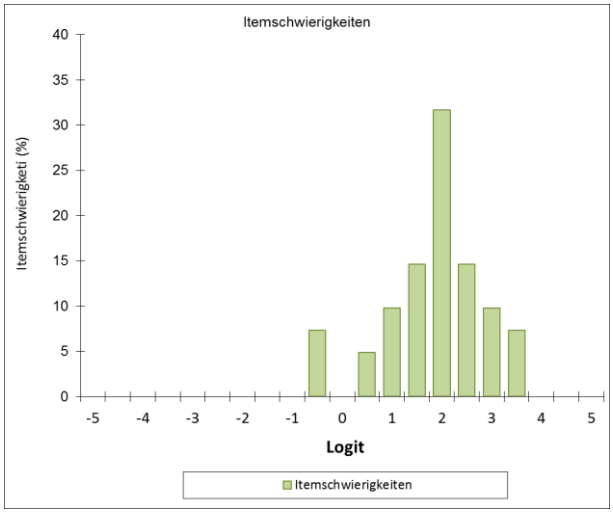

Abb. 2: Selektierte Items aus der Vorstudie

Aus den selektierten Items werden drei Testhefte erstellt, die als Messinstrument in der Hauptstudie über mehrere Messzeitpunkte genutzt werden sollen.

3.4 Fazit und weitere Betrachtungen

Für die „KoPhy"-Studie wurde ein gutes Messinstrument für einen Lernzuwachs im Kompetenzbereich Fachwissen benötigt. Daher wurden die Items von Winkelmann [5] neu analysiert und mit neugeschaffenen Items erweitert. Dies kann in gleicher Art auch mit ähnlichen Vorarbeiten geschehen. Insbesondere hat sich im vorliegenden Fall die Stärke dieser Analyseverfahren gezeigt: Auch bei Items, die geschulten Beobachtern zunächst nicht auffällig erschienen, konnten durch die psychometrische Analyse problematische Differenzierungen (DIF) nachgewiesen werden. Über anschließende gezielte Inhaltsanalysen ließen Verbesserungsmöglichkeiten solcher Items identifizieren. Auf diese Weise konnte der Test im Hinblick auch seine Fairness für unterschiedliche Personengruppen insgesamt verbessert werden. Es zeigt sich außerdem, dass auch mit Hilfe der klassischen Testtheorie erhobene Datenbestände für raschskalierte Auswertungen gute Grundlagen bilden.

Die Entwicklung neuer Items geschah in der beschriebenen Vorstudie anhand von Stammblättern und der Analyse der bestehenden Items. Die Güte der so neuentwickelten Items ist vergleichsweise hoch, 80% aller Items der Vorstudie konnten für die „KoPhy"-Studie selektiert werden.

Das so erzeugte Messinstrument wird anhand der Ergebnisse der Hauptstudie nochmals skaliert und validiert und kann dann als eigenständiger Fachwissenstest auch für andere Anwendungen genutzt werden.

4. Literatur

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

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

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

[4] 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

[5] 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.

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

[7] Klieme, E., Artelt, C., Hartig, J., Jude, N., Köller, O., Prenzel, M., Schneider, W., & Stanat, P. (2010). PISA 2009. Bilanz nach einem Jahrzehnt. Waxmann: Münster.

[8] Prenzel, M., Sälzer, C., Klieme, E., & Köller, O. (2013). PISA 2012. Fortschritte und Herausforderungen in Deutschland. Münster: Waxmann.

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

[10] 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.

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

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

[13] 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.