Philipp Straube[1]
Jurik Stiller[2]
Rüdiger Tiemann[2]
Volkhard Nordmeier[1]

[1]Freie Universität Berlin
[2]Humboldt-Universität zu Berlin

Ko-WADiS | Aspekte der Itemkonstruktion

Einleitung

Das der Itemkonstruktion im Rahmen des Projekts *Ko-WADiS* zugrunde liegende Kompetenzstrukturmodell (Straube et al. in diesem Band) wurde anhand bestehender Modelle adaptiert (Mayer, 2007; Upmeier zu Belzen & Krüger, 2010), die deduktiv aus der Theorie abgeleitet wurden. Bei der deduktiven Ableitung werden, ausgehend von theoretischen Überlegungen, die konkreten Handlungen beschrieben und ggf. durch Expertenurteile abgesichert (Hartig & Jude, 2007). Auf Grundlage des Kompetenzmodells wurde im Rahmen des Projektes Ko-WADiS ein Testinstrument zur Kompetenzerfassung entwickelt.

Grundsätzlich stehen dabei mehrere Testverfahren zur Verfügung. Dazu zählen unter anderem Selbsteinschätzungen, direkte Beobachtungen oder Paper-and-Pencil-Tests (dazu überblicksartig Hartig & Jude, 2007). Für den Bereich der Erkenntnisgewinnung beschreiben Schreiber, Theyßen, und Schecker (2009) darüber hinausgehend Instrumente, die die aufwändige Beobachtung von Experimentiersituationen durch Simulationsbaukästen am Computer ersetzen. Die Auswahl der Testinstrumente muss jedoch auch immer im Kontext der Stichprobe und der damit verbundenen Kosten gesehen werden (Stecher & Klein, 1997). Die Stichproben im Projekt Ko-WADiS umfassen zunächst 150 bis 400 Studierende der beteiligten Fächer an beiden beteiligten Universitäten pro Erhebung. Zusätzlich kommen punktuell die Studierenden der entsprechenden Mono-Bachelor- bzw. -Master-Studiengänge, die als Kontrollgruppe fungieren, hinzu. Der Weiteren werden – ebenfalls punktuell – Studierende der Universitäten in Wien und Innsbruck getestet, um den Einfluss eines nicht-kompetenzorientierten Studiums untersuchen zu können. Daraus ergibt sich eine Gesamtstichprobengröße von bis zu 1000 Studierenden. Ein Einsatz von direkten Beobachtungen ist daher nicht zu realisieren. Eine Selbsteinschätzung der Studierenden ist zwar ökonomischer im Einsatz, jedoch wird der Einsatz im Zusammenhang mit der Kompetenzerhebung kritisch beurteilt (Hartig & Jude, 2007). Der Einsatz der oben genannten Simulationsbaukästen wirft andere Probleme auf. So ist insbesondere fraglich, ob auf Hardwareseite die notwendigen Voraussetzungen erfüllt werden können. Konkret betrifft dies die Frage, ob für Kohorten mit mehreren hundert Studierenden ausreichend Desktop- oder Tablet-PCs zu organisieren sind. Zudem müssen die Simulationen programmiert werden und Aufgaben bestimmter Kompetenzfacetten eignen sich auch nicht für den Einsatz eines Simulationsbaukastens.

Diese Vorbetrachtung führte zu der Entscheidung, in diesem Projekt auf Paper-and-pencil-Tests zurückzugreifen. Auch wenn die Validität gerade im Bereich Experimentieren angezweifelt wird (Schreiber, Theyßen, & Schecker, 2009), so sprechen gerade die durch die hohe Standardisierung gewährleistete Testobjektivität und die ökonomische Einsetzbarkeit in Large-Scale-Studien für dieses Instrument (Hartig & Jude, 2007). Die Validität des Testinstruments wird durch eine gründliche theoretische Vorarbeit, an der sich die Aufgabenentwicklung orientiert, gewährleistet. Zusätzlich wird jede Aufgabe durch Experten beurteilt. Die Wahl fiel auf multiple-choice (single-select) Items (MC). Diese sind schneller auszuwerten, wodurch sie ökonomischer in large-Scale-Studien einzusetzen sind. Zudem weisen sie im Vergleich zu offenen Testaufgaben eine höhere Auswertungsobjektivität auf (Moosbrugger & Kelava, 2012). Durch die Vorgabe von

möglichen Antwortmöglichkeiten wird aber die Validität der Aufgaben in Frage gestellt, weshalb diese auf Grundlage einer gründlichen theoretischen Vorarbeit konstruiert werden müssen (Hartig & Jude, 2007).

