Volker Brüggemann
Volkhard Nordmeier

Freie Universität Berlin

Naturwissenschaftliches Denken im Lehramtsstudium - Computeradaptive Leistungsmessung -

Das Projekt ValiDiS

Das Projekt ValiDiS (Kompetenzmodellierung und -erfassung: Validierungsstudie zum wissenschaftlichen Denken im naturwissenschaftlichen Studium) hat das Ziel, die Entwicklung *naturwissenschaftlichen Denkens* bei Lehramtsstudierenden zu untersuchen. Es schließt an das Projekt Ko-WADiS (Hartmann et al. 2015) an. *Naturwissenschaftliches Denken* wird dabei als eine Kompetenz aufgefasst, die sich in Untersuchungs- und Modellierungsprozessen beobachten lässt (Straube 2016). Aufgeteilt werden diese Prozesse in sieben Handlungsfacetten, die sich aus der Kombination von Kompetenzmodellen zu Erkenntnisgewinnungsprozessen (Mayer 2007) sowie zur Arbeit mit Modellen (Upmeier zu Belzen und Krüger 2010) ergeben. Um diese Kompetenz zu messen, wurde im Projekt Ko-WADiS ein papierbasierter Multiple-Choice-Leistungstest entwickelt.

Das Ko-WADiS-Testinstrument befindet sich nun im Prozess der weiteren Validierung: Im Längsschnitt wird *naturwissenschaftliches Denken* über Studienverläufe hinweg (eindimensional) erfasst, wobei eine EAP/PV Reliabilität von .65 erreicht wird[1]. Vorläufig liegen nur Ergebnisse von Vergleichen verschiedener Kohorten im Querschnitt vor, da die Längsschnitt-Beobachtung nicht abgeschlossen ist. Hier zeigen sich erwartungskonform ansteigende Leistungen der Studierendengruppen in höheren Semestern. Ebenfalls kann man theoretischen Annahmen entsprechende Unterschiede zwischen bekannten Gruppen erkennen. Aufgrund dieser Datenlage wird davon ausgegangen, dass die Auslegung der Messwerte im Sinne eines Kompetenzmaßes valide ist (Straube 2016). Zusätzlich erscheint der Test in laufenden Interventionsstudien als sensitiv genug, um Kompetenzverläufe im Rahmen von einzelnen Lehrveranstaltungen aufzulösen.

Damit erscheint ein Einsatz in der Lehrevaluation als vielversprechend: Im Rahmen von kompetenzorientierten Studiengängen ist es wünschenswert, Lehrveranstaltungen nicht nur im Hinblick auf ihre strukturelle Güte, sondern auch auf die erreichte Kompetenzförderung hin zu untersuchen. Mit dem vorliegenden Instrument besteht die Chance, dies in den Naturwissenschaften (bezogen auf die Förderung von wissenschaftlichen Denkweisen) fächerübergreifend zu tun.

Aktuelle Herausforderungen

Beim angedachten Einsatz des Tests in Evaluationsszenarien ergibt sich aber eine praktische Hürde: Momentan absolvieren die Proband*innen den Test mit einer Gesamtlänge von 21 Items pro Heft in ca. 35 Minuten. Bei einer Lehrevaluation würde mit einer entsprechenden Vorbereitung sowie zweifacher Durchführung für eine Prä-Post-Messung mindestens ein ganzer Veranstaltungstermin in Anspruch genommen werden. Es ist daher für eine angestrebte Anwendung in solchen Situationen wünschenswert, die Testdauer zu verkürzen.

[1] Anm.: Dieser Wert ist den üblichen ‚Faustregeln' nach zwar als schlecht einzuordnen, ordnet sich aber in die Ergebnisse anderer Kompetenztests im Bereich *Erkenntnisgewinnung* ein (vgl. Wellnitz 2012; Woitkowski 2015).

Im Hinblick auf die noch zu optimierende Reliabilität erscheint eine Kürzung des Instruments zunächst als kritisch. Zudem wird vermutet, dass die Konzentration/Motivation gegen Ende der Erhebungen stark nachlässt (diese wurde im Längsschnitt nicht erhoben, es handelt sich lediglich um Beobachtungen der Testleiter*innen). Es ist also anzunehmen, dass die Items am Ende einer Erhebung weniger messgenaue Daten liefern. Da die Items je nach Fragebogenversion in der Reihenfolge vertauscht sind, würde sich ein solcher Effekt nicht anhand von ausgewählten, schlechteren Items zeigen: Es wird vermutet, dass er sich in einer sinkenden Reliabilität des gesamten Instrumentes niederschlägt. Zusätzlich scheinen die Proband*innen trotz allem bemüht, den Test vollständig zu beenden. Daher führt auch der Einbezug von Ratewahrscheinlichkeiten im Auswertungsmodell nicht zu einer Lösung dieses Problems. Sollte die geschilderte Vermutung stimmen, so könnte eine Verbesserung der Testeffizienz zu einer Steigerung der Reliabilität führen.

Adaptive Testverfahren

Eine Idee zur Steigerung der Testeffizienz ist die Nutzung adaptiver Verfahren (Weiss 1982). Um deren mögliche Vorteile gegenüber linearer Verfahren zu verdeutlichen, sollen nun beide Konzepte kurz skizziert werden.

In *linearen Testverfahren*, z. B. in Papierform, wird allen Teilnehmenden eine konstante Anzahl von Items in einer festen Reihenfolge präsentiert. Dabei ergibt sich eine besondere Eigenschaft, sofern das Instrument auf der Item-Response-Theory basiert: Im Itempool gibt es Aufgaben mit einer breiten Spanne an Schwierigkeiten und in der Gruppe der befragten Personen eine breite Verteilung an Fähigkeitsausprägungen. Gelöste Aufgaben geben aber nur eine verwertbare Information über Probandenfähigkeiten, wenn sie von ihrer Schwierigkeit zur Fähigkeit der Proband*innen passen. Um in einer Befragung alle Teilnehmenden genau beurteilen zu können, muss also jeder Fragebogen genug Items der verschiedensten Schwierigkeiten enthalten. Im Umkehrschluss folgt dann aber auch, dass ausnahmslos alle Proband*innen zahlreiche Aufgaben lösen, die nicht auf sie passen und nur wenig Information liefern.

Adaptive Testverfahren können diesen Umstand umgehen (SARI et al. 2016). Während der Testanwendung wird, nachdem erste Items bearbeitet wurden, die Fähigkeit des/der Probanden/in individuell von einem Algorithmus geschätzt. Dies geschieht auf der Grundlage zuvor festgesetzter Item-Kennwerte und den bisher gegebenen Antworten. Die geschätzte Personenfähigkeit wird verwendet, um im Folgenden optimal zu den jeweiligen Proband*innen passende Aufgaben auszuwählen (Frey 2012). Durch mehrfache Wiederholung dieses Vorgangs kann der Test die Schätzung und Item-Auswahl verfeinern und somit adaptiv auf den/die einzelne/n Probanden/in reagieren. Vergleichende Studien zeigen, dass adaptive Testverfahren gegenüber linearen Instrumenten die Testeffizienz deutlich erhöhen können (vgl. z. B. Weiss 1982).

Wie häufig die erwähnte Schätzung durchgeführt wird, ist von Test zu Test unterschiedlich. „Echte“ adaptive Tests führen sie nach jeder Aufgabe durch. Demgegenüber gibt es aber auch Multistage-Tests (MSTs) (Hendrickson 2007). Hier werden die Schätzungen immer zwischen Blöcken aus Aufgaben durchgeführt, den sogenannten Testlets. Die einzelnen Testlets werden so konstruiert, dass sie Aufgaben gleicher Schwierigkeit aus allen Inhaltsbereichen abdecken. Die Testlets fungieren also als eine Auswahl verschieden schwerer Versionen des gesamten Instruments.

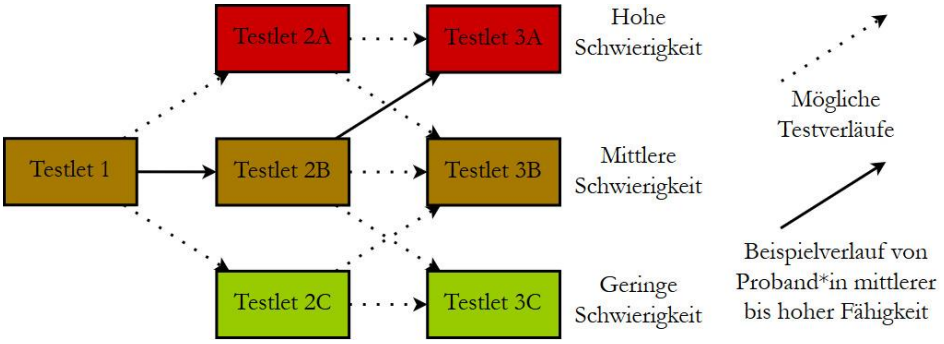

Abbildung 1: Adaptiver Multi-Stage-Test im 1-3-3 Design (Zheng und Chang 2014)

Vorhaben: Implementation eines adaptiven Testverfahrens
Wie oben beschrieben, steht dem regelhaften Einsatz des Ko-WADiS-Instruments (z. B. im Rahmen von Lehrevaluationen) derzeit noch die Testdauer im Wege. Zudem besteht die Hoffnung, dass eine Erhöhung der Testeffizienz in Kombination mit einer Verkürzung der Testdauer die Reliabilität des Instruments verbessern kann. Aus diesem Grund ist geplant, aus dem bestehenden Instrument eine adaptive Version zu entwickeln.