Konkretes Vorgehen

Im Projekt Ko-WADiS wurden die Testitems anhand des adaptierten Kompetenzstrukturmodells entwickelt (s.o.). Um eine Vergleichbarkeit der unterschiedlichen Items zu erreichen und der oben geforderten gründlichen theoretischen Vorarbeit Genüge zu tun, wurde eine Konstruktionsanleitung für die Items entwickelt. In dieser sind die verschiedenen Zellen des Kompetenzmodells konkret beschrieben und die unterschiedlichen Möglichkeiten der Aufgabenkonstruktion dargestellt. Außerdem wurde für jede Teilkompetenz ein einheitlicher Impuls festgelegt. Beispielhaft sei hier der Bereich *Hypothese* aufgeführt. Hier lautet der Impuls: „Stellen Sie eine naturwissenschaftliche Hypothese zu diesem Phänomen auf." Bewusst wird hier nach einer *naturwissenschaftlichen* Hypothese gefragt, um der postulierten fächerübergreifenden *naturwissenschaftlichen* Erkenntnisgewinnungskompetenz gerecht zu werden. Die Bereichs-Spezifität einer Aufgabe ergibt sich erst aus dem entsprechenden Kontext, der eindeutig einem der drei beteiligten Fächer zuzuordnen ist. Der dargestellte Impuls wurde gegebenenfalls geringfügig an die konkrete Aufgabe angepasst, zum Beispiel, wenn die, einer Untersuchung zugrunde liegende, Hypothese genannt werden sollte.

Ausgehend von den sieben Zellen des Kompetenzmodells (Mayer, 2007; Upmeier zu Belzen & Krüger, 2010) wurde zunächst angestrebt, pro Fach und Zelle zehn Items zu entwickeln. Dieses führt demnach zu 70 Items pro Fach. Ein gangbarer und auch empfohlener Weg zur Konstruktion von multiple-choice-Items ist der Einsatz von halboffenen Aufgaben (Bortz & Döring, 2006). Aus den Antworten der Studierenden lassen sich dann Distraktoren für MC-Items ableiten. In diesem Projekt wurde eine leicht abgewandelte Form dieser Methode verwendet: Es wurden offene Aufgaben eingesetzt, die den Studierenden mehr Freiheiten in der Gestaltung ihrer Antworten geben. Dadurch sollte eine weitere Bandbreite an Antworten erreicht werden. Die Auswertung der Antworten wird dadurch aber aufwändiger, gleichzeitig müssen umfassendere sprachliche Anpassungen vorgenommen werden (Burton, Sudweeks, Merrill & Wood, 1991). Zu bedenken ist dabei auch, dass die Beantwortung einer offenen Frage deutlich mehr Zeit in Anspruch nimmt, als die Beantwortung einer halboffenen oder geschlossenen Frage. Die Testhefte müssen dementsprechend kurz gehalten werden. Die Testhefte für diese Prä-Pilotierung umfassten 14 bis 15 Aufgaben. Zusätzlich wurden mehrere Fragen zu den einzelnen Aufgaben gestellt, die eine bessere Einschätzung erlauben sollten (Straube & Nordmeier, 2013). Die Bearbeitungszeit reichte von zirka einer halben Stunde bis deutlich über eine Stunde, wobei diese Daten Erfahrungswerte sind und nicht speziell protokolliert wurden. Auch der Zusammenhang zwischen Bearbeitungszeit und Qualität der Antworten wurde nicht weiter untersucht.

Insgesamt wurden über 140 Testhefte bearbeitet. Jede Aufgabe wurde damit mindestens zehn Mal bearbeitet, wobei für einige Aufgaben deutlich mehr Antworten vorlagen, weil eine gleichmäßige Ausgabe der Testhefte nicht immer gelang oder einige Studierende mehrere mögliche Antworten zu einem Item notierten. Die Antworten wurden zunächst auf ihre fachliche Richtigkeit hin analysiert und mehrmals vorkommende Antworten zusammengefasst. Aus den erhaltenen Antworten wurden pro Aufgabe drei Distraktoren und ein Attraktor gewonnen. Teilweise lagen so viele Antworten vor, dass aus einem Item zwei Items mit gleichem Aufgabenstamm (aber unterschiedlichen MC-Optionen) erstellt werden konnten. Die MC-Optionen wurden sprachlich vereinheitlicht (Burton, Sudweeks, Merrill & Wood, 1991). Anschließend wurden sie zunächst durch einen Experten des eigenen Faches, später durch mindestens einen Experten der anderen beteiligten Fächer und einen psychometrisch ausgewiesenen Experten (S. Hartmann, HU Berlin) bewertet und