Im Rahmen des Projekts ValiDiS werden derzeit erneut Modellrechnungen mit den gesammelten Daten durchgeführt, um auf Grundlage der bestmöglichen Personenschätzer die Itemkennwerte festzusetzen. Die erneute Modellierung ist vorgesehen, da bisher stets Gruppenschätzungen durchgeführt und keine einzelnen Proband*innen in den Fokus genommen wurden. Danach wird eine Itemselektion erfolgen, um die Testlets zusammenzustellen. Aktuell ist ein dreistufiges Design mit drei Schwierigkeitsbereichen geplant (1-3-3-Design, siehe Abbildung 1). Alle Proband*innen würden somit nur je 15 Items anstelle der bisherigen 21 bearbeiten.

Ausblick
Die grundlegende Testadaption soll bis Ende 2017 abgeschlossen sein. Danach folgt eine erste Pilotierungsphase zum Ende des Wintersemesters 2017/18, in der auf die Optimierung technischer Aspekte fokussiert wird: Weboberfläche, Algorithmus sowie die Erstellung einer automatisierten Datenbank. In einer zweiten Phase sollen dann im Sommersemester 2018 die einzelnen Testlets optimiert werden.

Der Pilotierung werden später vergleichende Studien zwischen der adaptiven und der papierbasierten Version des Tests folgen, um Unterschiede in Testeffizienz und Messgenauigkeit zwischen den beiden Formaten zu untersuchen.

Literaturverzeichnis

Frey, Andreas (2012): Adaptives Testen. In: Helfried Moosbrugger und Augustin Kelava (Hg.): Testtheorie und Fragebogenkonstruktion. Berlin/Heidelberg: Springer Berlin Heidelberg, S. 275–293.

Hartmann, Stefan; Mathesius, Sabrina; Stiller, Jurik; Straube, Philipp; Krüger, Dirk; Upmeier zu Belzen, Annette (2015): Kompetenzen der naturwissenschaftlichen Erkenntnisgewinnung als Teil des Professionswissens zukünftiger Lehrkräfte: Das Projekt Ko-WADiS. In: Koch-Priewe, Anne Köker, Jürgen Seifried und Eveleine Wuttke (Hg.): Kompetenzerwerb an Hochschulen: Modellierung und Messung. Zur Professionalisierung angehender Lehrerinnen und Lehrer sowie frühpädagogischer Fachkräfte. Bad Heilbrunn: Klinkhardt, S. 39–58, zuletzt geprüft am 10.03.2016.

Hendrickson, Amy (2007): An NCME Instructional Module on Multistage Testing. In: *Educational Measurement: Issues and Practice* 26 (2). Online verfügbar unter http://onlinelibrary.wiley.com/store/10.1111/j.1745-3992.2007.00093.x/asset/j.1745-3992.2007.00093.x.pdf?v=1&t=j63j9ui7&s=dc089570fc06bbf1a511a6a118ac93bf691861d9, zuletzt geprüft am 08.08.2017.

Mayer, Jürgen (2007): Erkenntnisgewinnung als wissenschaftliches Problemlösen. In: Dirk Krüger und Helmut Vogt (Hg.): Theorien in der biologiedidaktischen Forschung. Ein Handbuch für Lehramtsstudenten und Doktoranden. Berlin, Heidelberg: Springer-Verlag Berlin Heidelberg (Springer-Lehrbuch), zuletzt geprüft am 26.09.2017.

SARI, Halil Ibrahim; YAHSI-SARI, Hasibe; Corinne HUGGINS-MANLEY, Anne (2016): Computer Adaptive Multistage Testing. Practical Issues, Challenges and Principles. In: *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, S. 388. DOI: 10.21031/epod.280183.

Straube, Philipp (2016): Modellierung und Erfassung von Kompetenzen naturwissenschaftlicher Erkenntnisgewinnung bei (Lehramts-)Studierenden im Fach Physik. Berlin: Logos (Studien zum Physik. und Chemielernen, 209), zuletzt geprüft am 30.08.2016.

Upmeier zu Belzen, Annette; Krüger, Dirk (2010): Modellkompetenz im Biologieunterricht. Model competence in biology teaching. In: *Zeitschrift für Didaktik der Naturwissenschaften* 16, S. 41–58, zuletzt geprüft am 10.03.2016.

Weiss, David J. (1982): Improving Measurement Quality and Efficiency with Adaptive Testing. In: *Applied Psychological Measurement* 6 (4), S. 473–492. DOI: 10.1177/014662168200600408.

Wellnitz, Nicole (2012): Kompetenzstruktur und -niveaus von Methoden naturwissenschaftlicher Erkenntnisgewinnung. Zugl.: Kassel, Univ., Diss., 2012. Berlin: Logos-Verl. (Biologie lernen und lehren, 2).

Woitkowski, David (2015): Fachliches Wissen Physik in der Hochschulausbildung. Konzeptualisierung, Messung, Niveaubildung. Zugl.: Paderborn, Univ., Diss., 2015. Berlin: Logos-Verl. (Studien zum Physik- und Chemielernen, 185).

Zheng, Yi; Chang, hua-hua (2014): Multistage testing, on-the-fly multistage testing, and beyond. In: Ying Cheng und hua-hua Chang (Hg.): Advancing methodologies to support both summative and formative assessments (Chinese American Educational Research and Development Association book series).