gegebenenfalls überarbeitet (Schecker & Parchmann, 2006). Die Begutachtung der Items durch die fachfremden Experten sollte sicherstellen, dass diese auch durch Studierende der anderen Fächer prinzipiell – nur durch ‚Einsatz' ihrer Erkenntnisgewinnungskompetenz – lösbar sind. Im Anschluss wurden die Items an über 600 Studierenden der beteiligten Fächer pilotiert. Die Items wurden im Sinne des Multimatrix-Designs auf insgesamt 48 Testhefte verteilt.
Die Items wurden anhand von Item- und Modellfitparametern (Itemschwierigkeit, wMNSQ, T-Wert und klassische Trennschärfe) selektiert. Items, die schlechte Itemparameter aufwiesen, wurden aussortiert. Teilweise wurden diese erneut intensiv geprüft und überarbeitet. Außerdem wurde ausgehend von den Ergebnissen der Pilotierung (Stiller et al., in diesem Band) für einige Zellen des Kompetenzmodells Items nachkonstruiert. Dabei konnte auf die Erfahrungen aus der ersten Konstruktionswelle zurückgegriffen werden. So wurden neue Items ausgehend von Items konstruiert, die zufriedenstellende Itemparameter aufwiesen. Diese Items durchlaufen im Wesentlichen die oben dargestellten Qualitätssicherungsschritte und werden anschließend ebenfalls pilotiert.

Ausblick

Es wird angestrebt, bis zum Ende der längsschnittlichen Erhebung jede der Zellen mit mindestens sechs Items (mit zufriedenstellenden Itemparametern) abzudecken. Eine Pilotierung der nachkonstruierten Items muss zeigen, inwiefern dieses Ziel schon erfüllt ist und an welcher Stelle gegebenenfalls noch weitere Items benötigt werden. Die fertigen Items wurden Ende des Sommersemesters 2013 an Studierenden der beteiligten Fächer normiert. Für das Wintersemester 2013/14 ist eine Erhebung bei Studierenden des 1. Bachelor- und Mastersemester in den Lehramtsstudiengängen und den beteiligten Fachwissenschaften geplant (Beginn der längsschnittlichen Erhebung). Weitere Erhebungen finden jeweils halbjährig im 1. bzw. 4. Bachelor- und Mastersemester der beteiligten Fächer statt.

Das Projekt wird im Rahmen des Programms „Kompetenzen im Hochschulsektor" (KoKoHs) durch das BMBF gefördert.

Literatur

Bortz, J., & Döring, N. (2006). *Forschungsmethoden und Evaluation*. Berlin: Springer.

Burton, S., Sudweeks, R., Merrill, P., & Wood, B. (1991). *How to Prepare Better Multiple-Choice Test Items: Guidelines for University Faculty*: Brigham Young University Testing Services and the Department of Instructional Science.

Hartig, J., & Jude, N. (2007). Empirische Erfassung von Kompetenzen und psychometrische Kompetenzmodelle. In J. Hartig & E. Klieme (Eds.), *Bildungsforschung: Vol. 20. Möglichkeiten und Voraussetzungen technologiebasierter Kompetenzdiagnostik. Eine Expertise im Auftrag des Bundesministeriums für Bildung und Forschung* (pp. 17–30). Bonn, Berlin: BMBF.

Mayer, J. (2007). Erkenntnisgewinnung als wissenschaftliches Problemlösen. In D. Krüger & H. Vogt (Eds.), *Theorien in der biologiedidaktischen Forschung. Ein Handbuch für Lehramtsstudenten und Doktoranden* (pp. 177–184). Berlin, Heidelberg: Springer-Verlag Berlin Heidelberg.

Moosbrugger, H., & Kelava, A. (2012). *Testtheorie und Fragebogenkonstruktion* (2nd ed.). *Springer-Lehrbuch*. Berlin, Heidelberg: Springer-Verlag Berlin Heidelberg.

Schecker, H., & Parchmann, I. (2006). Modellierung naturwissenschaftlicher Kompetenz. *Zeitschrift für Didaktik der Naturwissenschaften, 12*, 45–66.

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

Stecher, B. M., & Klein, S. P. (1997). The Cost of Science Performance Assessments in Large-Scale Testing Programs. *Educational Evaluation and Policy, 19*, 1–14.

Straube, P., & Nordmeier, V. (2013). Ko-WADiS – Kompetenzmodell der Erkenntnisgewinnung. In V. Nordmeier & H. Grötzebauch (Eds.), *PhyDid B, Didaktik der Physik, Beiträge zur DPG-Frühjahrstagung*.

Upmeier zu Belzen, A., & Krüger, D. (2010). Modellkompetenz im Biologieunterricht. *Zeitschrift für Didaktik der Naturwissenschaften, 16*, 41–57.